ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

MOLECULAR DIAGNOSIS WITH EXOME SEQUENCING RE-ANALYSIS OF UNDIAGNOSED AND RARE DISEASES

SEZER AKYÖNEY

M.Sc. THESIS

DEPARTMENT OF BIOSTATISTICS AND BIOINFORMATICS

SUPERVISOR
Assoc. Prof. Özden Hatırnaz Ng

SECONDARY SUPERVISOR
Dr. Özkan Özdemir

ISTANBUL-2022

ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

MOLECULAR DIAGNOSIS WITH EXOME SEQUENCING RE-ANALYSIS OF UNDIAGNOSED AND RARE DISEASES

SEZER AKYÖNEY

M.Sc. THESIS

DEPARTMENT OF BIOSTATISTICS AND BIOINFORMATICS

SUPERVISOR
Assoc. Prof. Özden Hatırnaz Ng

SECONDARY SUPERVISOR
Dr. Özkan Özdemir

ISTANBUL-2022

Department:	Biostatistics and Bioinformatics
Program:	Biostatistics and Bioinformatics
Thesis Title:	Molecular Diagnosis of Undiagnosed and Rare Disease Patients with Whole Exome and Whole Genome Re-Analysis
Student's name and Surname:	Sezer Akyöney
Date of Defence:	/ /

This is to certify that I have examined this copy of master thesis. I have found that she/he prepared after fulfilling the specified requirements in the associated legislations before the final examining committee whose signatures are below.

Jury Member (Head of the Defense)	Title, Name Surname Instituition	Signature
Jury Member (Thesis Supervisor)	Title, Name Surname Instituition	Signature
Jury Member (Thesis Co-Supervisor)	Title, Name Surname Instituition	Signature
Jury Member	Title, Name Surname Instituition	
Jury Member		

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 ethicel 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.

23/06/2022

Sezer Akyöney

PREFACE AND ACKNOWLEDGEMENT

First and foremost, I would like to express my heartfelt gratitude to my supervisors, Assoc. Prof. Özden Hatrnaz Ng and Özkan Özdemir, PhD who provided all kinds of support in the writing and making of this thesis, correcting my mistakes.
Secondly, I would like to thank to Prof. Yasemin Alanay for providing patients for this thesis and working simultaneously with us in the evaluation of patients. I would like to thank ACURARE and its director Prof. Uğur Özbek, who provided us with the appropriate environment and equipment for this thesis. In addition, I would like to thank Berk Ergün, who showed his talent in the coding and design of ACUGEN. I would also like to thank Merve Gündoğdu, who supported me in my wet-lab experiments in my limited time. Most importantly, I would like to express my endless thanks and gratitude to my mother and to my father, who I lost while doing this thesis, for providing all kinds of financial and moral support so that I could complete my education and this thesis. Finally, I would like to express my endless thanks to Didem Çakırsoy, who showed the insignificance of distances with her love and support during this intense and challenging period.

This thesis study supported by Acıbadem Mehmet Ali Aydınlar University Scientific Research Projects Commission Presidency with the project numbered 2021/02/01.

TABLE OF CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

1000G	1000 Genome Project
ACMG	American College of Medical Genetics and Genomics
ASCII	American Standard Code for Information Interchange
BAM	Binary Alignment Format
BCL	Binary Base Call
BED	Browser Extensible Data
bp	Base pair
BQSR	Base Quality Score Recalibration
BWA	Burrows-Wheeler Alignment
CI	Confidence Interval
CNN	Convolutional Neural Network
CNV	Copy Number Variation
ddNTP	Dideoxynucleotidetriphospates
dH_2O	Distilled water
DNA	Dioxynucleicacid
dNTP	Dioxynucleotidetriphospates
GATK	Genome Analysis Toolkit
GB	Gigabyte
GC	Guanin-cytosine
GME	Great Middle East Project
gnomAD	The Genome Aggregation Database
GO	Gene Ontology
GRCh	Genome Reference Consortium
GWAS	Genome-Wide Association Study
HGP	Human Genome Project
HGVS	Human Genome Variation Society
HPC	High-Performance Computing
HTML	Hypertext Markup Language
INDEL	Insertion-Deletion
mRNA	Messenger Ribonucleic acid

ng	Nanogram
NGS	Next Generation Sequencing
Nm	Nanomol
Oe	Observed/Expected
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biotechnology
PCR	Polymerase Chain Reaction
RAM	Random Access Memory
RD	Rare Diseases
RNA	Ribonucleic acid
RPKM	Reads Per Kilo Base Per Million Mapped Reads
SAM	Sequence Alignment Map
SCF	Sanger Chromatograph Format
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variation
ssDNA	Single Strand Deoxyribonucleic acid
SV	Structural Variant
TF	Transcription Factor
TG	ThirGenerationSequencing
UCSC	University of California, Santa Cruz
UV	Ultra Violet
VCF	Variant Calling Format
VUS	Variant with uncertain significance
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
µl	Microliter
µM	Micromolar

LIST OF FIGURES

LIST OF TABLES

ÖZET

Tanısız ve Nadir Hastalıklarda Yeniden Ekzom Dizileme Analizi ile Moleküler Tanı

Nadir hastalıklar, Avrupa'da 2000 kişiden birini etkileyen hastalıklar olarak tanımlanmaktadır. Literatüre göre halihazırda bilinen 7000 farklı hastalık nadir olarak kabul edilmektedir ve nadir hastalıkların %80'i genetik temellidir. Daha önce tanımlanan hastalıklar ile açıklanamamış veya bugüne kadar yeterli tanı testlerinin bulunmaması nedeniyle teşhis koyulamayan nadir hastalıklardan etkilenen bireyler de mevcuttur. Tanı yöntemlerindeki gelişmeler, genetik bozukluklarda tanı oranlarını artırmıştır. Günümüzde yeni nesil dizileme, tüm genomun yüksek çözünürlüklü taranmasını sağlayan, moleküler teşhiste en yaygın kullanılan yöntemlerden biridir. Bununla birlikte, yeni nesil dizileme uygulamalarının başarı oranları %50'nin altındadır. Geliştirilen yeni teknolojilerin yanı sıra artan literatür bilgileri, gelişen biyoinformatik analiz araçları ve daha kapsamlı filtreleme yöntemleri ile bugüne kadar üretilmiş yeni nesil dizileme verilerinin başarı oranlarını artırmak mümkündür. Bu araştırmanın bir parçası olarak, nadir görülen hastalıkların teşhisi için özel olarak tasarlanmış bir biyoinformatik iş akışı geliştirdik. Farklı algoritmalarla geliştirilen birden fazla varyant çağırıcıyı birleştirerek, bir genomdaki tüm tek nükleotid varyantlarını, insersiyon-delesyon varyantlarını ve yapısal varyantları bulmayı hedefledik. Ayrıca klinisyenler tarafından hazırlanan derin fenotipleme ile gen listelerimizi büyüttük ve daha geniş bir filtreleme gerçekleştirdik. Çalışmalar sonucunda birçok hasta tanısız kalmaya devam etti. Bunun dışında iki hastada kesin tanı alırken, diğer üç hastanın varyantı doğrudan hastalıkla ilişkilendirilemedi, ancak güçlü adaylar olarak bildirildi. Bu çalışmada derin fenotipleme ve genişletilmiş filtrelemenin önemini ve nadir hastalıklar için geliştirdiğimiz ACUGEN iş akışının daha önce tanı alamamış hastaların yeniden analiz yoluyla tanı alabileceğini gösterdik.

Anahtar kelimeler: Nadir hastalıklar, biyoinformatik, tekrar analiz, tanısız hastalıklar, moleküler tanı.

ABSTRACT

Rare diseases are described as affecting one in 2000 people in Europe. According to the literature, there are already known 7000 different conditions considered rare, and 80% of rare diseases have a genetic background. There are rare disease patients who have not been diagnosed with previously described disorders or because adequate diagnostic tests are not available to date. Improvements in diagnostics methods have increased the rates of diagnosis of genetic diseases. Nowadays, next-generation sequencing is one of the most widely used molecular diagnostics methods that provide high-resolution screening of the entire genome. Nevertheless, the success rates of next-generation sequencing applications are less than 50%. Apart from the new technologies developed, it is possible to increase the success rates of the next generation sequencing data produced until today, with the increasing literature information, development of bioinformatics analysis tools and more comprehensive filtering methods. We developed a pipeline specifically designed for the diagnosis of rare diseases as part of this research. By combining multiple variant callers developed with different algorithms, we aimed to find all single nucleotide variants, insertion-deletion variants, and structural variants on a genome. In addition, we expanded our gene lists with deep phenotyping prepared by clinicians and performed more extended filtering. Many patients went undiagnosed as a result of the studies. Aside from that, two patients had a definitive diagnosis, and three other patients' variants could not be associated with the disease directly, but they were reported as strong candidates. We demonstrated the importance of deep phenotyping and extended filtering in this study and how the ACUGEN pipeline we developed for rare diseases enables previously undiagnosed patients to be diagnosed through re-analysis.

Keywords: Rare diseases, bioinformatics, re-analysis, undiagnosed diseases, molecular diagnostics.

1. INTRODUCTION

In contrast to common diseases, a rare disease affects a small number of individuals, but it is known that rare diseases in total can create a mid-sized continent. To consider a disease as rare differs in different regions. In Turkey, rare diseases are defined as diseases that affect one in two thousand people in the population, the same as in Europe (1). There are approximately 7000 rare diseases defined in the literature, and in addition to these, there are still a number of undiagnosed rare disease patients remaining (2). Patients who have a condition that has not been previously described or for which a diagnostic test is not yet available are defined as undiagnosed disease patients (3).

Today, next-generation sequencing (NGS) technologies play an important role in the diagnostics of genetic diseases. NGS provides high-resolution screening of a part of the entire genome (4). Most genetic diseases are caused by the alterations in the coding regions of the genome, and sequencing of these coding regions, whole-exome sequencing (WES), has helped to define hundreds of genetic disorders. Nevertheless, the diagnosis rate of the WES is around 30% (5). Whole-genome sequencing (WGS) is generally used when whole-exome sequencing is insufficient. WGS is the most comprehensive method applied today to display intronic and structural variants, especially in the absence of a significant variant in the coding regions (6).

Variant filtering and interpretation are crucial steps for the molecular diagnosis of rare diseases. Generally, a rare disease patient is diagnosed around six years after many negative test results (7).

The development of bioinformatics approaches and medical literature and re-analysis of the generated data may avoid unnecessary testing. Also, deep phenotyping and advanced filtering applications may improve the diagnosis rates (8). For a successful reanalysis, updating the applied pipelines and annotations according to the recent developments in bioinformatics is very important.

Here we aimed to improve the diagnostic rate of WES/WGS analysis in patients with negative test results or with results that cannot explain the phenotype. We developed a new analysis flow consisting of the most recent advanced tools and aimed to improve the diagnosis rate with the support of deep phenotyping and advanced filtering.

2. BACKGROUND

2.1. Rare Diseases

The common disease-common variant (CD-CV) hypothesis states that common disease-causing alleles, or variants, will be found in all human populations that display a certain condition. In the coding and regulatory sequences of genes, common variations (not necessarily disease-causing) are known to exist. Some of these variations, according to the CD-CV hypothesis, increase susceptibility to complex polygenic disorders. Each gene mutation that affects the phenotype of complex disease will have a minor additive or multiplicative effect. Some diseases may be rare in one region but not in another (e.g., Familial Mediterranean Fever is considered common on the Mediterranean coast but in no other regions.). Also, there are many common diseases with rare variants (e.g., subtypes of some cancers like acute lymphoblastic leukemia)(9) Rare diseases (RD) are defined as diseases that affect 1/2000 people in Europe (1). According to literature, 80% of RDs have a genetic aetiology, but according to Orphadat,a only 39% of RDs' genetic background is identified (3). Except for rare diseases with genetic background, there are very rare forms of infectious diseases also grouped as rare diseases. RDs are primarily chronic, progressive, and severe groups of diseases and mostly occur in the first years of life. Today, the cause of most of the RDs remains unknown (10).

There are around six to seven thousand rare diseases that have been described in the medical literature (1). Our understanding of the genetic factors that might cause rare genetic diseases has vastly improved over the last 30 years (11). Chromosomal rearrangements, copy number variations, trinucleotide repeats, indels, SNVs, mitochondrial mutations, and epigenetic modifications have been added to our knowledge to understand these factors (12). These research insights brought new diagnostic techniques with them. Karyotyping and Sanger sequencing techniques were the gold standard for the last 40 years in diagnostics of genetic disorders and the diagnosis rate improved with the installment of microarrays and NGS (13). On the other hand, there is still a high number of rare disease patients who remain

undiagnosed due to a lack of access to the right kind of expertise or testing (14). According to patient surveys, patients with rare diseases require an average of 7.6 years in the United States and 5.6 years in the United Kingdom to receive a diagnosis (15).

Science can provide some answers for all rare disorders. Biological samples can now be used to diagnose hundreds of rare diseases. The construction of registries advances our understanding of these diseases' aetiology. Researchers are increasingly using networks to discuss their findings and progress their careers more quickly. The perspectives presented by European and national policies in the field of rare diseases in several European countries provide new optimism.

2.2. Undiagnosed Diseases

A rapidly growing area of genomic medicine is establishing a diagnosis for individuals with complex phenotypes (or combinations of phenotypes) that have eluded conventional medical examination. The term undiagnosed is used for diseases that, despite all the current clinical and laboratory applications, the patients cannot receive a clear diagnosis. Also, these patients may be misdiagnosed due to complex and heterogeneous symptoms and incomplete patient examination. Early diagnosis is crucial for prognosis and survival. Misdiagnosis will prolong the time of diagnosis, and wrong drug use may cause irreversible damage to the patient. Most undiagnosed patients have a rare disease that couldn't be diagnosed (15). The reason most undiagnosed diseases are considered rare is that differential diagnoses between rare diseases are so narrow. Only one different mild phenotype may change the cause of the disease.

According to Undiagnosed Diseases Network International (UDNI), Undiagnosed rare diseases are disorders that defy a referring physician's diagnosis; some patients wait years for a final diagnosis. Undiagnosed rare diseases might include unidentified disorders with similar symptoms, diseases with well-defined phenotypes, diseases with unknown molecular causes, or diseases caused by unknown, non-genetic sources (16).

The National Institutes of Health Undiagnosed Diseases Program and, more recently, the Undiagnosed Diseases Network revealed early achievements, prompting the creation of the global UDNI endeavour, which now includes programs in 16 countries. The utilization of genomics as a key component of the diagnostic process is a common feature of these programs (17).

As evidenced by pediatric applications to the National Institutes of Health (NIH) patients may remain for years without a diagnosis under the Undiagnosed Diseases Program, in the range of 4–6 years of age and 16–18 years of age (18).

2.3. The Human Genome Project

The Human Genome Project (HGP) was an international, collaborative scientific project whose goal was to map and comprehend all human genes completely. It is still the largest collaborative biological project in the world. The project formally debuted in 1990 and was declared largely completed on April 14, 2003, but only comprised roughly 85 percent of the genome. In May 2021, the level of "complete genome" was achieved, with only 0.3 percent of nucleotides remaining with potential concerns. In January 2022, the missing Y chromosome was added. The project was planned by the National Health Institute and National Science Academy Committee in 1988. It started with the “first five years plan” in April 1990, and scientists who were working on the project published the physical map of the human genome. In December 1999, the 22nd chromosome, consisting of 33.5 million letters, was first decoded.

According to the HGP, there are approximately 20,500 coding human genes. The HGP's final product has provided a wealth of precise knowledge about the structure, organization, and function of the whole set of human genes to the rest of the world.

In February 2001, the International Human Genome Sequencing Consortium published the first draft of the human genome in Nature and Science Journals. The genome's three billion base pairs were sequenced to nearly 90% completeness.

HGP has different outputs besides the human genome sequence. Characterize the whole genomes of several additional organisms that are commonly utilized in biological research, such as mice, fruit flies, and flatworms. Because most organisms have many related or "homologous" genes with similar functions, these efforts complement one another. As a result, determining a gene's sequence or function in a model organism, such as the roundworm *C. elegans*, has the potential to explain a homologous gene in humans or one of the other model species (Table 1).

Table 1. The achievements of the HGP (19).

Area	Goal	Achieved	Date
Genetic Map	2- to 5-cM resolution map (600 - 1,500 markers)	1-cM resolution map(3,000 markers)	September 1994
Physical Map	30,000 STSs	52,000 STSs	October 1998
DNA Sequence	99.99 percent of the human sequence comprising genes was completed with 99.99 percent accuracy.	99 percent of the human sequence comprising genes was completed with 99.99 percent accuracy.	April 2003
Capacity and Cost of Finished Sequence	Sequence 500 Mb/year at < $0.25 per finished base	Sequence >1,400Mb/year at <$0.09 per finished base	November 2002
Human Sequence Variation	100,000 human SNPs mapped	3.7 million mapped human SNPs	February 2003
Gene Identification	Complete human cDNAs	15,000 complete human cDNAs	March 2003
Model Organisms	Whole-genome sequences of *E. coli, S .cerevisiae, C. elegans, D. melanogaster*	*E. coli, S. cerevisiae, C. elegans,* and *D. melanogaster* complete genome sequences, as well as whole-genome drafts of *C. briggsae, D. pseudoobscura,* mouse, and rat	April 2003
Functional Analysis	Develop genomic-scale technologies	DNA microarrays using high-throughput oligonucleotide synthesis	1994
		Eukaryotic, whole-genome knockouts (yeast)	1996
		The two-hybrid technique for protein-protein interaction has been scaled up.	1999

The Genome Reference Consortium (GRC) produced the current human reference genome in 2013 and updated it in 2019 (GRCh38.p13) (20). The current reference genome, GRCh38 comprises 151 Mbp of unidentified sequence, including

pericentromeric and subtelomeric areas, terminal segmental duplications, ampliconic gene arrays (eight palindromes, three inverted repeats, two arrays of no long open reading frames, and testis-specific Y repeats), and ribosomal DNA (rDNA) arrays, all of which are required for essential cellular activities (21). Recently, the Telomere-to-Telomere (T2T) Consortium presented T2T-CHM13, a 3.055 billion–base pair human genome sequence that includes gapless assemblies for all chromosomes except Y. T2T-CHM13, corrected errors in previous references, and presented nearly 200 million base pairs of sequence containing 1956 gene predictions, 99 of which are predicted to be protein-coding (22).

2.4. Reference Genome

A reference genome (also known as a reference assembly) is a digital nucleic acid sequence database established by scientists to represent the set of genes in a unique idealized individual organism of a species. Reference genomes do not correctly represent the set of genes of any single individual organism because they are generated through the sequencing of DNA from a number of individual contributors (23). A reference, on the other hand, gives a haploid mosaic of various DNA sequences from each donor. The majority of the population who have had their full genome sequenced, such as James D. Watson, who is the first sequenced individual, had their genome assembled in this method (21). Certain changes to the reference genome sequences are taking place as a result of breakthroughs in the scientific community. Patches to the current genome or a new genome version may be developed, depending on the size of these advancements. Patches are scaffold sequences that have been accessioned and indicate assembly updates. There are two types of patches; fix patches and novel patches. Changes to current assembly sequences are represented by fixedpatches. Error fixes (handled by methods such as base changes, component replacements/updates, switch point updates, or tiling path adjustments) or assembly improvements (such as sequence expansion into gaps) are the most common. The inclusion of new alternate loci to the assembly is represented by novel patches. These are alternative sequence representations of chromosomal sequences. The accessions for the innovative patch

scaffolds will persist in the next major release, and the scaffolds will be known as alternate loci (24).

2.5. Human Genome

The human genome is composed of 3 billion nucleotides, which carry the genetic information required to determine all elements of life. DNA is a polymeric nucleic acid macromolecule consisting of three different sorts of units: a five-carbon sugar (deoxyribose), a nitrogen-containing base, and a phosphate group. Polymerization of nucleotides results in long polynucleotide chains linked together from the fifth carbon to the third carbon (5' to 3') direction that form a double-helix structure with the opposite strand (25). The two groups of nitrogenous bases comprising the two groups of nucleotide bases are purines and pyrimidines. Guanine (G) and adenine (A) are purine nucleotide bases that distinguish deoxyribonucleotides (deoxyadenosine and deoxyguanosine) and ribonucleotides (deoxyguanosine and deoxyguanosine) (adenosine, guanosine). DNA and RNA are made up of nucleotides.

Cytosine (C), thymine (T), and uracil (U) are pyrimidine derivatives that can be found in nucleic acids. These bases and their complementary purines form hydrogen bonding in DNA and RNA. The purines adenine (A) and guanine (G) pair up respectively with the pyrimidines thymine (T) and cytosine (C) in DNA.

Coding Regions of the Genome

Each nucleated cell in the body has its own copy of the human genome, which contains approximately 20,000 coding genes and approximately 30,000 non-coding genes, which are out of the scope of this study. Genes are the basic units of DNA, encoding the inherited structural and regulatory elements of the organisms. Genes play important roles in health and disease. The genes are ordered linear along the chromosomes, with each gene having a specific location or locus. Each chromosome carries a different subset of genes that are ordered linearly along with its DNA. Members of a pair of chromosomes (also known as homologous chromosomes or

homologues) have genetic information that is identical; they usually have the same genes in the same order. On the other hand, alleles are the alternative forms of a DNA sequence at a specific locus. For many genes, there is a single dominant allele, known as the wild-type or common allele, that is found in more than half of a population. The variant (or mutant) alleles differ from the wild-type allele due to the existence of a change (26).

Depending on their functions, there are two groups of genes in the genome: protein-coding genes and non-coding genes (27). There is no one-to-one correspondence between proteins and genes. A gene may encode more than one protein that has different functionalities by an alternative splicing mechanism (28). Genes also have regions that are coding and noncoding. Coding regions are translated to mRNAs, called exons, and noncoding regions of a gene are called introns (Figure 1). With the splicing mechanisms, a gene can be transcribed to different mRNAs, which can be translated by different exon combinations (29).

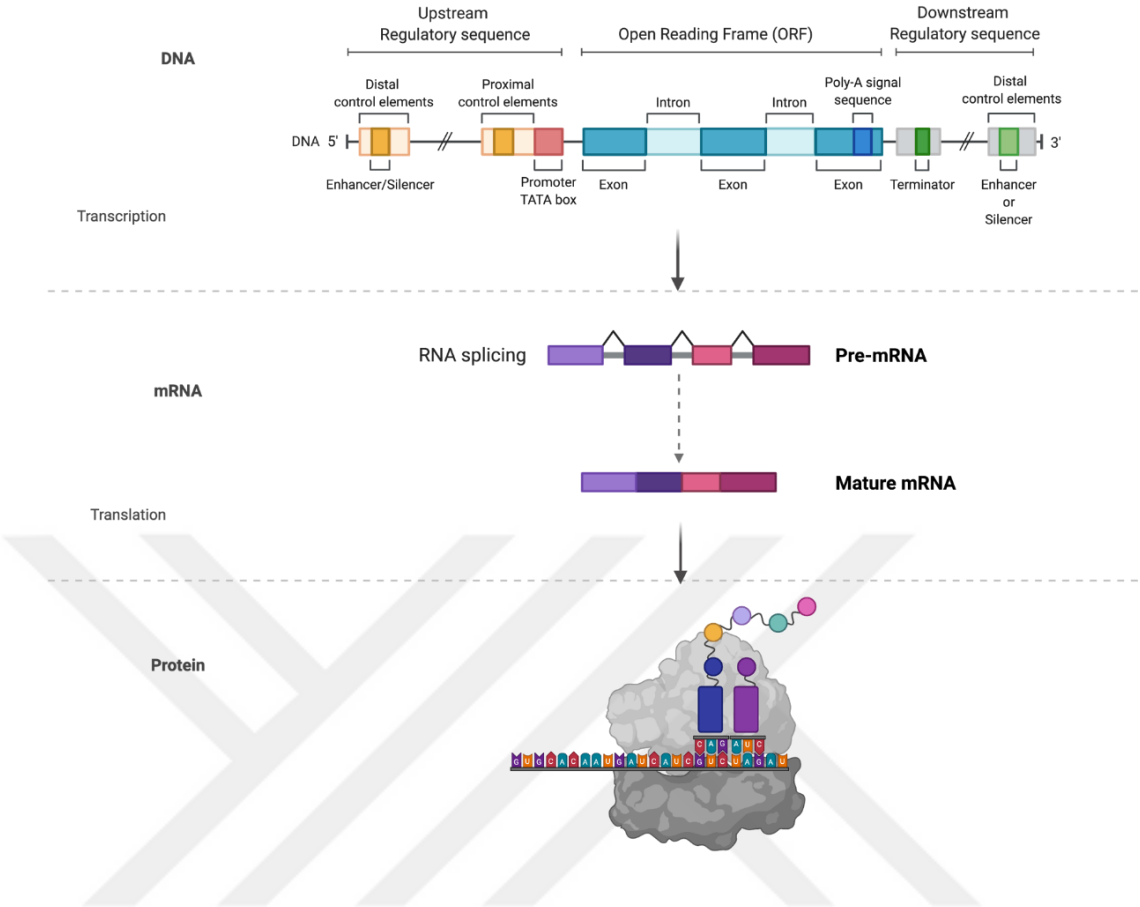


Figure 1. The typical gene structure in humans and the central dogma.

The figure shows the regulatory locations, which include distal control elements, enhancers/silencers (gold), and promoter TATA box (light coral), which contains several Transcription Factor (TF) binding sites. Next to the promoter site, the transcribed region has been shown with its exons (light sea green), introns (ice blue), and Poly-A signal sequence (steel blue). Then, regulatory sequence Following the transcription and posttranscriptional modifications, mature mRNA occurs (30). At last, the transcription of the protein (31) (This figure is drawn with BioRender application).

2.5.1. Variations

Despite the fact that two people's genomes are 99.5 percent the same, each person has a distinct phenotype, with the exception of monozygotic twins. (32). This genetic and phenotypic diversity mainly occurs due to the variations. The frequencies of these variations may vary among different populations. If a locus in a population has two or more relatively common alleles (defined by convention as having an allele frequency > 1 percent), that locus is said to be polymorphic (33). According to HGVS, however, the terms "polymorphism" and "mutation" are no longer in use due to their unclear meanings in colloquial usage. Polymorphism can be misleading because, in some areas, it refers to a non-disease-causing sequence variation, while in others, it refers to a variant identified at a frequency of more than 1% in a population. Hence, “variant”, “change,” “alteration” terms are suggested for changes in the DNA sequences.

Variations are mainly classified into three groups. Variants that change the number of chromosomes are called genome variants, such as aneuploidies. Also, “regional variants” that change only a part of a chromosome and might change the copy number of sub-chromosomal segments or a structural rearrangement, are called chromosomal variations. Lastly, modifications of the DNA sequence, which involve substitution, deletion/duplication, and insertion that range from single nucleotide to 100 kb, are called gene or DNA variations (34).

This thesis primarily covers the gene variations, and these variations will be explained in the following section in detail.

2.5.1.1. Types of gene variations

Gene variants can be categorized according to their underlying mechanism and their effect on the gene involved (Table 2)(35).

Table 2. Type of gene variations, their mechanisms, and their size range.

Variation	Rearrangement type	Size range
Single nucleotide changes	Single nucleotide polymorphisms, point mutations	1 bp
Small insertions/deletions	Binary insertion/deletion events of short sequences	1-50 bp
Short tandem repeats	Microsatellites and other simple repeats	1-500 bp
Fine-scale structural variations	Deletions, duplications, tandem repeats, inversions	50 bp to 5 kb
Retroelement insertions	Short interspersed elements, long interspersed elements, long terminal repeats, endogenous repeat viruses	300 bp to 10 kb
Intermediate-scale structural variation	Deletions, duplications, tandem repeats, inversions	5 kb to 50 kb
Large-scale structural variations	Deletions, duplications, large tandem repeats, inversions	50 kb to 5 Mb

2.5.1.1.1. Nucleotide substitutions

A single nucleotide substitution (or point mutation) can alter the translation in a triplet of bases in a coding sequence, resulting in the nonsynonymous replacement of one amino acid by another in the gene product. Because they change the coding strand of the gene to indicate a new amino acid, these mutations are referred to as missense mutations (33). Although not all missense mutations cause a change in the protein's function, the resulting protein may fail to function properly, be unstable and quickly degraded, or fail to locate through its intracellular position (36). Synonymous variants (also known as a silent substitution) are the changes that affect the DNA sequence without causing alteration in protein sequence (37).

Nonsense mutations are point variations in a DNA sequence that cause one of the three termination codons to replace the regular codon for an amino acid. Because mRNA translation stops when a termination codon is reached, a mutation that transforms a coding exon into a termination codon causes mRNA translation to stop halfway through the coding sequence. A mutation of this type will result in an aberrant protein product with extra amino acids at its carboxyl terminus, as well as a disruption of regulatory activities normally given by the 3′ untranslated region downstream of the normal stop codon (33,36). Premature termination codons in genes, on the other hand, rarely result in the shortened protein that could be expected. Cells have a process called nonsense-mediated decay (NMD) that recognizes and degrades mRNAs with premature termination codons. As a result, a nonsense mutation usually prevents the gene from being expressed (38).

A frameshift variant is a genetic variation induced by indels in a DNA sequence that is not divisible by three nucleotides. Because of the triplet nature of gene expression by codons, insertion or deletion can change the reading frame, resulting in a translation that is completely different from the original. The protein will be more changed if the deletion or insertion occurs earlier in the sequence (39). A frameshift variant differs from a single-nucleotide polymorphism, which occurs when a nucleotide is altered rather than inserted or deleted. In general, a frameshift variant causes the codons read following the mutation to code for different amino acids. The frameshift mutation will also change the first stop codon in the sequence ("UAA", "UGA", or "UAG"). The polypeptide produced could be exceptionally short or long, and it will almost certainly be ineffective (33,36).

Splice site mutations occur when nucleotide(s) are inserted, deleted, or substituted in the precise spot where splicing occurs during the conversion of precursor messenger RNA into mature messenger RNA. Initial RNA transcripts encounter a series of alterations before being processed into mature mRNAs (or final forms of noncoding RNAs), including transcription factor binding, 5′ capping, polyadenylation, and splicing. All these processes in RNA maturation are dependent on certain RNA evolutionary conserved sequences. Two types of splicing variations have been

identified in the case of splicing. For introns to be excised from unprocessed RNA and exons to be spliced together to generate mature RNA, specific nucleotide sequences near the exon-intron (5′ donor site) or intron-exon (3′ acceptor site) junctions must be present. Normal RNA splicing at either the splice donor or acceptor site is interfered with (and in some cases abolished) by mutations that disrupt these necessary nucleotides. Base substitutions, which do not impact the donor or acceptor site sequences themselves but instead form alternative donor or acceptor sites that compete with the usual sites during RNA processing, are the second type of splicing mutation. Inactivating a splice site usually results in the loss of a gene's function, or at the very least, all isoforms that use that site, but the exact biochemical mechanisms are difficult to predict. Exons are sometimes skipped; intronic sequences are sometimes maintained in mature mRNA; a neighbouring cryptic splice site is frequently employed. Cryptic splice sites are sequences within a primary transcript (in exons or introns) that resemble real splice sites but aren't close enough for the cell to recognize them as such. Within a cryptic site, a nucleotide variation may increase the similarity enough to turn it into a functional site. This will cause the transcript to be incorrectly processed. A sequence modification, on the other hand, may reduce the strength of a true splice site; thus, a neighbouring cryptic site is preferred (33,38)

2.5.1.1.2. Insertion and deletion variations

Variations caused by insertion or deletion (INDEL) of anywhere from a single base pair up to around 1000 bp, although larger indels have been found, make up the second class of polymorphism (40). A frameshift variant will occur in coding regions of the genome unless the length of an indel is a multiple of three. A point mutation is the contrary of an indel. A point variation is a type of substitution that changes one of the nucleotides without changing the overall number in the DNA. Indel inserts and deletes nucleotides from a sequence, whereas a point mutation replaces one of the nucleotides without changing the overall number in the DNA (41). An indel variant in the coding region of an mRNA causes a frameshift during translation, which might result in an improper (premature) stop codon in a different frame. In coding regions, Indels that are not multiples of three are rare, although they are very common in non-coding

regions (42). Each individual has between 192 and 280 frameshifting indels (43). Indels are believed to account for between 16% and 25% of all sequence polymorphisms in humans (44). In fact, indel frequency is far lower than that of single nucleotide polymorphisms (SNPs) in most known genomes, including humans, except for highly repetitive regions such as homopolymers and microsatellites.

Copy number variation (CNV) is a term used to describe a molecular process in which the number of times a sequence of the genome is repeated changes amongst individuals of the same species (45–47). Inversions and balanced translocations, as well as genomic imbalances (insertions and deletions), are examples of CNVs. The human genome's segmental replication architecture is crucial to comprehending structural diversity. Segmental duplications (also known as low copy repeats) are blocks of DNA 1–400 kb in length that occur at several locations across the genome and have a high level of sequence similarity (>90 percent) (48,49).

2.6. ACMG Guidelines and Variant Interpretation

In 2008, The American College of Medical Genetics and Genomics (ACMG), developed guidance for the interpretation of sequence variants (50). Since then, sequencing Technologies and our knowledge have evolved. According to improvements in the area, ACMG revised its standards and guidelines for the interpretation of variants. These guidelines broadly apply to genotyping, single genes, panels, exomes, and genomes, which are all common genetic tests used in clinical laboratories. To describe variations found in Mendelian disorders, this guideline suggests using standard terms such as 'pathogenic', 'likely pathogenic', 'uncertain significance', 'likely benign', and 'benign' with an evidence-based decision tree; i.e. it is a guideline to classify known disease-gene associations. Evidence of pathogenicity is rated as very strong, strong, moderate, and supporting (Figure 2). In exchange, weight criteria for benign evidence are stand-alone, strong, and supportive. The criteria are then combined using the scoring rules in Figure 2 to determine which of the five tiers to use (51,52).

	BENIGN		PATHOGENIC			
	Strong	Supporting	Supporting	Moderate	Strong	Very Strong
Population Data	MAF is too high for disorder BA1/BS1 OR observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalance in affecteds statistically increased over controls PS4	
Computational and Predictive Data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4 Missense in gene where only truncating cause disease BP1 Silent variant with non predicted splice impact BP7 In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein lenght changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional Data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path, missenses common PP2	Mutational hot-spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation Data	Nonsegregation with disease BS4		Cosegregation with disease in mutiple affected family members PP1			
De novo Data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity & maternity confirmed) PS2	
Allelic Data		Observed in trans with a dominant variant BP2 Observed in cis with a pathogenic variant BP2		For recessive disorders, detected in trans with a pathogenic variant PM3		
Other Database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5			
Other Data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly spesific for gene PP4			

Figure 2. Evidence framework of ACMG Guideline, 2015 (52).

This diagram arranges each of the criteria for a benign (left side) or pathogenic (right side) assertion by the type of evidence and the strength of the criteria. BS stands for benign strong; BP stands for benign supporting; FH stands for family history; LOF stands for loss of function; MAF stands for minor allele frequency; PM stands for pathogenic moderate; PP stands for pathogenic supporting; PS stands for pathogenic strong; PVS stands for pathogenic very strong.

2.7. Milestones of Sequencing Technologies

2.7.1. First generation sequencing

The development of Sanger's 'chain-termination' or dideoxy technique in 1977 was a critical advancement that transformed the progress of DNA sequencing technology forever (53). In time the method was called Sanger sequencing. The Sanger sequencing is a DNA sequencing method that is based on including chain-terminating dideoxynucleotides selectively by DNA polymerase. This sequencing method was developed by Frederick Sanger and his colleagues in 1977. Sanger sequencing is still the most used sequencing method, despite the widespread usage of NGS technologies

and Sanger sequencing is still used for shorter sequences and is the diagnostic gold standard for the co-segregation of variants. Sanger sequencing method requires a single strand DNA template, DNA polymerase, deoxynucleotriphosphates (dNTPs) and modified dideoxynucleotide triphosphates (ddNTPs) and DNA primers. All ddNTPs have fluorescent dye according to their nucleotides. The samples loaded on the polyacrylamide gel irradiate under UV light, and the sequence can be read. Also, capillary electrophoresis may be applied to read the sequence (Figure 3).

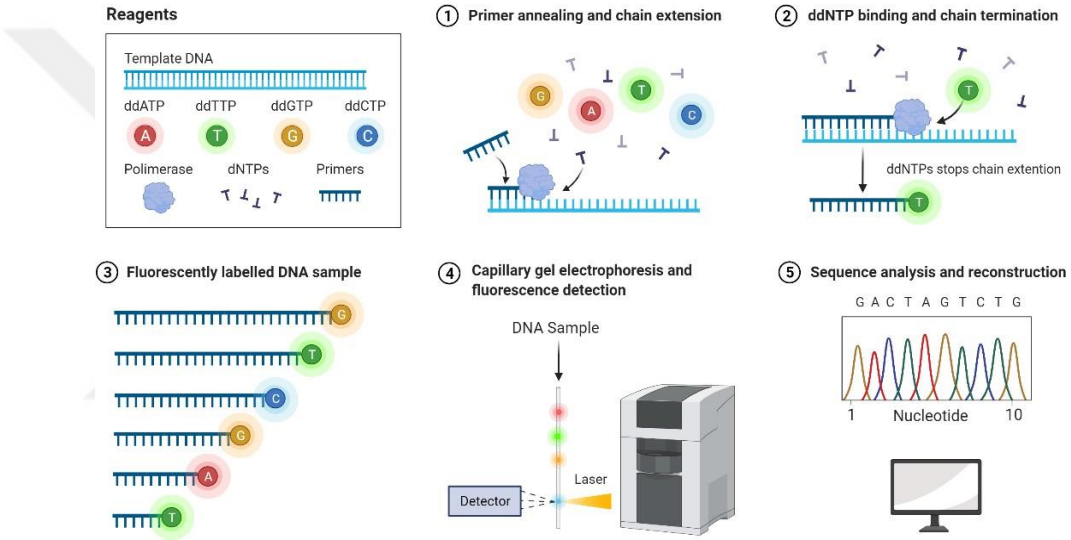


Figure 3. Sanger sequencing workflow.

(This figure was drawn with the BioRender application)

Sanger sequencing creates SCF chromatograms as an output in which each peak expresses a nucleotide (Figure 4). Nucleotide changes can be noticed with the reference genome comparison. The chromatogram view also allows to interpretation variants zygosities (Figure 4).

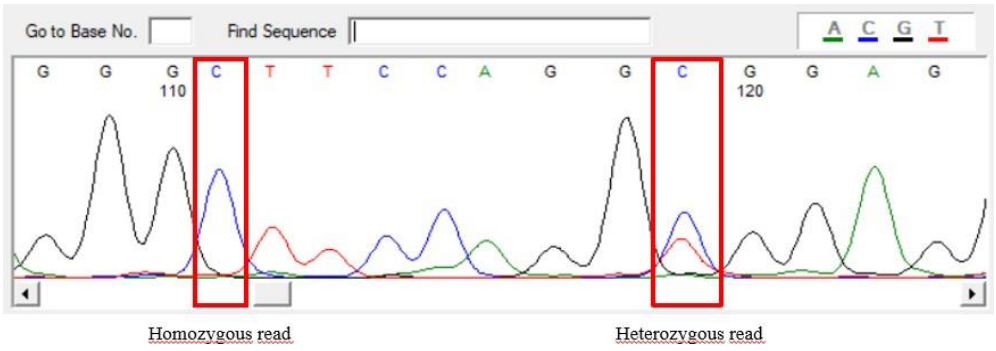


Figure 4. DNA sequence chromatogram format visualization.

The image of homozygous and heterozygous variants on the chromatogram are indicated.

2.7.2. Next-generation sequencing (NGS)

Massively parallel sequencing, also known as next-generation sequencing (NGS) or second-generation sequencing, is one of several high-throughput techniques for DNA sequencing that uses the concept of massively parallel processing (54). In recent years, several massively parallel sequencing technologies have become available, allowing for larger-scale genomic sequence generation (55).

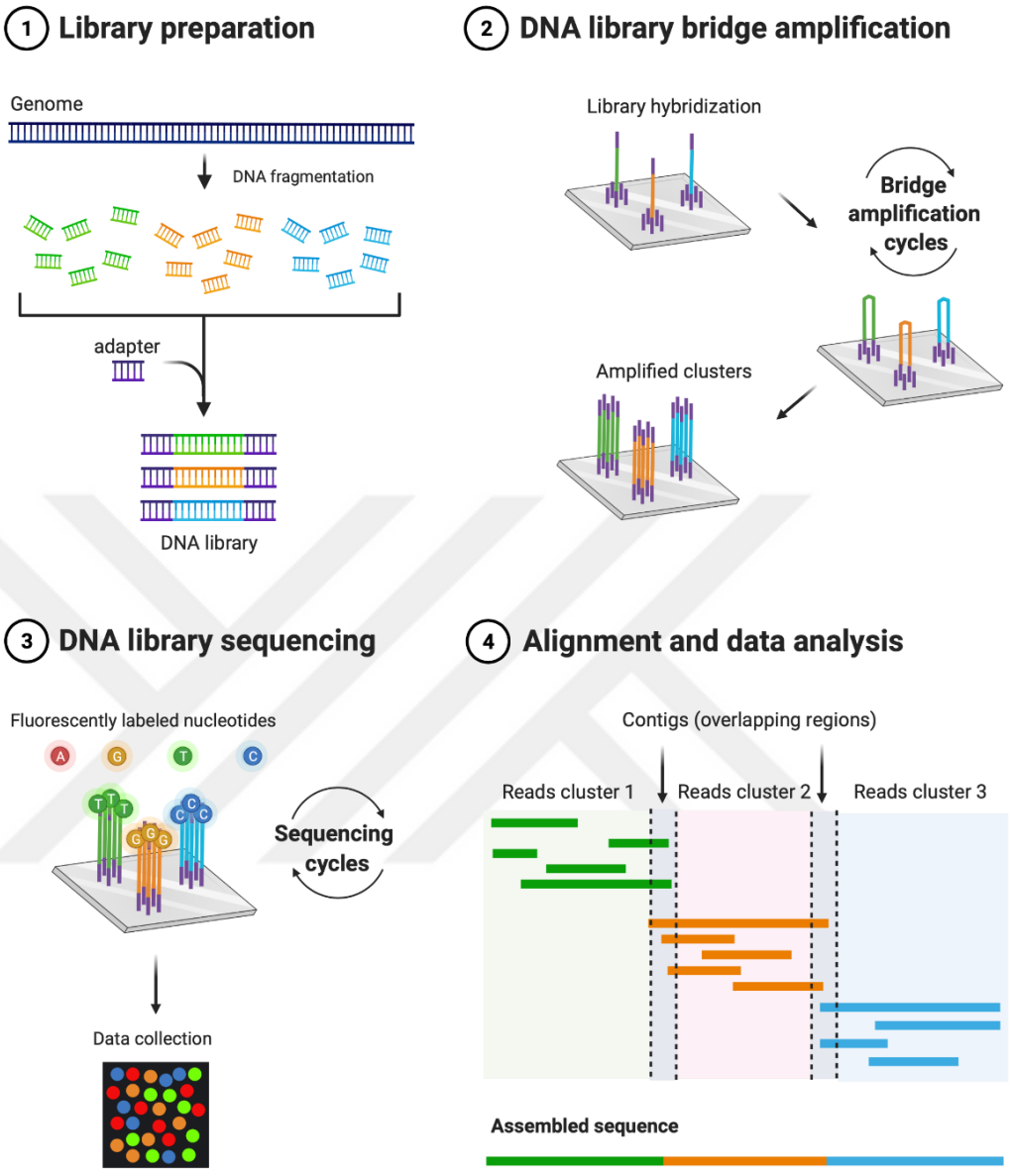


Figure 5. Methodology of the Illumina short-read sequencing (sequencing by synthesis).

1) The genome fragments and sequencing adapters ligate these randomly fragmented genomic DNAs. DNA libraries are created in this manner. 2) Attach single-stranded fragments to the inside surface of flow cell channels at random. To begin solid-phase bridge amplification, add unlabeled nucleotides and enzymes. On the solid-phase substrate, the enzyme integrates nucleotides to form double-stranded bridges. Single-stranded templates remain anchored to the substrate after denaturation. Each channel of the flow cell generates several million dense clusters of double-stranded DNA. 3) After adding four tagged reversible terminators,

primers, and DNA polymerase, the first sequencing cycle begins. The emitted fluorescence from each cluster is collected after laser activation, and the initial base is recognized. The incorporation of four tagged reversible terminators, primers, and DNA polymerase is repeated in the following cycle. The image is acquired as before after laser excitation, and the identity of the second base is recorded. One base at a time, the sequencing cycles are performed to identify the sequence of bases in a fragment. 4) Sequencing differences are found after aligning the data and comparing it to a reference (This figure is drawn with BioRender.) (56).

NGS can be applied to genetic materials such as DNA and RNA and provides specific sequencing of the whole genome, whole exome, and or a region of interest with the index primers and barcodes used to prepare the sequencing libraries. Also, NGS is involved in functional genomics, transcriptomics, oncology, evolutionary biology, and medicine. It has made great contributions to life sciences in many fields, such as diagnosis and discovery of new disease-related genes and classification and discovery of novel organisms. Despite the fact that next-generation sequencing (NGS) is a more recent sequencing approach, Sanger sequencing is still significant, depending on the sequencing target and the size of the region to be sequenced (Figure5) (53,57–59).

2.7.2.1. Targeted enrichment

NGS plays a leading role in much of today's genetic research. Depending on the study aim, it has different application types. Targeted enrichment is one of these methods and it is the most suitable and preferred method in terms of price/time cost today. Targeted enrichment is the sequencing of specific regions of the genome, and it can be shaped according to the aim of sequencing (57). Genes or genomic regions which are associated with complex traits can be performed with less data and less cost for pharmacogenetics and pathway analyses. The value of exome sequencing becomes even greater when it is assumed that such pathogenic mutations occur in exonic regions for cost and timing. There are basic standards to catch at targeted enrichment which causes different difficulties such as coverage of the targeted region and read depths, availability to re-analysis, ratios of in-target/out-target reads, and determination of necessary DNA amount,etc. If these standards can be provided, data can be analysed,

and it can be re-analysed later. There are two popular targeted sequencing approaches hybridization capture, and amplicon sequencing (Figure 6).

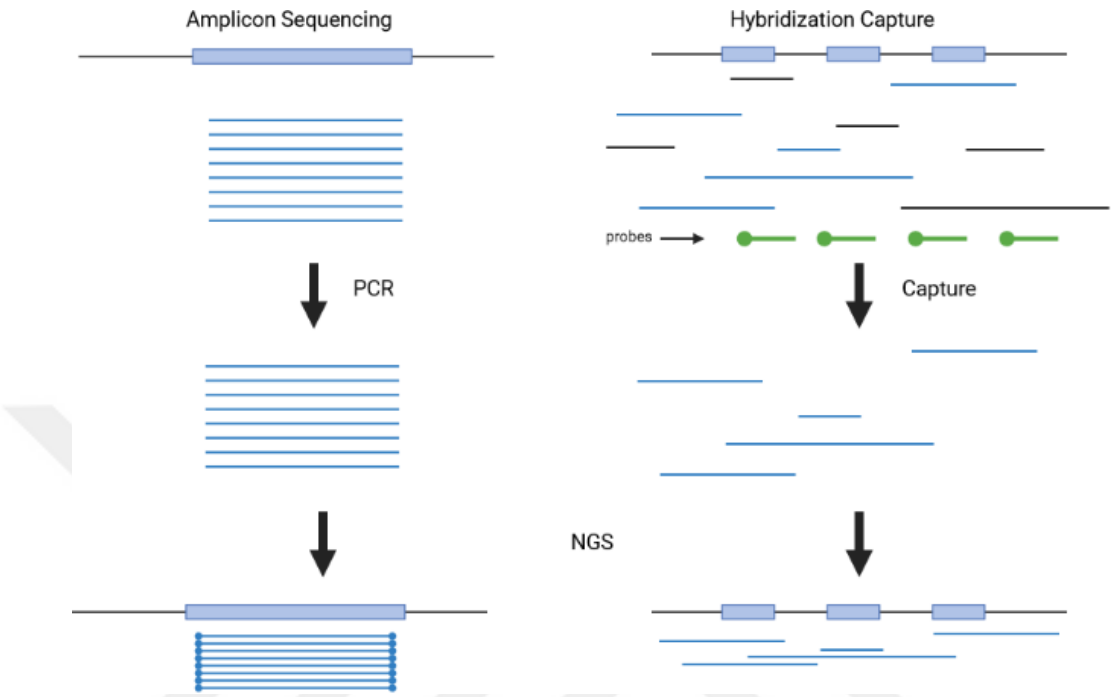


Figure 6. The differences between amplicon sequencing and hybridization capture.

The DNA products of a PCR are known as amplicons. Amplicons are created by PCR, pooled, and sequenced in NGS amplicon sequencing. Amplicon sequencing can discover variations at very low levels and frequencies since NGS-based targeted sequencing leads to very extensive coverage of a specific region of interest. The approach enables sample multiplexing, allowing hundreds of PCR fragment sequences to be determined at the same time (60). Target enrichment, also known as hybridization capture, is a method of targeted next-generation sequencing (other methods of targeted sequencing can include the use of amplicons or molecular inversion probes). DNA samples are transformed into sequencing libraries before hybridization capture (61).

NGS can sequence millions of fragments with hybridization capture. The best applications of hybridization capture are exome sequencing, gene discovery, oncology and genotyping, etc (60). Exome sequencing is a sequencing method that sequences exon regions and exon-intron boundaries of a genome. Today, exome sequencing is one of the important targeted enrichment methods in the diagnostics of monogenic diseases. Exons are regions that cover 1.9% of the entire genome and are involved in

protein-coding. It is known that the majority of genetic diseases defined today are caused by mutations in exons (61). According to Yang's 2014 article, whole-exome sequencing has around a 25% diagnostic rate (62). The diagnostic rate for WGS is 45.6%, according to Bertoli-Avella, 2021 article (63).

2.7.2.1.1. Whole exome sequencing

Whole exome sequencing is one of the applications of targeted enrichment designed to capture coding regions on the genome. The WES application aims to identify variants that are directly related to protein translation. In ClinVar, more than 90% of pathogenic variants are located in coding DNA sequences (60).

WES has several advantages and disadvantages compared to whole-genome sequencing (WGS). WGS remains expensive and requires high-performance computing to analyze. WES is financially more affordable, creates a lower amount of data than WGS, which makes it easy to analyze, and reduces the risks of identifying incidental findings that make it more ethical to use it (61). The most important disadvantage of WES is not applicable for identifying structural variants like chromosomal rearrangements, long insertions, deletions, inversions, and copy number variants more accurately.

WES analysis has several steps that are split into bioinformatics steps and filtering steps. Bioinformatics steps were explained in the “FASTQ to VCF” and “Annotation” sections. The filtering step is to filter the variants that are prone to pathogenicity from the annotated VCF by using the information obtained as a result of annotation from the list of variants obtained. The aim of variant filtering is to find the variant or variants that may cause the disease of the studied individual.

2.7.2.2. Whole genome sequencing

WGS is a method for analyzing complete genomes. Identification of hereditary diseases, describing the mutations that drive cancer progression, and tracking disease

outbreaks have all benefited from genomic knowledge. Whole-genome sequencing has become a valuable tool for genomics research due to rapidly falling sequencing prices and the capacity to create massive amounts of data using today's sequencers (64).

NGS devices with WGS capabilities can sequence an entire genome in less than 30 hours. A single entire genome's DNA is fragmented and assembled into a single sequencing library, which is then sequenced in one run. In a normal human genome, there are around 2–3 million SNPs and tiny INDELs, with about 15–20,000 of them occurring in the coding area (65).

WGS can cover up to 98 percent of the human genome, while WES can cover almost 95 percent of coding areas but only 1–2% of the genome. WES has a lower cost per sample than WGS, a better depth of coverage in target locations, fewer storage requirements, as well as quicker data analysis.

2.7.3. Third generation sequencing

Third-generation sequencing (TGS) is also known as long-read sequencing. The major difference between short-read and long-read sequencing is the fragment sizes of each read. NGS reads have approximately 100 to 300 bp reads for each fragment, while TGS sequencers can read approximately 10 kb bp reads for each fragment (66). Even with the application of state-of-the-art bioinformatic techniques, structural variations (SVs), repetitive elements, excessive guanine-cytosine (GC) content, or sequences with many homologous elements in the genome are challenging to define using NGS. Because of the limits of NGS-based human disease research, scientists are looking for new ways to improve diagnostic accuracy and speed up detection times in genetic diseases (67). TGS is a single molecule and real-time sequencing method developed by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) in 2011 (68). Single-molecule real-time (SMRT) technology is used on the PacBio platform. There is no need for PCR in the DNA library preparation because a closed and circular ssDNA template may be duplicated mechanically. The fluorescence signals are activated by a laser during the sequencing process as soon as a tagged dNTP is

integrated into DNA (69). In time also, Illumina announced their new long-read sequencers, which are named “Infinity Long Reads” (70).

2.8. NGS Analysis

The data obtained from the sequencing device passes through the pre-process, process, and analysis steps. High computing power is required to realize these steps. However, in order for the NGS data to be large and archived these data, these data must go through a group of pre-processes. In the pre-processing step, the data obtained from the sequencing device is made ready for it to go through a group of bioinformatics stages. In terms of Illumina devices, the rawest data obtained from the sequencing device are Binary Base Call (BCL) format files. BCL formatted files are very large in size compared to other raw data. For this reason, BCL files are compressed by converting to FASTQ formats. FASTQ formatted files are much smaller data compared to BCL. In this way, they can be stored and transferred more easily (71).

In the process step, the FASTQ format file is made ready for analysis by going through the stages of alignment, sorting, removing duplicates, recalibrating of bases, and variant calling.

VCF files obtained as a result of variant calls are filtered by annotating the data collected from the databases in the analysis step.

2.8.1. Data types

2.8.1.1. FASTQ format

FASTQ is a text format to keep biological sequence and also its quality scores (Figure 7). Every letter which encodes sequence and quality score, is encoded with American Standard Code for Information Interchange (ASCII).

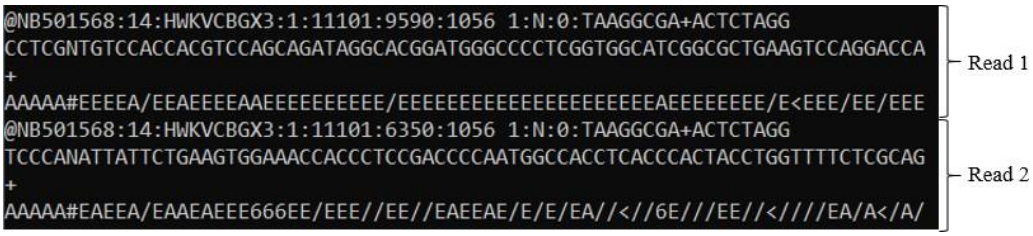


Figure 7. An example of a FASTQ file with multiple reads.

It was developed by the Wellcome Trust Sanger Institute to store FASTA formatted sequences and its quality information. FASTQ format has four different line types (Table 3). Each four lines describes one read (72).

Table 3. Line types of FASTQ format.

Line 1	@title and description
Line 2	sequence
Line 3	+
Line 4	quality

2.8.1.2. SAM and BAM format

The binary alignment map (BAM) format presents aligned raw data for genomic data. BAM is the compressed format of Sequence Alignment Map (SAM). SAM is a text-based format that keeps aligned sequence information along with the quality metrics. These formats support both short and long reads and more than one sequencing platfooutpututs. Each alignment has a section and a header. Headers start with "@" symbol. These symbols help recognize alignments one by one SAM files keep too much space for archiving which causes more storage needs. For improving the performance and lowering the storage needs, SAM files can be compressed to BAM. SAM/BAM files are not sorted. For advanced processes (especially indexing) data should be sorted by its coordinate and chromosome (73).

2.8.1.3. Variant calling format

Variant calling format (VCF) is a text based and tab-delimited data format that contains SNPs, INDELs, structural variants, and their information. A standard VCF file has eight different columns. The first column (CHROM) represents the chromosome where the variant is localized. The second column (POS) represents the starting position of the variant on the chromosome. The third column (ID) represents the unique identifier of the variant if it exists. Forth (REF) and Fifth (ALT) columns show the nucleotide changes. The REF column has the reference nucleotide at that position; the ALT column shows the changed nucleotide. The sixth column (QUAL) has a phred-scaled quality score for each variant read. QUAL scores are integer numbers, and high scores mean high confidence that the variant is true. The FILTER column, which is seventh, contains the information of whether the variant is acceptable to analyse for your research or not. FILTER column requirements should be given while variant calling step (QUAL score filter, read-depth filter etc.). If a variant passes every filter, the FILTER column indicates "PASS." The eighth column (INFO) represents the sequencing information of the variant. The INFO column may indicate more than 5 different pieces of information like ancestral allele, allele count, allele frequency for each changed nucleotide etc (74).

2.8.2. Processing of NGS data

Multiple format changes occur before the NGS data becomes analysable. These format changes are necessary processes to make the analysis more accurate and to make the data more workable.

2.8.2.1. FASTQ to VCF

There is more than one step in the process of NGS data. These processes are called a "pipeline" together. Almost every pipeline has differences according to its purpose of usage, but there are common steps that have been shared by every pipeline (Figure 8).

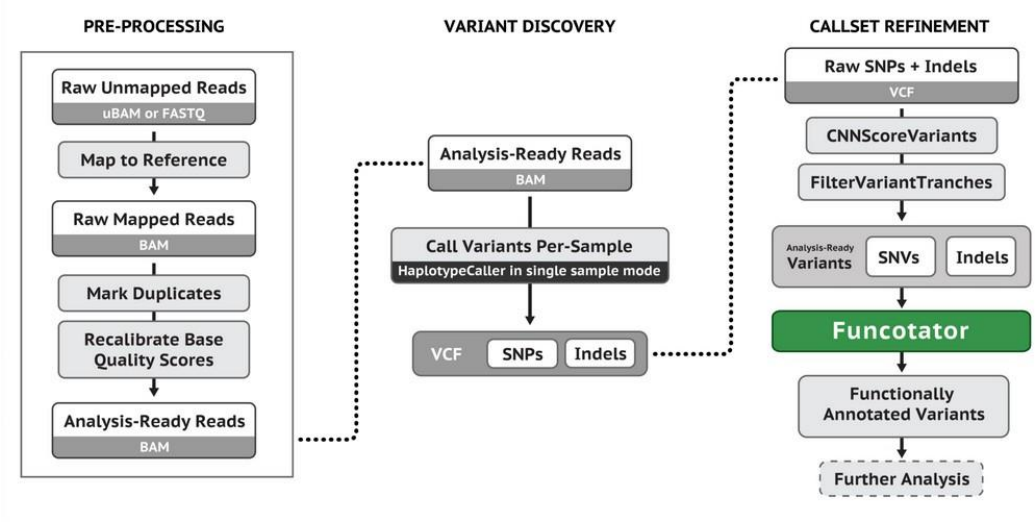


Figure 8. GATK Best Practices pipeline for germline genomic NGS analysis (75,76).

All genomic pipelines have the mapping step. FASTQ files contain millions of fragment sequences without an order. Mapping is an algorithm-based step that localizes every fragment to proper places on the reference genome and creates a SAM file. SAM files are too big to archive and slow down the process (a WES SAM is approx. 40 GB). Also, SAM files contain unnecessary information for further steps and keep a lot of space. SAM to BAM conversation is clearing that unnecessary information and compressing the SAM file. A WES BAM file is around 6-10 GB.

BAM files must be sorted to their coordination for indexing the data. It's a necessary process which lowers the time of further steps.

Marking duplicates is a process of locating and tagging duplicate reads in BAM files. Duplicate reads originate from a single fragment of DNA, which arises during sample preparation steps. Marking and removing these duplicates provides eliminating false variants before variant calling.

Marked and removed duplicates BAM may still consist of false-positive mismatches. The base recalibration step provides a table based on specified covariates

by using database VCF files (e.g., dbSNP, gnomAD, etc.). With this step, poor base quality mismatches are noticed with statistical methods.

The variant calling step provides the VCF with mismatches, which are called variants now, VCFs contain the location of the variants and the nucleotide changes but not detailed information about them. In order to obtain this detailed information and make an easier analysis, the annotation step should be applied (76).

2.8.2.2. Annotation

A variant may have too much information about itself in the databases, and it will take enormous time to research every database on the internet. Annotation is a process of collecting needed functional elements and information about a variant. There is no general rule for annotation. Every researcher creates an annotation list according to their need. Annotation for genomic studies may consist of allele frequency information, in-silico pathogenicity scores, information from disease databases, ontology information, the clinical significance of a variant if it exists, etc. The annotation also provides variant classification information about their function and location on the genome, and their Human Genome Variation Society (HGVS) nomenclature, which is needed for filtering and further processes like literature review (77).

2.9. Annotation Databases

The human genome contains only four bases, but each letter has different functions depending on the region it is in, the number of repetitions it contains, and the combination of certain letters (33). That is why every nucleotide of every location has a unique feature. Genomic databases are collecting, storing, and publishing these features, but every feature is a different research area. Almost for every research area, there are one or more databases that exist.

2.9.1. Allele frequency databases

All diploid people in a population carry two alleles for each gene. Allele frequency (AF) is the ratio of alleles in a population to the total. Allele frequency databases use thousands of exome and genome sequencing data from different regions (populations) all around the World. All found variants are counted and calculated in their allele frequencies in their population and total. The Genome Aggregation Database (gnomAD) and 1000 Genome Project (1000G) are the most used databases for general research (78,79). More data means more accurate allele frequency. That is why gnomAD and 1000G are the most used databases for allele frequency. It's important that allele frequencies show differences between populations. There are so many variants that are inherited by only one population. This makes regional AF information more important. Turkey does not have any genome project output right now. Greater Middle East Variome Project (GME) consists of data from Turkey; also, Iranome, a genome Project from Iran which has a Turkmen population, may be preferable for studies from Turkey (80,81). The gnomAD allele frequency database also provides Gene Constrained information. Gene constrained is a metric for calculating a gene's tolerance for nonsynonymous, synonymous, and LoF variants. The observed/expected (oe) number of loss-of-function variations in that gene is used to calculate the constraint score in gnomAD. Expected numbers are calculated using a mutational model that considers sequencing context, coverage, and methylation. The observed/expected (oe) ratio is a continuous assessment of a gene's tolerance to a certain type of variation. A gene with a low oe value is subjected to more selection for that class of variation than one with a higher value. The precision of the oe values varies a lot from one gene to the next since counts are dependent on gene size and sample size. As a result, the 90 percent confidence interval (CI) for each of the oe values in addition to the oe value was presented. The 90 percent confidence interval must be considered when determining how constrained a gene is (79).

2.9.2. In silico pathogenicity prediction databases

The human genome has pathogenic and non-pathogenic variants in its content. Some of them have already been related to diseases, and there is still no information about many of them. Classifying a variant as pathogenic requires clinical trials, segregation studies, and more in vitro studies. This is almost impossible to do that classification for each variant seen in the human genome. Variants may affect protein with amino acid change, splicing, and start-stop site changes. These changes may cause disease or nothing. In silico pathogenicity prediction tools are developed because of these needs. Every tool has a specific way of predicting a variant (e.g., polar-charged amino acid changes, conservation, function, etc.) (82). In silico, pathogenicity prediction scores might be used by accessing their websites or with annotation to NGS data. There is no standard between scores. Almost each tool uses its scoring scale, and each score has a deleterious threshold given by developers. dbNSFP is a database that collects most using in-silico pathogenicity prediction tools together and ranks them between 0 and 1. The closer the score is to 1, the more likely it is to be pathogenic, and the closer to 0, the more likely it is to be benign, according to dbNSFP ranking (83).

2.9.3. The Gene Ontology

The Gene Ontology is a computer description of the current scientific understanding of the roles of genes across a wide range of creatures, from humans to bacteria. Understanding how specific genes contribute to an organism's biology at the molecular, cellular, and organismal levels is one of the fundamental aims of biomedical research. Furthermore, experimental data from an organism can frequently reveal similarities with other organisms, particularly if the organisms have related genes inherited from a common ancestor. Gene Ontology (GO) cooperated on a uniform classification framework for gene function in 1998 as a collaboration, based on the work of academics investigating the genomes of three model species (*Drosophila melanogaster, Mus musculus and Saccharomyces cerevisiae*). GO provides comparable definitions of homologous gene and protein sequences throughout the evolutionary range in a flexible and dynamic manner (84).

3. MATERIALS AND METHODS

3.1. Patients

In the study, WES (n=22) and WGS (n=3) data of 25 patients were enrolled in the study. There were 21 male and 4 female patients with a median age of 19 (min. 1- max. 60) has the approval of the ethics committee (Ethics committee decision no. ATADEK-2020-10/7). In this study, the data of 25 patients in total were used with the approval of the ethics committee (Ethics committee decision no. ATADEK-2020-10/7). The demographic and clinical features of the patients were summarized in Figure 9 and Table 4.

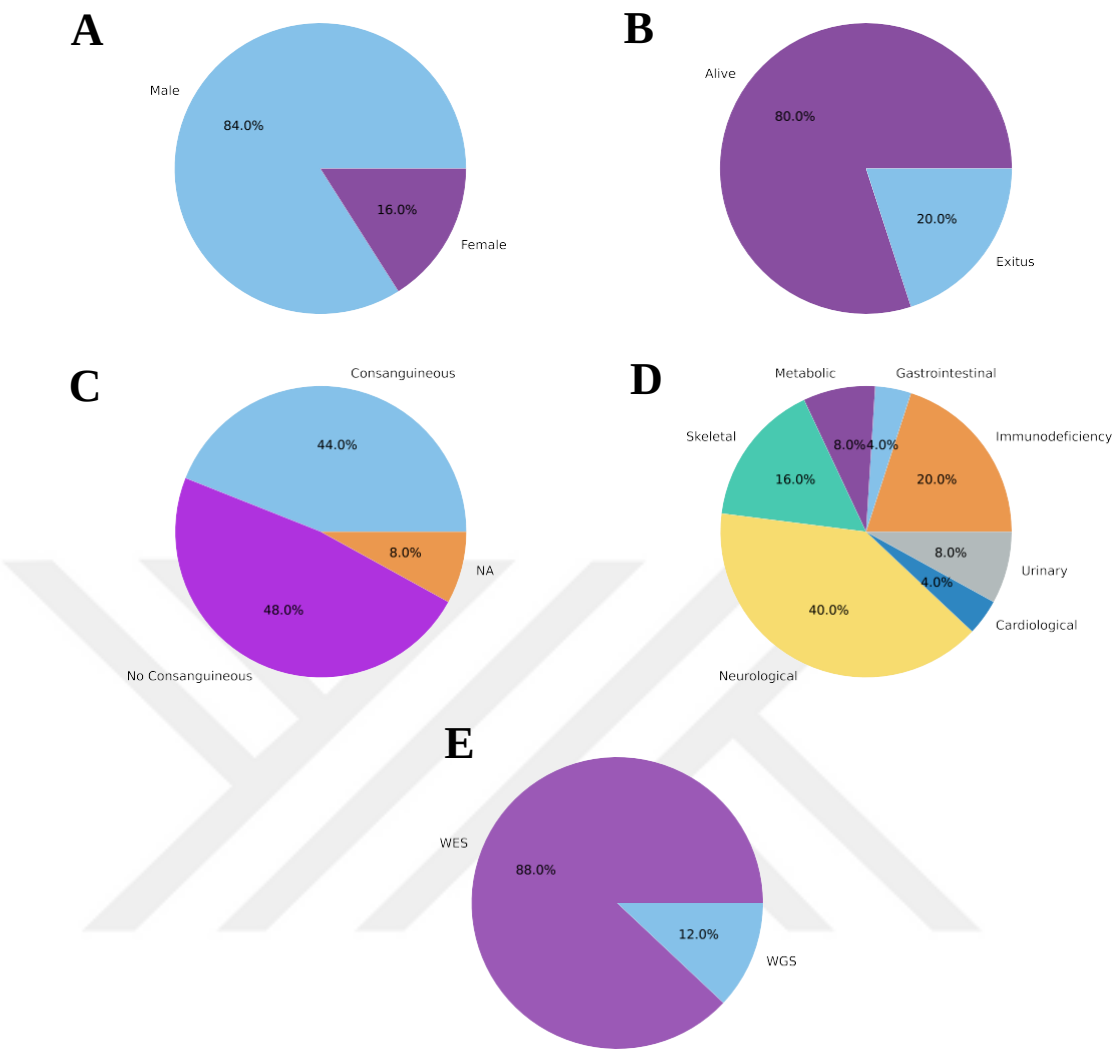


Figure 9. Distribution of the patients.

A) Gender distribution of the patients. B) Status distribution of the patients. C) Consanguineous distribution of the patients. D) Phenotype distribution of the patients. E) Data type distribution of the patients.

Table 4. Demographic and clinical features of the patients enrolled in the study.

Patient ID	Date of Birth	Sex	Patient's status	Phenotype information	Cons.
WES001	25.03.2005	M	Exitus	Primary immundeficiency	Yes
WES002	19.03.2010	M	Alive	Inflammatory Bowel Disease, Ulcerative Collitis	No
WES003	9.12.1962	F	Alive	Hypertension with Hypokalemia	NA
WES004	Not provided	M	Exitus	Primary immundeficiency	Yes
WES005	19.03.2007	M	Alive	Osteochondrodysplasia	NA
WES006	24.06.2017	M	Alive	Neurodevelopmental delay	No
WES007	16.07.2018	M	Alive	Spondylocostal dysostosis	Yes
WES008	Not provided	M	Exitus	Primary immundeficiency	Yes
WES009	28.04.2018	M	Alive	Seizures	No
WES010	25.12.1979	M	Alive	Cardiomyopathy	Yes
WES011	15.06.2004	M	Alive	Sensorineural hearing impairment	No
WES012	15.05.2000	F	Alive	Combined immuno deficiency	Yes
WES013	20.02.1991	M	Alive	Primary immundeficiency	No
WES014	4.08.2010	M	Alive	Poor fine motor coordination	No
WES015	10.10.2016	M	Alive	Agnesia of corpus callosum	No

Table 4. Demographic and clinical features of the patients enrolled in the study (continue).

Patient ID	Date of Birth	Sex	Patient's status	Phenotype information	Cons.
WES017	22.01.2018	M	Exitus	Global developmental delay	Yes
WES018	14.07.2020	M	Alive	Microcephaly, thin corpus callosum	No
WES019	16.09.2017	M	Alive	Delayed motor milestones	Yes
WES020	10.07.2018	M	Alive	Epiphyseal dysplasia	Yes
WES021	4.03.2013	M	Alive	Absence of expected normal physiological development	No
WES022	14.06.1985	M	Alive	Focal segmental glomerulosclerosis	Yes
WGS001	15.12.2014	M	Alive	Nephrocalcinosis	Yes
WGS002	17.04.2020	F	Exitus	Encephalopathy and Spasticity	No
WGS003	23.08.2008	F	Alive	Delayed gross motor development	No

3.2. Collection of the Raw WES and WGS Data

Raw data were collected from five different centers. There were 23 FASTQ and 2 BAM files available for analysis. Clinical information of the patients were collected with their consent. There were both WES and WGS data from undiagnosed patients (negative results) or patients with clinically irrelevant (does not match with the phenotype) results. The data transfer and delivery was performed by USB sticks or portable hard drives. Collected data were stored in a portable hard drive and in the local servers of Acıbadem University, Rare Disease Application and Research Center (ACURARE) with a backup.

3.3. Genome Analysis Pipeline-ACUGEN

The re-analysis was carried out using ACUGEN, an in-house bioinformatics pipeline built by the ACURARE team (https://github.com/sakyoney/acugen). ACUGEN is written in the Python programming language. It includes four SNP-INDEL variant calling tools: Freebayes, BCFtools, DeepVariant, HaplotypeCaller, and Parliament2 for SV calling (85–89), which includes four alternative SV calling tools. ACUGEN also provides four separate VCF files as output to improve the accuracy of genuine variant calling. Individual VCF files from different variation callers, as well as a consolidated VCF file, are included. FASTQC, Burrows-Wheel Alignment (BWA), Samtools, Picard, Genome Analysis Toolkit (GATK), Freebayes, BCFtools, DeepVariant, Parliament2, and mosdepth are among the bioinformatics tools included in it (Figure 10). ACUGEN's speed is one of its most noteworthy attributes. Tasks that require less processing power can be completed simultaneously thanks to the additional parallelism function, which allows many phases to occur at the same time. These tools, which require fewer threads, save time by making use of the computer's idle threads. ACUGEN is parallelized using the divide-and-conquer algorithm, which ensures that low-thread and memory-intensive steps are divided among all threads and memories. Divide and conquer is an algorithm design approach in computer science. A divide-and-conquer algorithm breaks down a problem into two or more sub-problems of the same or related type until they are simple enough to be

solved directly (90). The sub-problems' solutions are then integrated to produce a solution to the original problem. Many efficient algorithms are based on the divide-and-conquer technique, including sorting, multiplying large numbers, discovering the closest pair of points, syntactic analysis, and computing the discrete Fourier transform (FFT). Divide-and-conquer algorithms are well-suited to multi-processor machines, particularly shared-memory systems because data transfer between processors does not need to be planned ahead of time because different sub-problems can be handled by different processors (91).

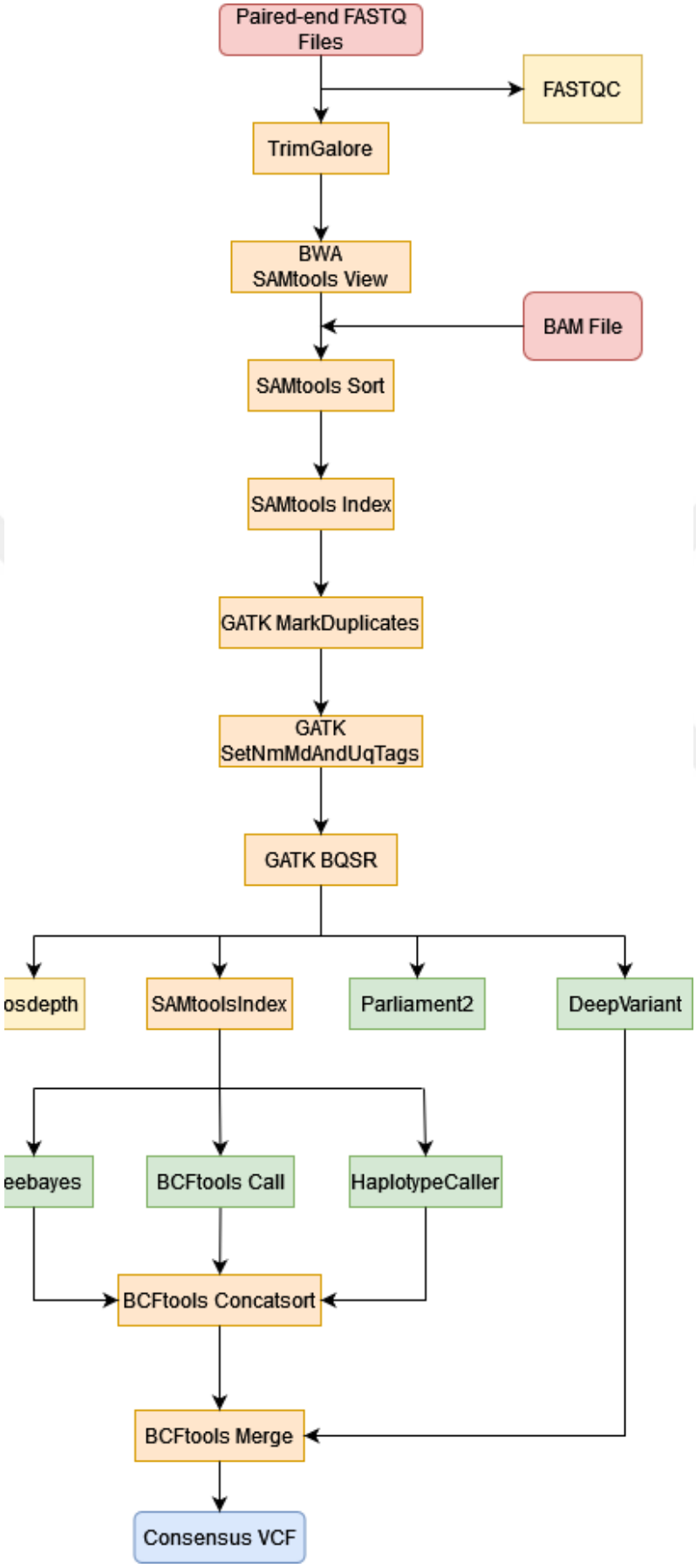

Paired-end FASTQ Files
FASTQC
TrimGalore
BWA
SAMtools View
BAM File
SAMtools Sort
SAMtools Index
GATK MarkDuplicates
GATK
SetNmMdAndUqTags
GATK BQSR
osdepth
SAMtoolsIndex
Parliament2
DeepVariant
eebayes
BCFtools Call
HaplotypeCaller
BCFtools Concatsort
BCFtools Merge
Consensus VCF

Figure 10. Flowchart of the ACUGEN pipeline.

Red cells show the input data, yellow cells show the quality control steps, orange cells show pre-process and mid-steps, and green cells show the variant calling steps.

3.3.1. Tools in ACUGEN

FASTQC

FASTQC is quality control for raw high throughput sequencing data. FASTQ, SAM, and BAM files may be imported to FASTQC. It provides an HTML-based report for each data with a set of analyses like base sequence quality, sequence quality scores, etc. ACUGEN uses the FASTQC tool for quality control of the raw sequencing data. In this study "per base sequence quality" was specifically checked for the data to be accepted for analysis (Figure 11). The number of base locations binned together is determined by the read length; for example, with 150bp reads, the plot will provide aggregate statistics for 5bp windows. Longer reads will have larger windows, while shorter reads will have smaller windows. The mean quality score at each base position/window is represented by the blue line. Illumina has created a primer on sequencing quality scores. The median quality score at that position/window is represented by the red line within each yellow box. The inner-quartile range for the 25th to 75th percentile is represented by the yellow box. The 10th and 90th percentile scores are shown by the top and lower whiskers, respectively (92).

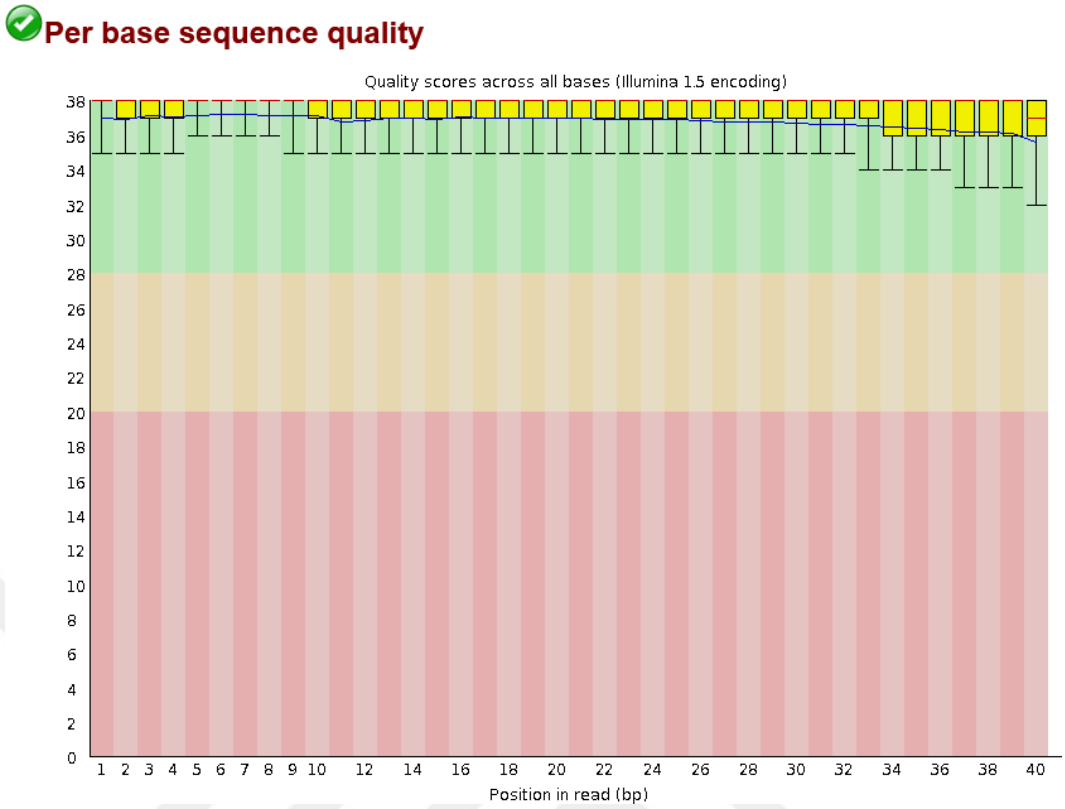


Figure 11. Per base sequence quality report of an Illumina FASTQ file of FASTQC.

Per Base Sequence Quality is a box-and-whisker graphic depicting aggregated quality score information along with all reads in the file at each point. The X-axis is not consistent; it begins with bases 1-10 being reported individually, then bins bases over a window of a specific number of places wide.

Burrows-Wheeler Aligner (BWA)

BWA is a package for mapping sequences against a large reference genome. It consists of three algorithms for different mapping lengths of sequences. BWA-MEM algorithm is the most used algorithm of BWA. BWA-MEM is created for sequences that are longer than 100 bp; also the most updated algorithm of the BWA, so according to creators, it's the most accurate and faster algorithm of all. BWA uses FASTA or FASTQ files as input and uses SAM format files as output (93). ACUGEN uses the BWA-MEM algorithm because all of the data contains longer than 100 bp fragments each.

Samtools

Samtools is a tool for manipulating aligned sequencing data like SAM, BAM, and CRAM. Samtools is one of the core tools for ACUGEN, and it uses three options. These options are “view” for SAM to BAM conversation, “sort” for sorting data according to their coordinates, and “index” for indexing sorted BAM files (73).

Picard

Picard is a Java-based tool for manipulating high throughput sequencing data such as SAM, BAM, and VCF. Raw aligned sequence data contains short fragments of polymerase chain reaction (PCR) and library preparation primer duplicates. ACUGEN uses Picard for marking and removing PCR and library preparation duplicates in reads for decreasing false-positive nucleotide variants in VCFs (94).

Genome Analysis Toolkit (GATK)

GATK is a Java-based tool for variant discovery in high-throughput sequencing data, developed by Broad Institute. It consists of more than 50 flags in itself and, the ACUGEN uses BaseRecalibrator for generating a recalibration table of read groups, reported quality scores from databases, machine cycle an,d nucleotide context. This provides calculation of the probability of error of poor base quality of reference mismatches. Also, ACUGEN uses ApplyBQSR for applying recalibration tables to our data, and HaplotypeCaller for variant calling (95).

BCFtools

BCFtools is a tool for manipulating VCF files. ACUGEN uses BCFtools’ “mpileup” and “call” arguments for variant calling. The mpileup argument generates a VCF file containing genotype likelihoods for an alignment. In ACUGEN, following the mpileup argument, the call argument is executed. The call argument calls

nucleotide changes in the mpileup output and generates a new individual VCF file (86).

Freebayes

Freebayes is another tool for variant calling in the ACUGEN pipeline. Freebayes' variant calling is based on Bayesian statistics which models multi-allelic loci sets of individuals with non-uniform copy number (85). Freebayes creates one of the individual VCFs in ACUGEN.

DeepVariant

DeepVariant is a deep learning-based variant caller that receives aligned reads (in BAM or CRAM format), generates pileup image tensors from them, uses a convolutional neural network to classify each tensor, and finally delivers the findings in a standard VCF or gVCF file (96).

3.4. Benchmark of ACUGEN

We performed a benchmark study to test accuracy and efficiency of ACUGEN. The benchmark study was conducted by using the data in the 1000 genomes database coded "NA12878", which is accepted as gold quality genome (97). A tool called hap.py was used to compare the gold quality VCF created by GATK best practice pipeline and a VCF data created by ACUGEN. The results of this tool compare true positive, false positive, and false negative variants between two VCFs (98). Hap.py outputs the total number of variants, true positive, false positive, false negative, non-assessed calls, transitions and transversions of SNP and INDELs for each VCF (Table 5).

Table 5. Contingency table describing the matches and mismatches situations between Truth data and Pipeline output.

The term of "ref" stands for reference allele, "alt1" stands for alternative allele 1 and "alt2" stands for alternative allele 2 which is divided into two for the possibility of seeing different variants on the location. TP; true positive, FN; false negative, FP; false positive, FP.AL; false positive allele mismatch, UNK; unknown.

		Truth		
	Genotype	ref/ref	ref/alt1	Outside BED
Pipeline	ref/ref	-	FN	-
	ref/alt	FP	TP	UNK
	ref/alt2	-	FP.AL	-

3.5. Annotation and Filtering Process

Annotation is one of the most important steps for molecular diagnosis of undiagnosed and rare diseases. The databases which are used in genomics, are updating frequently. If an analysis workflow stands out of date, this may result patients to remain undiagnosed.

In this study for the annotation step a free variant annotation and filtering software named VarAFT (Aix Marseille University, France) was used (99). VarAFT provides a very useful and user-friendly interface and different filtering approaches according to the patient's features (Figure 12). Also, it provides annotation very easily without requiring a script. VarAFT's annotation is done by Annovar. All annotation files were provided from Annovar's webpage (100,101).

Figure 12. The user interface of variant annotation and filtering tool VarAFT.

In the filtering step, unlike the other steps, a standard method was not followed. Different filtering priorities were applied according to the epicrisis of each patient like consanguinity situation, different gene lists, inheritance models etc. The filtering steps of this study are shown in Figure 13.

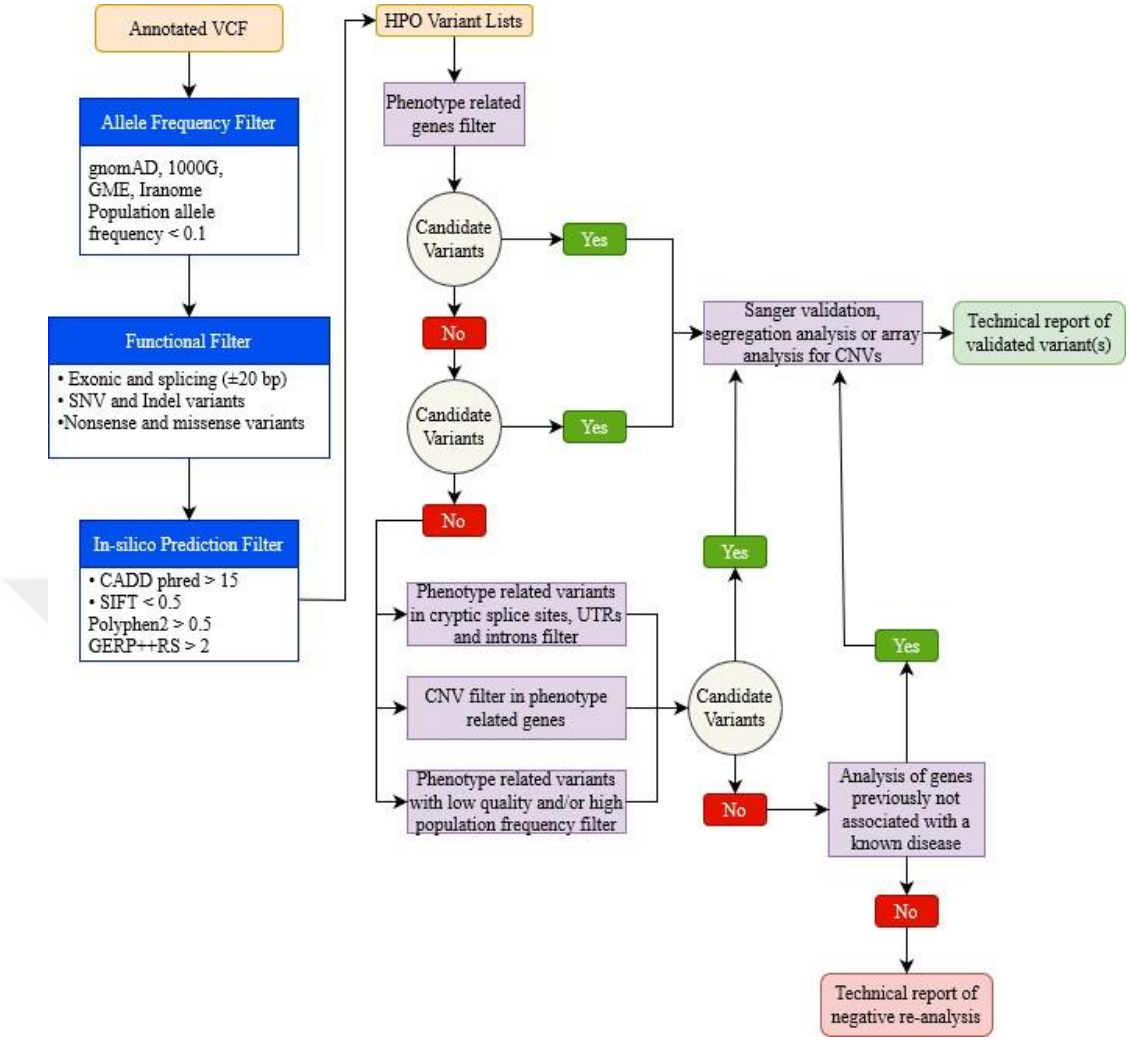


Figure 13. Representation of the filtering method used in this thesis on the flowchart.

3.6. Copy Number Variation Analysis of WES Data

Copy number variation (CNV) analysis was performed with Conifer v0.2.2 (102). CoNIFER counts how many sequencing reads align to exons before calculating a normalized RPKM value (103). Each RPKM value is standardized by the population's median and standard deviation. The SVD transformation is used to reduce systematic bias using the Z-RPKM values. After smoothing the final SVD-ZRPKM signal, a threshold method is used to find the duplication/deletion breakpoints. The calculation step of RPKMs needs a file called probes. Probes file is a tab-delimited text file that contains chromosome, start, stop, and gene name targets for exon capture. Publishers of Conifer suggest using a BED file of the used capture kit for WES, but, the capture

kit bed files used for sequencing were not used. In our analysis, we used the USCS Table Browser's enlarged bed file (which includes splice sites and UTR regions) (104). The control RPKMs were created from four healthy individual WES data. For discovery of deletions and duplications, SVD-ZRPKM threshold values were set -1.5 to 1.5.

3.7. Structural Variant Analysis of WGS Data

The calls of the structural variants for WGS were analysed by using the parliament2 tool, which is embedded in the ACUGEN pipeline. In the output obtained from this tool, the results of four different structural variant callers are collected in a common VCF. For the annotation of these variants, the structural variant annotation tool called AnnotSV was used (105). AnnotSV provides a lot of information, such as population allele frequency of structural variants found in VCF, genes affected, and previously identified diseases associated with it. Only the genes that are phenotype-associated were analysed by this tool.

3.8. Validation and Segregation of Candidate Variants

Following the filtering steps the variants that are considered as candidates were presented to a variant evaluation board consisting of responsible physician, paediatrician, medical geneticists, molecular geneticists and bioinformatician. The board evaluates the candidate variants with the phenotypes and the ones that are found relevant are validated by Sanger sequencing and the family segregation analysis was performed, as suggested by ACMG guideline (Evidence PS2). For validation and family segregation analyses peripheral blood samples are collected from the patient and family members (affected and unaffected) and genomic DNA is isolated.

3.8.1. DNA isolation, PCR and direct sequencing

A total of 5-10ml of peripheral blood were collected from the index case, parents, siblings and any other family member who was affected, if possible. DNA isolation

was performed with Genemark Genomic DNA Purification Kit (Genemark, Taiwan) as described previously by the manufacturer. All samples were archived in the Biobank Unit of Acibadem University. DNA quality and quantity were measured by Nanodrop One (Thermo Fisher, USA).

The variant-specific primer pairs were designed by Primer3plus according to their melting temperature and GC content ratio (106). Self-dimer status of primer pairs and hairpin formation status were checked by the OligoAnalyzer tool of IDT (107). Prior synthesis and amplification, an in-slico PCR was performed by UCSC In-Silico PCR tool (108). Following PCR optimizations index case and the family members, when available, were amplified (Table 6) and specific PCR products were directly analysed by Sanger Sequencing. Sanger sequencing service is provided by Acıbadem Labmed, Acıbadem Health Group. "ab1" formatted files analysed with CLC Workbench (QIAGEN, USA). If the analysis results confirm the variant, the changing image is added to the report.

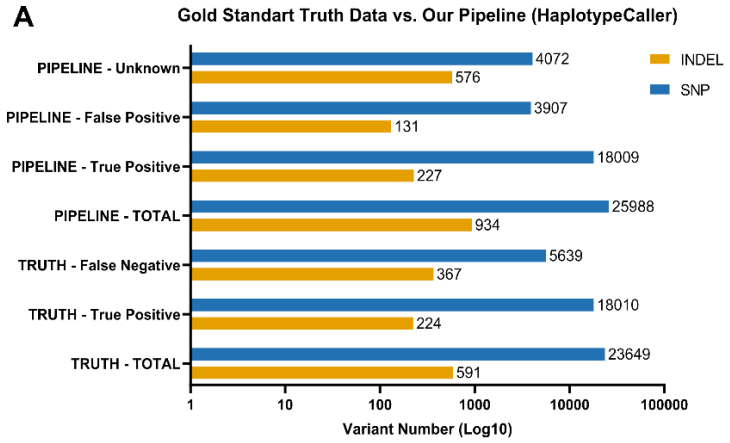
Table 6. List of substances and their concentrations which are used in PCR.

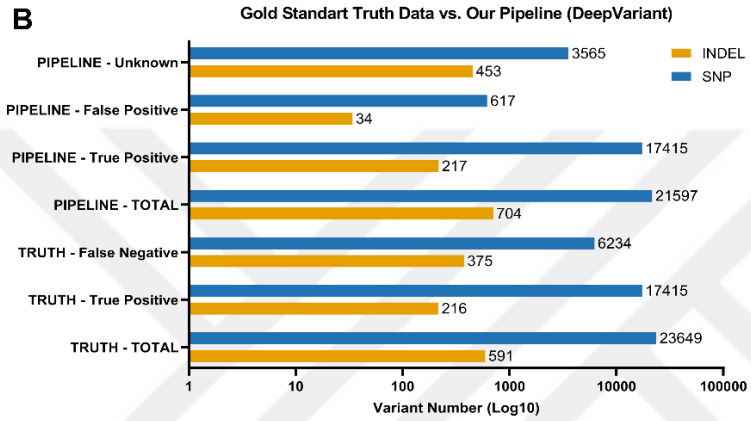
Substance	**Concentration**
Genemark 5X PCR Master Mix	5X
Primers	100uM
dH2O	-
DNA	10 ng

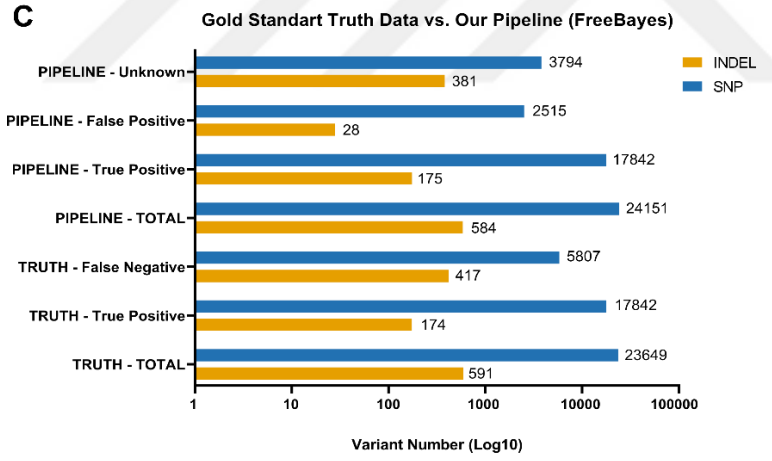
4. RESULTS

4.1. Benchmark Results

Using the hap.py tool, four variant callers and multi-caller VCFs in the ACUGEN were compared with the NA12878 dataset which is known as Truth data. All variant callers give true positive results equal to Truth data for SNPs. However, this was different for the INDEL variations. In INDEL variant calls, all variant callers except the Mpileup tool detected more true positive variants than Truth data. The Mpileup tool detected an equal amount of true positive variants with Truth data (Figure 14).


A
Gold Standart Truth Data vs. Our Pipeline (HaplotypeCaller)
PIPELINE - Unknown
PIPELINE - False Positive
PIPELINE - True Positive
PIPELINE - TOTAL
TRUTH - False Negative
TRUTH - True Positive
TRUTH - TOTAL
4072
576
3907
131
18009
227
25988
934
5639
367
18010
224
23649
591
INDEL
SNP
1
10
100
1000
10000
100000
Variant Number (Log10)


B
Gold Standart Truth Data vs. Our Pipeline (DeepVariant)
PIPELINE - Unknown
PIPELINE - False Positive
PIPELINE - True Positive
PIPELINE - TOTAL
TRUTH - False Negative
TRUTH - True Positive
TRUTH - TOTAL
3565
453
617
34
17415
217
21597
704
6234
375
17415
216
23649
591
INDEL
SNP
1
10
100
1000
10000
100000
Variant Number (Log10)


C
Gold Standart Truth Data vs. Our Pipeline (FreeBayes)
PIPELINE - Unknown
PIPELINE - False Positive
PIPELINE - True Positive
PIPELINE - TOTAL
TRUTH - False Negative
TRUTH - True Positive
TRUTH - TOTAL
3794
381
2515
28
17842
175
24151
584
5807
417
17842
174
23649
591
INDEL
SNP
1
10
100
1000
10000
100000
Variant Number (Log10)

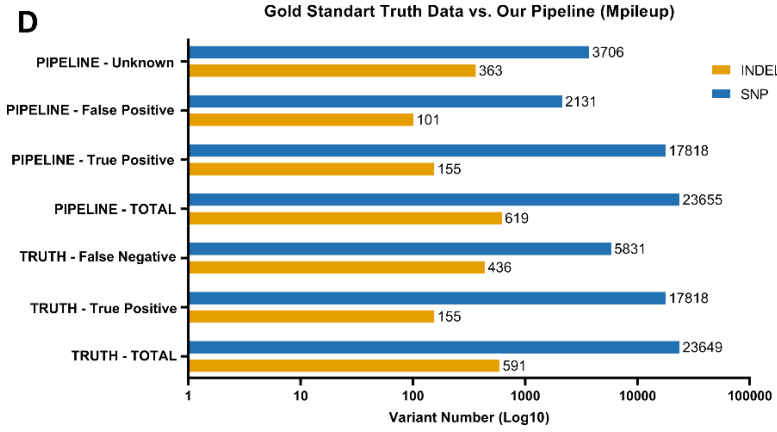

D
Gold Standart Truth Data vs. Our Pipeline (Mpileup)
PIPELINE - Unknown
PIPELINE - False Positive
PIPELINE - True Positive
PIPELINE - TOTAL
TRUTH - False Negative
TRUTH - True Positive
TRUTH - TOTAL
3706
363
2131
101
17818
155
23655
619
5831
436
17818
155
23649
591
INDEL
SNP
1
10
100
1000
10000
100000
Variant Number (Log10)

Figure 14. Comparison of Gold Standard Truth Data vs our pipelines individual VCFs.

A) Gold Standard Truth Data compared with HaplotypeCaller. B) Gold Standard Truth Data compared with DeepVariant. C) Gold Standard Truth Data compared with Freebayes. D) Gold Standard Truth Data compared with Mpileup The terms true positive, false negative, false positive, unknown are defined in Table X.

4.2. Clinical Information of Collected Data

The NGS data and clinical information of 45 index cases were referred to ACURARE from various centers via physicians with an informed consent for re-analysis. All the obtained data were evaluated in terms of suitability for the study. Patients who could not obtain detailed phenotype and clinical information required for differential diagnosis were excluded from the study

4.3. Quality Control Results

According to the quality check results 25 out of 45 cases were suitable for re-analysis. The quality check was performed regarding data quality, comprehensive phenotype information, and the type of the data. Quality control reports were added to the study as appendix (Appendix 3).

The mean depth of coverage was calculated as another quality control parameter apart from the raw data quality control. For this reason, the minimum mean coverage depths for the data to be included in the study were accepted as 30X (±2) for WES and 10X (±1) for WGS. The mosdepth tool used for depth calculations includes a python script for plotting outputs.

4.4. Variant Filtering Results

Twenty-two WES and three WGS data that passed the quality control were filtered. Among 25 index cases, in 9 of them a candidate variant that may be related to the phenotype was determined (Table 7). The cases with a candidate vaiant and their evaluation are summarized in the following section.

Table 7. Features of candidate variants determined in nine patients.

Patient No	Chromosomal Position	RS ID	Zygosity	Gene Function	AF	Candidate Variants	ACMG Classification	CNV results	Confirmation
WES001	chr6-135511354 chr6-135539075	- rs200859137	Het. Het.	Missense Missense	- 0.0000399	*MYB* NM_005375:c.396G>T *MYB* NM_005375:c.1880G>C	VUS VUS	Negative	Confirmed
WES002	chr17-80198415	rs771451722	Het	Missense	0.00000813	*CARD14* NM_001366385:c.1675C>T	VUS	Negative	Confirmed
WES003	chr16-1257426	rs771994752	Het.	Missense	0.0000243	*CACNA1H* NM_021098.c.3059C>T	VUS	Negative	Confirmed
WES005	chr16-88841094	rs758479097	Het.	Missense	0.00000399	*GALNS* NM_001323544:c.338C>T	Pathogenic	Negative	Confirmed at an another center
WES007	chr7-2526921	rs1187978169	Hom.	Missense	0.0000361	*LFNG* NM_001040167:c.1073G>A	Likely Pathogenic	Negative	Confirmed

Table 7. Features of candidate variants determined in nine patients. (continue)

Patient No	Chromosomal Position	RS ID	Zygosity	Gene Function	AF	Candidate Variants	ACMG Classification	CNV results	Confirmation
WES018	chr3-47410304 chr3-47412144	rs759704017 -	Het. Het.	Nonsense Missense	0.00000423 -	*PTPN23* NM_015466:c.2506C>T *PTPN23* NM_015466:c.4124A>C	Pathogenic Pathogenic	Negative	Confirmed in another center
WGS001	chr3-190402388 chr2-227253582	- rs1310347317	Het. Het.	Frameshift Missense	- 0.00000401	*CLDN16* NM_006580:c.166delG *COL4A3* NM_000091:c.709C>T	Likely Pathogenic Likely Pathogenic	Negative	Negative Confirmed
WES021	chr12-111347567	-	Het.	Missense	-	*CUX2* NM_015267:c.3703G>A	VUS	Negative	Confirmed
WGS003	chr19-35729208	-	Het.	Missense	-	*KMT2B* NM_014727:c.4829A>G	Pathogenic	Negative	Confirmed
WES022	chr1-94103130	rs62646862	Het.	Missense	0.00259	*ABCA4* NM_000350:c.455G>A	VUS	Negative	Confirmed

Patient WES001

WES001, a male patient with a suspicion of primary immune deficiency, was referred to our study through his clinician with a previously negative report of WES analysis. The patient presented leukopenia, anemia, severe B-cell lymphocytopenia, abnormal granulocyte morphology, and panhypoagammaglobunemia (Figure 15). From HPO lists, 631 genes associated with phenotype were retrieved and first the homozygous variants were prioritized due to consanguinity in the family. Following the analysis steps three possible candidate variants were discovered. Two VUS variants were detected in the *MYB* gene NM_005375:c.396G>T and NM_005375:c.1880G>C. According to the Sanger sequencing results of the NM_005375:c.396G>T variant in the *MYB* gene, the mother and father were heterozygous, as the proband was found to have the variant homozygous. For the other *MYB* variant, NM_005375:c.1880G>C, the mother and proband were heterozygous carriers while the father was wildtype. Further studies are needed to confirm the disease-causing effects of *MYB* variation (Figure 16).

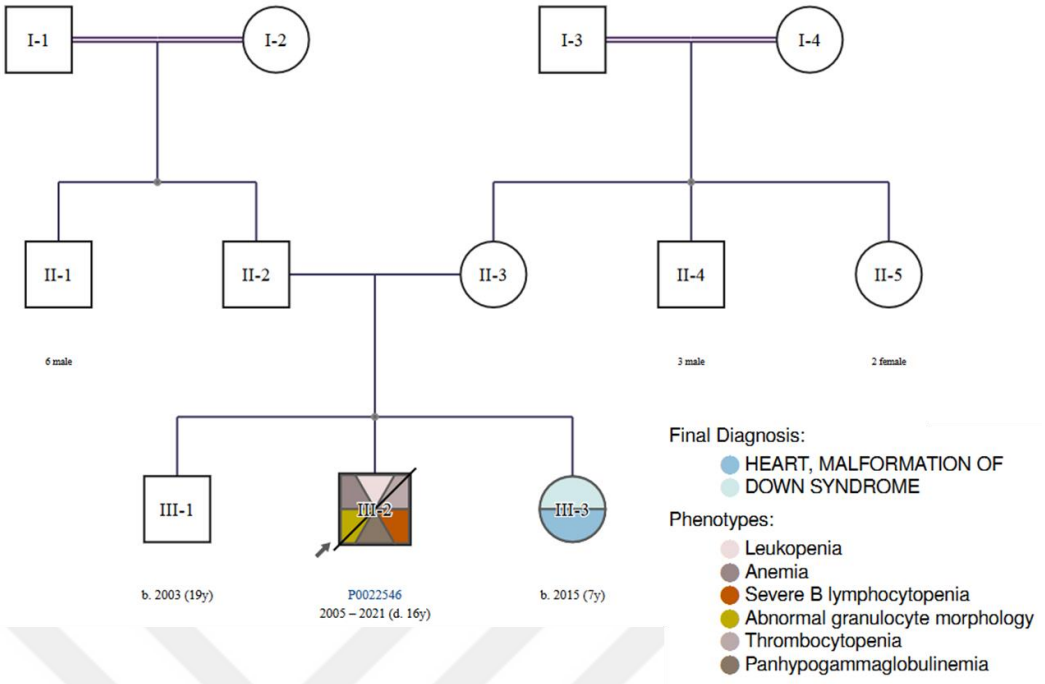


Figure 15. Pedigree and phenotype information of Patient WES001.

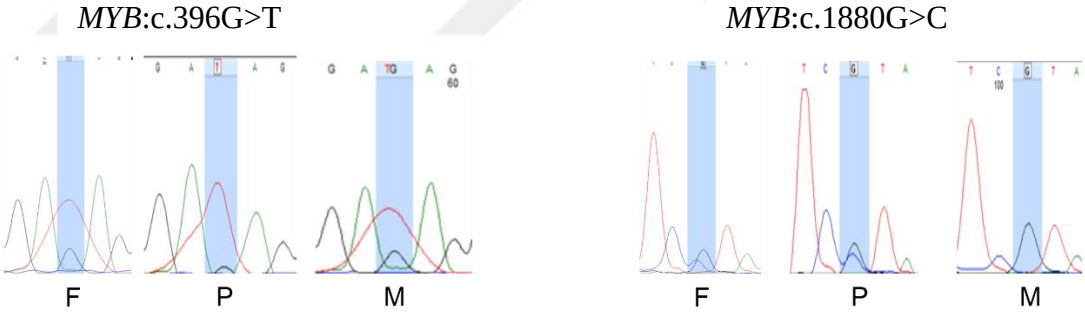


Figure 16. Sanger sequencing results of Patient WES001.

Letter F represents father, letter P represents proband and letter M represents the mother.

Patient WES002

WES002, a 12-year-old male patient, is admitted to the clinic with the diagnosis of inflammation of the large intestine. Additionally, the patient presented ulcerative colitis and psoriasiform dermatitis. During the pre-analysis genetic counselling, multiple members with ulcerative colitis and psoriasiform dermatitis are observed in the family including the mother and relatives from her side (Figure 17). Only 44 genes were acquired from HPO, which are associated with the patient's phenotype. Because there was no consanguineous marriage and numerous affected people, heterozygous variations were thoroughly explored. Following the filtering three potential pathogenic variations were discovered. NM_024110.4:c.1675C>T heterozygous variant was detected in the *CARD14* gene (Appendix 4). It was chosen as a strong candidate since the variant could explain not only the inflammation but also the ulcerative colitis phenotype seen in the patient. For WES002, blood samples were requested from the parents and the proband, but only blood samples from the parents could be obtained. For this reason, Sanger sequencing was applied only to the samples of the parents. As a result of direct sequencing for the NM_024110.4:c.1675C>T variant, the father was heterozygous carrying the variant, while the mother was observed as wildtype (Figure 18) and the variant was eliminated.

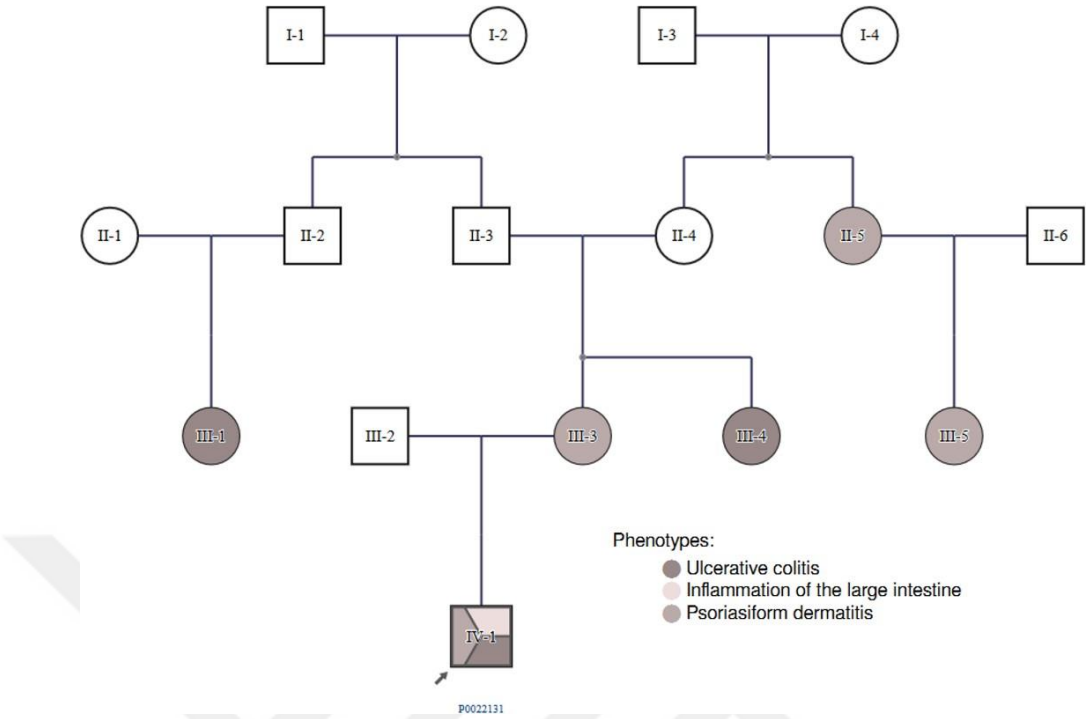


Figure 17. Pedigree and phenotype information of Patient WES002.

CARD14: c.1675C>T

Figure 18. Sanger sequencing results of Patient WES002's parents.

Patient WES003

WES003, a 60-year-old female patient, is admitted to the clinic with complaints of hypertension with hypokalaemia, dissolution, and weakness following meals, especially eating tropic fruits (Figure 19). The result of the WES analysis performed by the clinician was reported as negative. In the light of this information, a re-analysis was performed. From HPO lists, 369 genes associated with phenotype were retrieved. Since there were no consanguinity in the family and the phenotype was milder, as the first step the heterozygous variations were checked. Heterozygous NM_021098.2.c.3059C>T variant was detected in the *CACNA1H* gene and this was confirmed by Sanger sequencing (Figure 20). Additional enzymatic tests were suggested to strengthen this finding. Since the index was an adult no samples were available from the parents or the siblings.

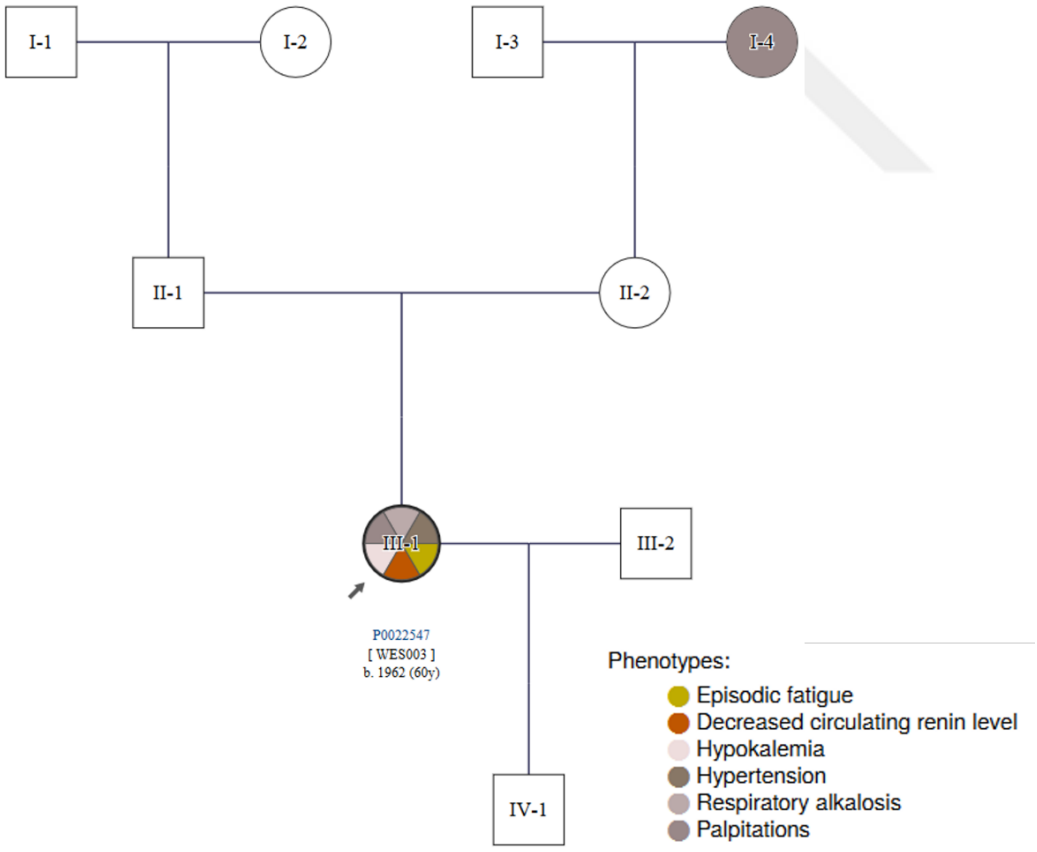


Figure 19. Pedigree and phenotype information of Patient WES003.

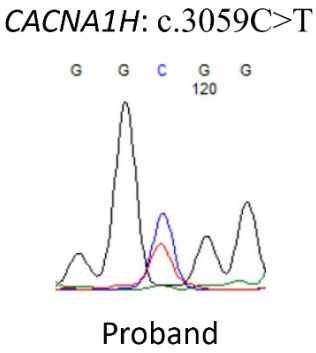


Figure 20. Sanger sequencing results of Patient WES003.

Patient WES005

WES005, a 15-year-old male patient with a suspicion of osteochondrodysplasia, was admitted to the clinic due to disportioned leg length and scoliosis (Figure 21). In the first WES analysis no candidate gene was reported. From HPO lists, 1316 genes associated with phenotype were retrieved. Only one variation that may have a pathogenic effect was found after filtering, a heterozygous NM_001323544:c.338C>T variant in the *GALNS* gene (Figure 22) (Appendix 5). The sanger sequencing was performed in another center and the variant was confirmed by Sanger Sequencing.

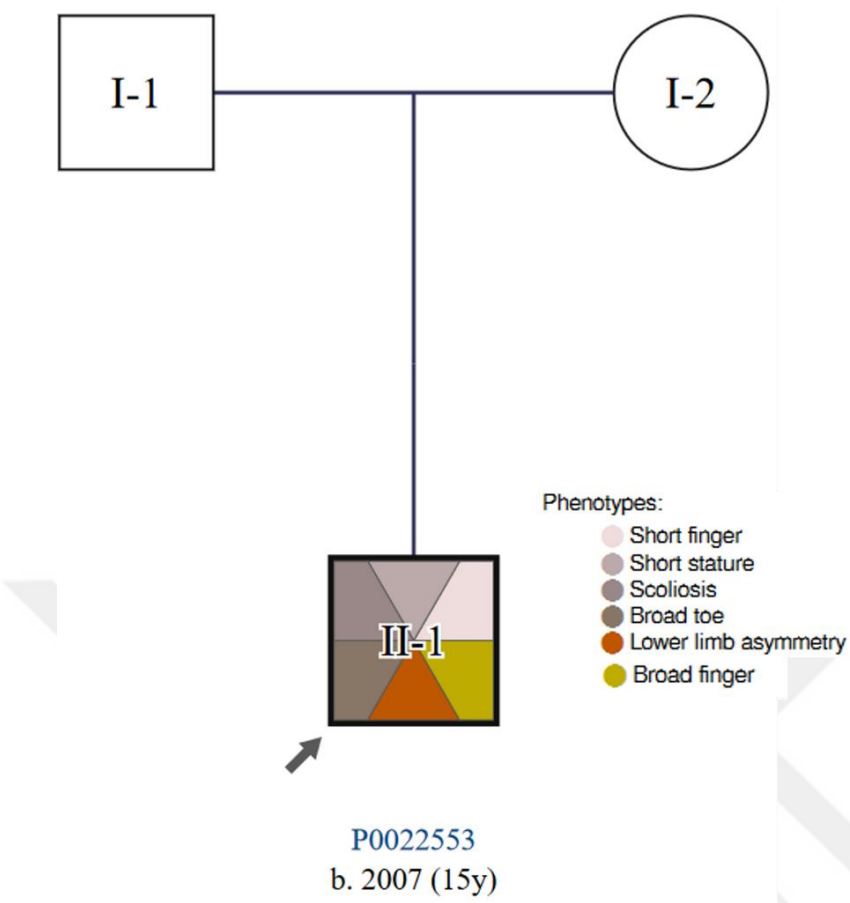


Figure 21. Pedigree and phenotype information of Patient WES005.

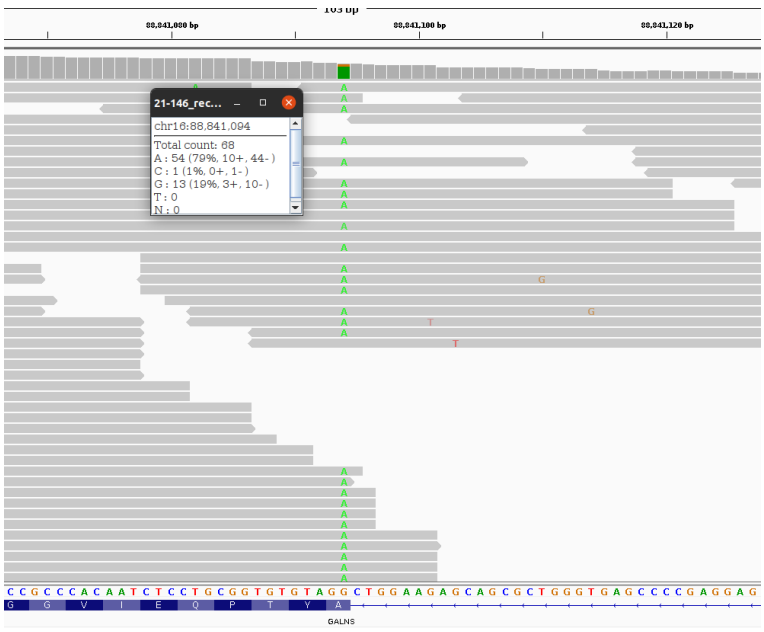


Figure 22. IGV visualisation of the variant found in *GALNS* gene.

Patient WES007

WES007 was a 4-year-old male patient, who was born to a consanguineous marriage with a prediagnosis of spondylocostal dysostosis (Figure 22). From HPO lists, 21 genes associated with phenotype were retrieved. Homozygous filtering based on consanguineous marriages in the family was prioritized and, homozygous NM_001040167:c.1073G>A variant was detected in the *LFNG* gene. It was noteworthy that the variant was homozygous due to its consanguineous marriage in the family, and that it was the only variant explaining the phenotype among nine other variants in the final list. From the Sanger sequencing results for WES007 and his parents, it was observed that the proband and mother were homozygous for the NM_001040167:c.1073G>A variant in the *LFNG* gene, while the father was heterozygous (Figure 23) and the variation was eliminated. There were no another candidate after applied the expanded filtering.

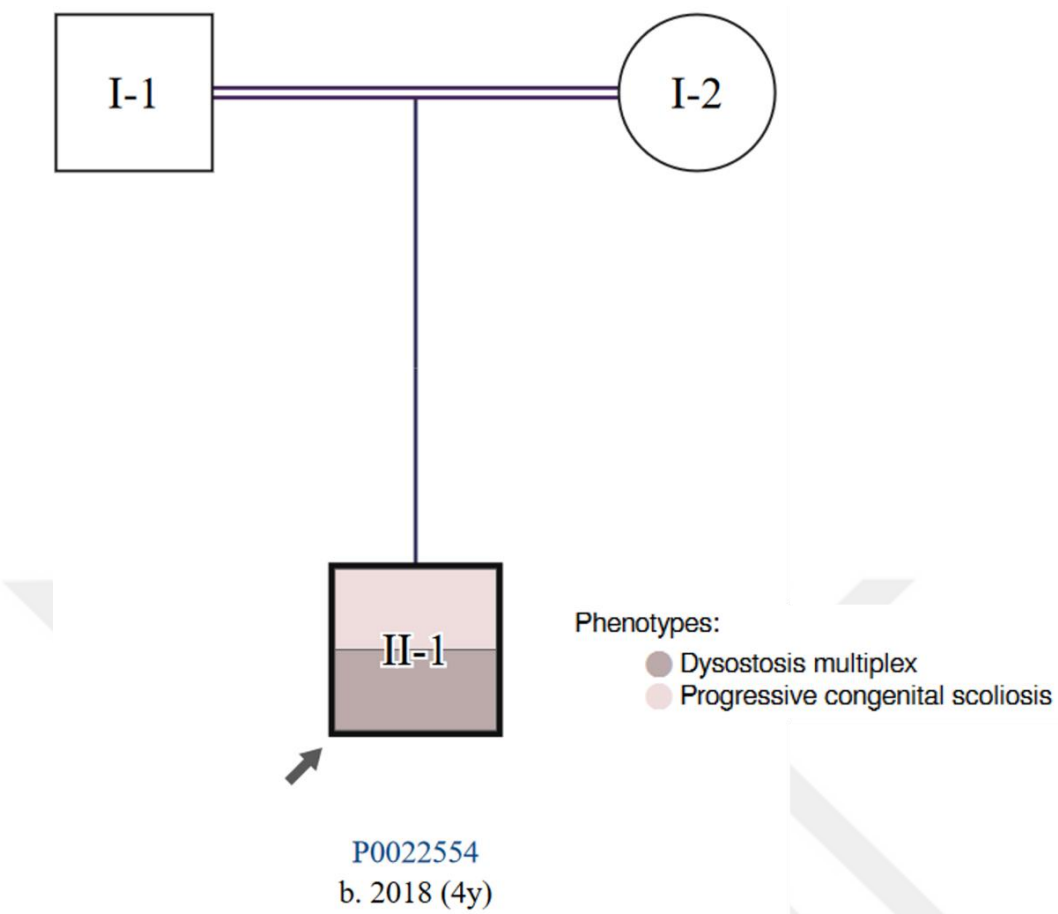


Figure 23. Pedigree and phenotype information of Patient WES007.

LFNG: c.1073G>A

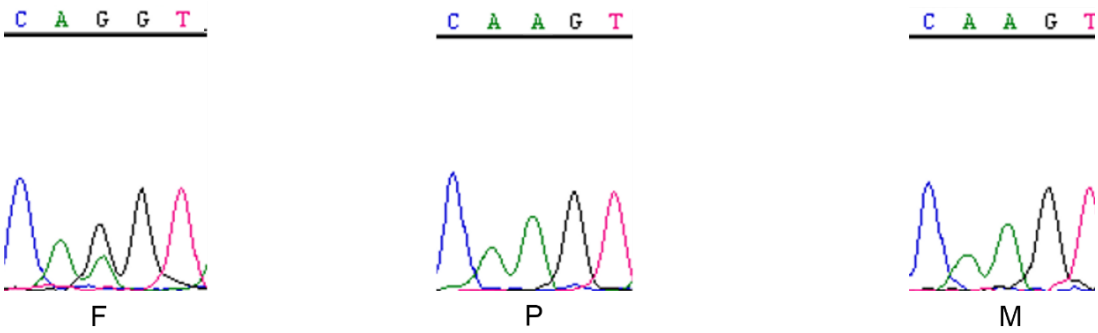


Figure 24. Sanger sequencing results of Patient WES007 and parents.

Letter F represents father, letter P represents proband and letter M represents mother.

Patient WES018

WES008, was a 2-year-old male patient with microcephaly, abnormal globus pallidus morphology, seizures, optic atrophy, abnormal myelination, dysphagia, hyperintensity in cerebral white matter, abnormal corpus callosum morphology (Figure 24). From HPO lists, 2052 genes associated with phenotype were retrieved. It was stated that there was no consanguinity between the mother and father. As a result of filtering NM_015466:c.2506C>T and NM_015466:c.4124A>C variants were detected in the *PTPN23* gene (Figure 26) (Appendix 6). The autosomal recessive inheritance pattern of pathogenic variants in the *PTPN23* gene has drawn attention for the compound heterozygosity status. Patient samples could not be reached, but segregation analysis was performed in a different center and the variants NM_015466:c.2506C>T and NM_015466:c.4124A>C were segregated among the parents.

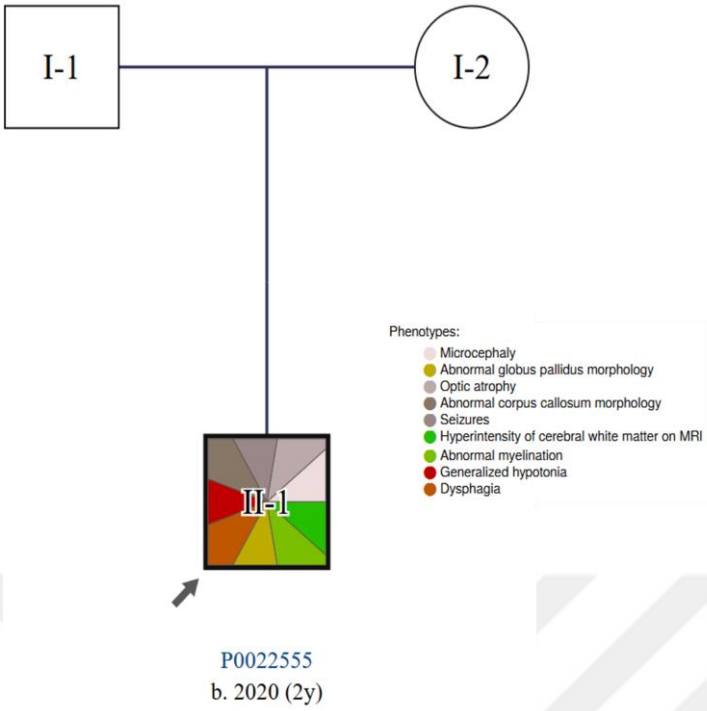


Figure 25. Pedigree and phenotype information of Patient WES018.

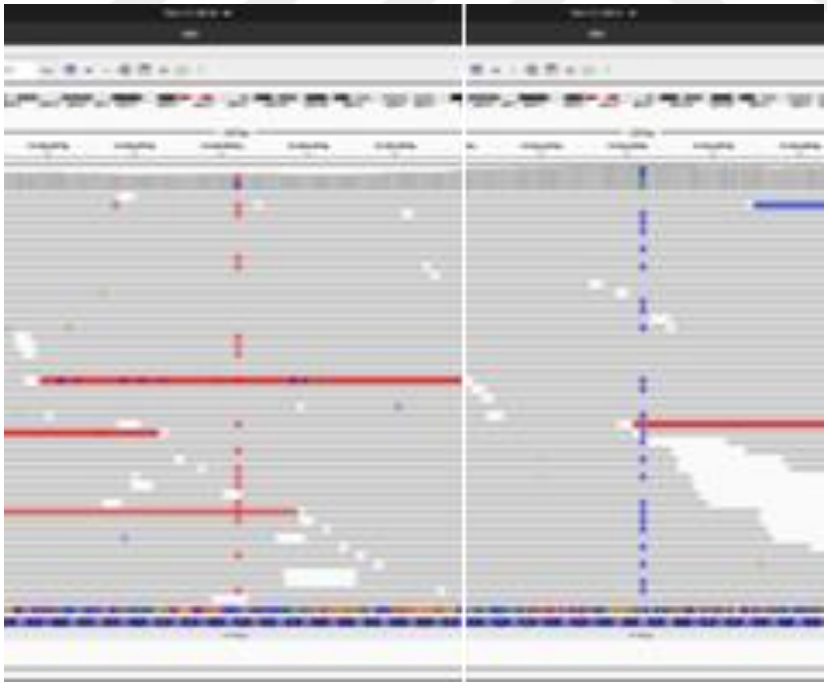


Figure 26. IGV visualisation of NM_015466:c.2506C>T (left) and NM_015466:c.4124A>C (right) variants.

Patient WGS001

WGS001, an 8-year-old male patient presents to the clinic with the phenotype of nephrocalcinosis. The index showed hypercalciuria and hypouricemia as well and there were additional family members with the same findings. The index was born in a consanguineous marriage and there was more than one consanguineous marriage in the family (Figure 25). From HPO lists, 91 genes associated with phenotype were retrieved. The homozygous filtering resulted with benign variations so the heterozygous filtering was continued. As a result of the filtering two candidate variants that may be clinically relevant were identified as two final variants. These variants were; NM_006580:c.166delG in the *CLDN16* gene and NM_000091:c.709C>T in the *COL4A3* gene. Sanger sequencing results for the heterozygous NM_006580:c.166delG *CLDN16* gene variant could not be confirmed, while the NM_000091:c.709C>T variant in the *COL4A3* gene was confirmed. Proband and mother were positive for heterozygous variant, while the father was observed as wildtype (Figure 26) and the variation was eliminated.

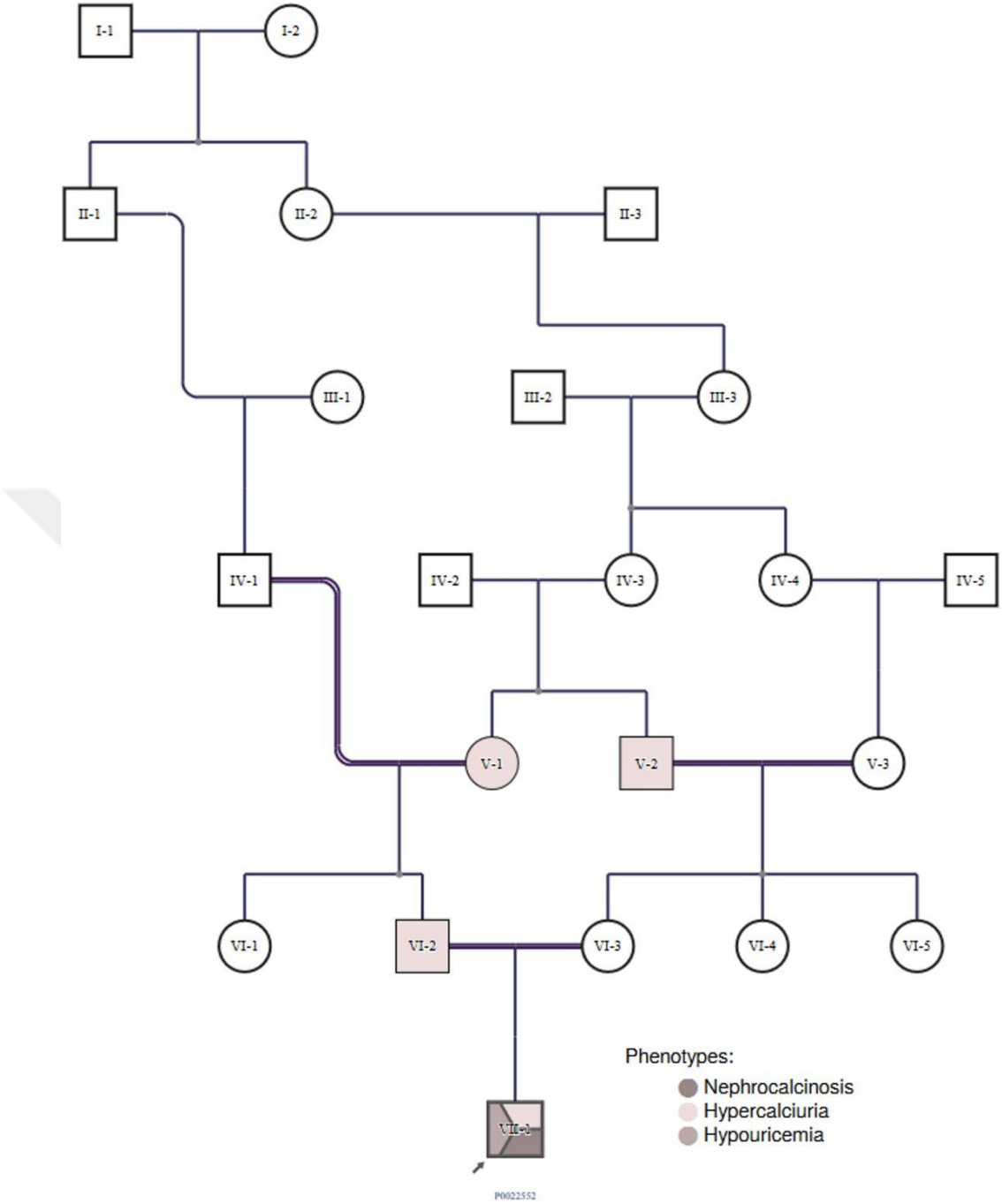


Figure 27. Pedigree and phenotype information of Patient WGS001.

COL4A3: c.709C>T

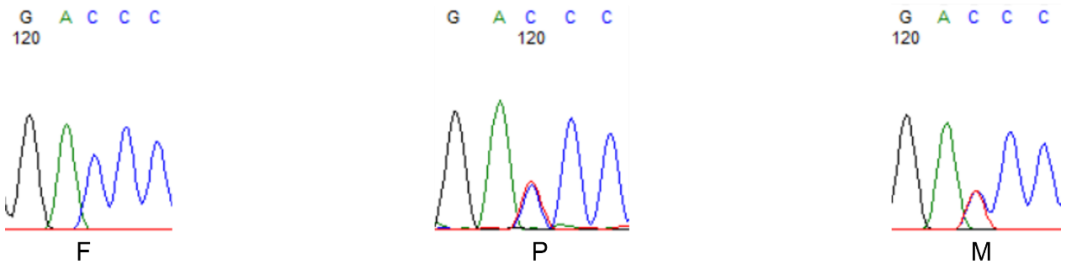


Figure 28. Sanger sequencing results of Patient WGS001 and parents.

Letter F represents father, letter P represents proband and letter M represents mother.

Patient WES021

WES021 9-year-old male patient with phenotype information absence of expected normal physiological development. From HPO lists, 2070 genes associated with phenotype were retrieved. Since there was no other affected individual and no consanguineous marriage, heterozygous variations were evaluated in the pedigree with the possibility of a de novo variant. Heterozygous NM_015267:c.3703G>A variant was detected in the *CUX2* gene as a result of filtering (Figure 27) (Appendix 7). According to Sanger sequencing results for WES021, the mother was wildtype, while the father and proband were heterozygous for the NM_015267:c.3703G>A variant in the *CUX2* gene (Figure 28).

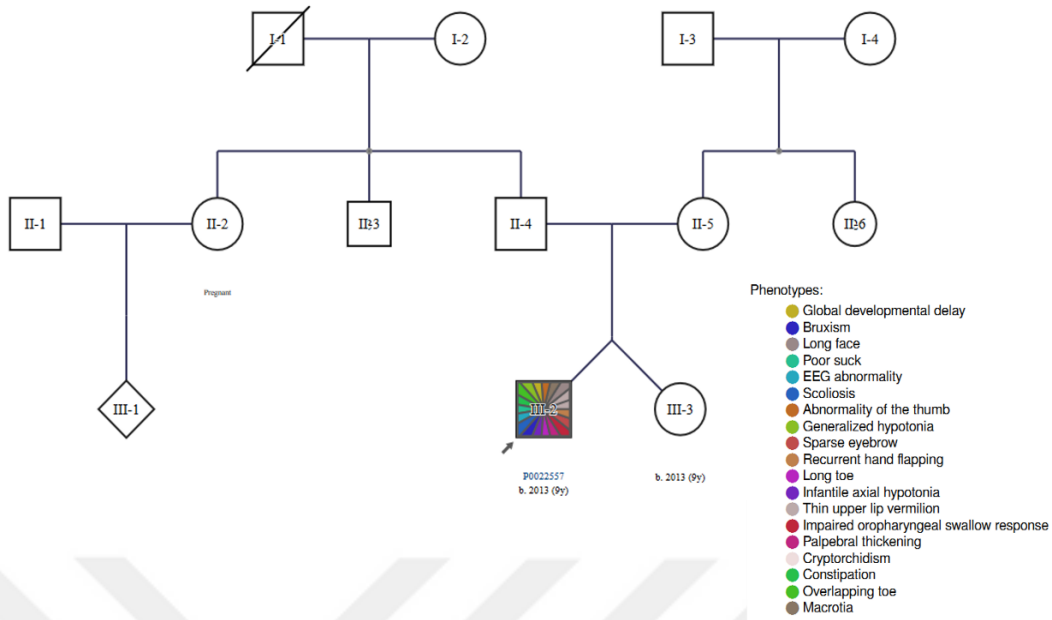


Figure 29. Pedigree and phenotype information of Patient WES021.

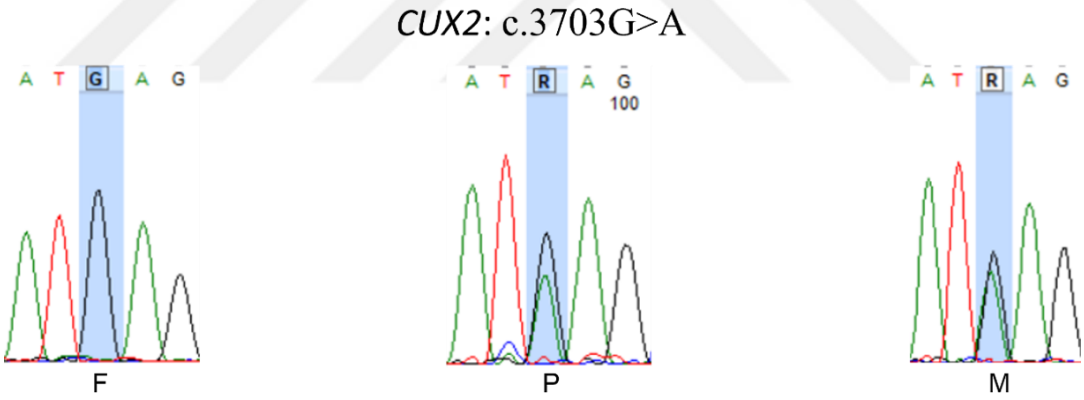


Figure 30. Sanger sequencing results of Patient WES021 and parents.

Letter F represents father, letter P represents proband and letter M represents mother

Patient WGS003

WGS003 is a 13-year-old female patient with severe neurodevelopmental phenotypes (Figure 29). From HPO lists, 1404 genes associated with phenotype were retrieved. Heterozygous NM_014727.3:c.4829A>G variant was found in the *KMT2B* gene other than five different variants, which may be clinically relevant as a result of filtering performed with delayed gross motor development phenotype (Figure 29). Confirmation was done at another center, and the clinical relevance was approved by the clinician. As a result of the analysis the father was wildtype for the NM_014727.3:c.4829A>G variant, while the mother was heterozygous.

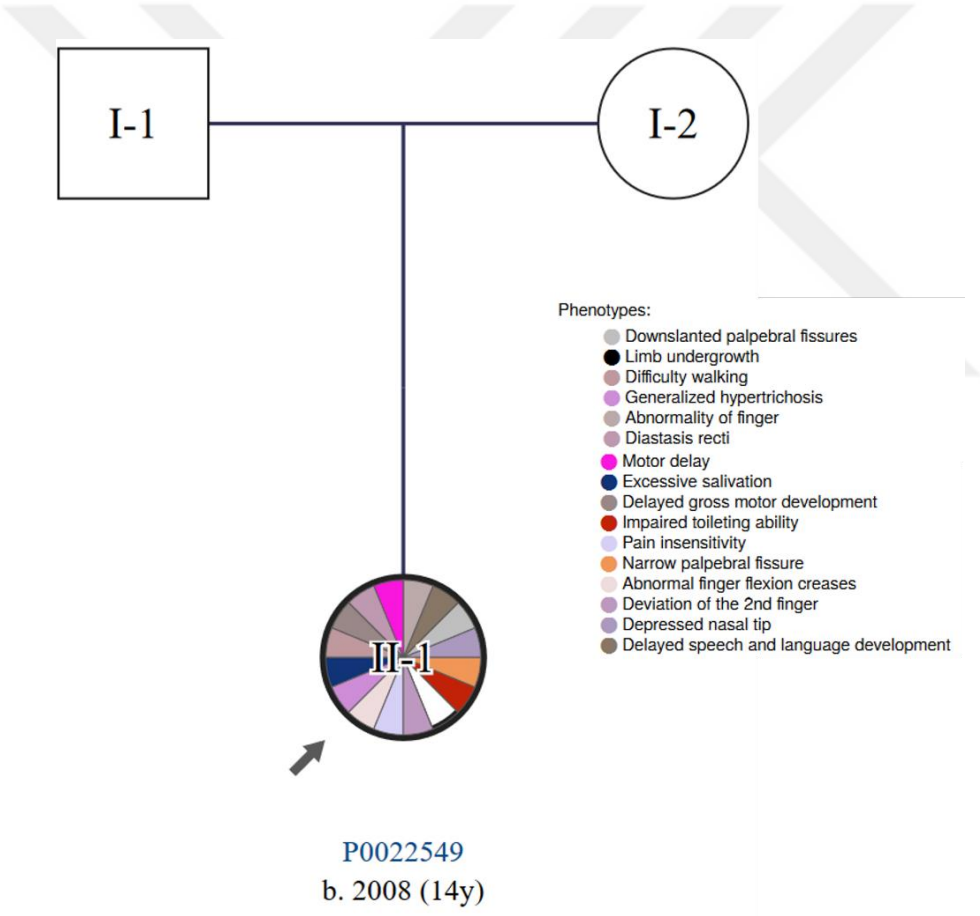


Figure 31. Pedigree and phenotype information of Patient WGS003.

Patient WES022

WES022, male patient with focal segmental glomerulosclerosis phenotype (Figure 30). From HPO lists, 836 genes associated with phenotype were retrieved. Due to the patient's parents' consanguineous marriage, homozygous variations were evaluated first, and no variant with a pathogenic effect was discovered. NM_000350:exon5:c.455G>A variant was detected in *ABCA4* gene. It was shown to be the most closely related to the patient's phenotype among the eight variants identified. According to Sanger sequencing results for WES022, *ABCA4* heterozygous NM_000350:exon5:c.455G>A variant is positive for proband and father, while the mother is wildtype (Figure 31).

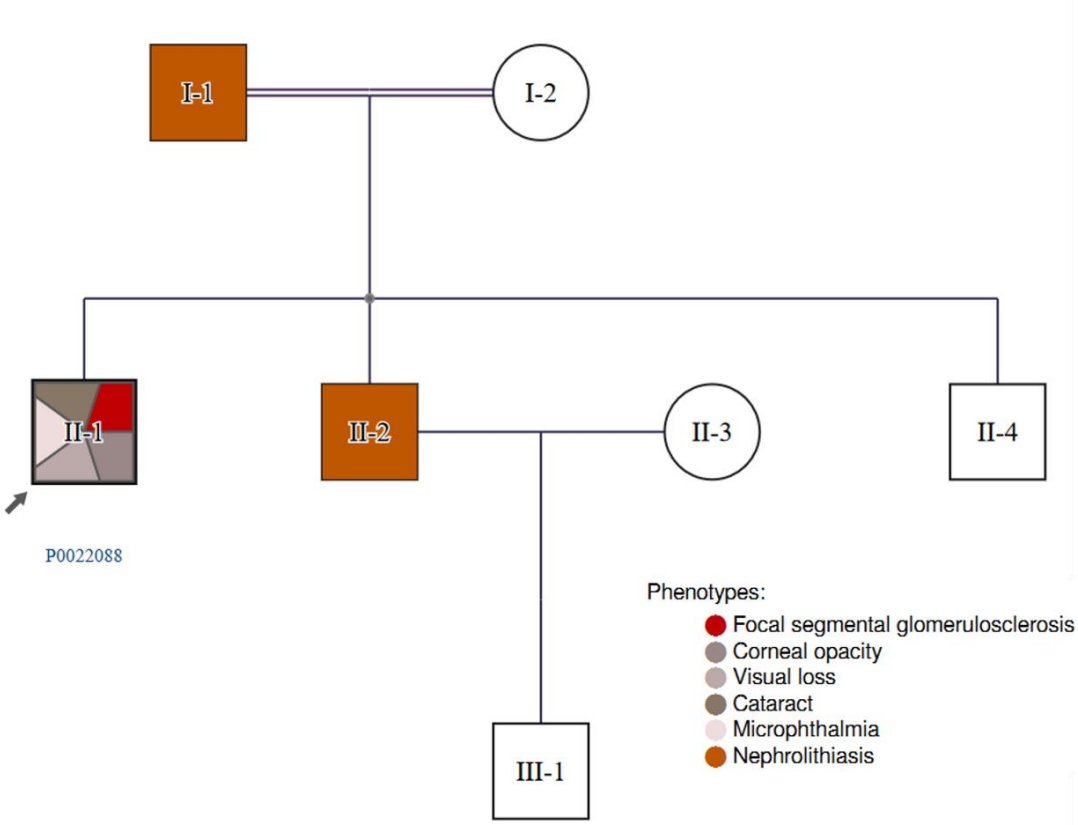


Figure 32. Pedigree and phenotype information of Patient WES022.

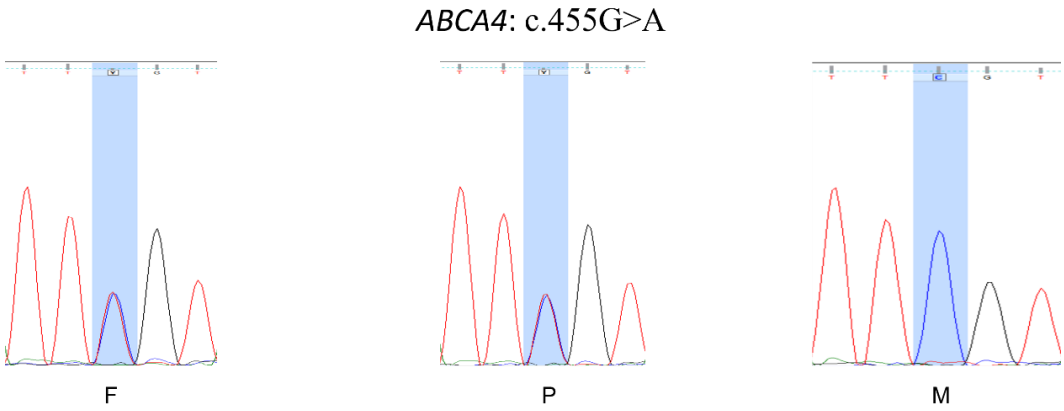


Figure 33. Sanger sequencing results of Patient WES022 and parents.

Letter F represents father, letter P represents proband, and letter M represents the mother.

No variant that could explain the phenotype was found in the reanalysis of 16 patients. "Negative" technical reports were prepared for these patients and these reports were forwarded to the clinician. These data were archived on ACURARE servers for reanalysis 6 months later.

4.5. Copy Number Variants – Structural Variants Results

Whole exome sequencing is still not a reliable tool for detecting CNVs. However, there are different approaches to determine CNVs via WES, where no additional data is available. Here we used WES data of six healthy individuals to compare the healthy reads to patients' data. For this purpose, the SVD-ZRPKM values, which are described in section 3.6, were obtained from these exome data with the patient data. However, we did not find any copy number variants in any of the samples.

For the whole genome data, we analysed the data obtained from the parliament2 tool, which works in our pipeline, by annotating with the AnnotSV tool (89,109). As a result of this analysis, no structural variants that could explain the clinic were found.

As a result of this thesis study, only two of 25 patients were definitively diagnosed (Table 8).

Table 8. Final report table of whole patients.

Patient ID	Final Report
WES001	Negative
WES002	Negative
WES003	Negative
WES004	Negative
WES005	**Solved**
WES006	Negative
WES007	Negative
WES008	Negative
WES009	Negative
WES010	Negative
WES011	Negative
WES012	Negative
WES013	Negative
WES014	Negative
WES015	Negative
WES016	Negative
WES017	Negative
WES018	**Solved**
WES019	Negative
WES020	Negative
WES021	Negative
WES022	Negative
WGS001	Negative
WGS002	Negative
WGS003	Negative

5. DISCUSSION

Diagnosing rare diseases is still a challenging journey today. Due to the inability to diagnose and the complexity of the diseases, many samples are collected, and data are obtained during the diagnostic process. WES is today the most widely used analysis method in the diagnosis of rare diseases. The raw data can be stored, and with the regular improvements in bioinformatics and literature, WES can be reanalysed at different time points. One advantage of reanalysis is that before leading to other approaches, with the recent updates data can be evaluated, and additional cost and stress on the patient can be avoided. However, for reanalysis to be useful significant advances in technology and literature are required. For this reason, regular re-analyses in 6 months to 1-year periods (110), depending on the workforce and computer power, would be efficient, especially for the regions where NGS applications are still costly.

Currently, there is new data for running WGS as a first-line diagnostic tool or at least a trio WES analysis to improve the diagnostic success. WGS and WES were not significantly different in terms of diagnostic efficacy. The likelihood of diagnosis was considerably higher for trios than singletons in studies with within-cohort comparisons of WGS/WES. The diagnostics utility of WGS was 41% and 36% for WES when heterogeneity was precluded. Thus, heterogeneity reduced studies shows diagnostics utility of WGS as 42% and as 38% for WES (111). The advantage of WGS is that one can have the information for both coding and non-coding regions and also structural variants. On the other hand, the generated data is much larger than WES, and needs more storage area, higher computational power and expertise. Additionally, it is still a costly method to be used as a first line diagnostic tool. Still for the sake of the patients and their families, WGS can be suggested where available.

There are several reasons why rare diseases are difficult to diagnose. The first challenge is that the diseases are not often encountered. Hence, there is a lack of knowledge on phenotypic presence or molecular mechanism. Most rare diseases are chronic disorders with heterogenous features, which can be misleading for the right diagnosis as well. With these aspects additional and powerful tools are needed in the

analysis of the data generated from rare diseases. In this thesis, a pipeline suitable for the diagnosis of rare diseases was developed. This pipeline is an automated pipeline that works to avoid missing candidate variants in a data using variant callers that are successful in different variant types. Being an automated pipeline, it enables every user with sufficient computing power to provide the raw data and obtain the VCF file at the end of the process. Working with a short command line also makes this pipeline a user-friendly tool. There are several standardized tools for analysing short-read NGS data, which are mentioned in section 2.7.7. (73,112,113). These are the tools used in almost every pipeline today and are the most successful tools in their field. ACUGEN works with these standardized and successful tools in its field. By integrating the data from each tool, it finally makes it available for analysis. One of the biggest feature that distinguishes ACUGEN from other pipelines is that it uses four different variant callers. For this reason, the differences between the data in the intermediate stages and the other pipelines were not included in this study, and the differences in these data may only arise due to version transitions. These differences are also stated in the version updates of the relevant tools.

NA12878, which accepted as the “Truth data”, is a dataset that is being used for many benchmark studies (88,114–116). In order to show the performance of the ACUGEN pipeline, the VCF prepared with the GATK best practices pipeline was compared with the VCFs coming out of the pipeline we designed in this study (117). In this way, variant results previously called with a different version of HaplotypeCaller were also compared. SNP and INDEL variant numbers were also compared separately to get more accurate results. This was also an opportunity to show which caller was better in determining which variant type. While benchmarking, we selected only variants that received PASS from the callers’ quality control. All variant callers we use in the pipeline, except DeepVariant, found more SNP variants than Truth data (Figure 14). For the INDEL variations, all variant callers except Freebayes found more INDEL variants. The major reason why DeepVariant finds fewer variants is that it calls variants with high accuracy, and also the convolutional neural network (CNN) can account for the readings' complicated dependencies between reads and creates VCFs accordingly. In the DeepVariant study, it was emphasized that deep

learning models such as CNN are better than statistical methods for variant calling (87). Due to a difference in the BED file we utilize in the ACUGEN process, several variants are classed as Unknown when the figures from the benchmark research are examined. When compared against truth data, the majority of these variants, which are referred to as unknown in the pipeline, appear as false negatives. When looking at the false-positive numbers, we can see that ACUGEN identified variations that were not called in the Truth set. Some of these variants are likely due to alignment mistakes, but when the overall numbers (for HaplotypeCaller and Freebayes) are compared, it is clear that HaplotypeCaller calls more variants on SNPs and INDELs, whereas Freebayes calls more variants on SNPs. However, in this study, it is one of the purposes to ensure the discovery of all variants that may be missed during data processing, especially since it is studied in cases that cannot be resolved repeatedly. According to this inference, the importance of having more than one caller in the pipelines designed becomes evident.

The quality of biological material used in sequencing technologies is of great importance (118). The quality of the data to be analysed is also important in order to find the relevant result. For these reasons, incomplete or incorrect phenotype information causes delays in diagnosis and false positive results in rare diseases (119). Elaboration of the given phenotype information will increase the statistical success as it will expand the gene pool screened. Clinicians working more cooperatively with bioinformaticians and taking an active role in variant filtering can eliminate such problems.

Low-quality data yields a large number of false positives and it can also hinder the discovery of existing variants (120). For this reason, trimming is done in the first steps of the designed pipelines. Especially for the discovery of rare variants it is important to call every variant that can be found. However, the trim step may cause these variants to be filtered out as well. For this reason, only high-quality data with detailed phenotype cases were included in the study. The data that were not suitable for the study were also briefed on the standards of this study and the physician was informed why the data was opted out. Coverage has also become a parameter used in

quality controls. WES data used in this study had an average of 30X coverage. It has also been observed that 50X and 100X coverage sequencing is not common, considering the economic conditions throughout our country. High read depth prevents false negative and false positive calls in data (121). In our country, genomic sequencing service is generally provided for 30X average coverage depth for WES and 10X for WGS. However, these calculations are not included in most of the reports of the data we collect, or previous analysis reports could not be obtained.

WES001, it is seen that the parents are not consanguineous, although there are consanguineous marriages in the family. Based on this information, heterozygous variants which may cause the formation of de novo and compound heterozygosity conditions were primarily examined, and then homozygous variants were examined. Compound heterozygosity was considered for two of the three candidate variants obtained in the *MYB* gene. More than one interpretation can be made because the c.396G>T variant is seen in the parents and the proband, and the c.2243G>C variant is seen only in the father and the proband. The presence of two heterozygous variants in the father and the father being healthy reveal two results. One is that the father carries both variants on the same allele; in this case, it can be interpreted that the c.396G>T variant in proband is inherited from the father, and the c.2243G>C variant is inherited from the mother, or that the variants have no pathogenic effect. The third variant, *ELF4*:c.1520C>T, the confirmation of the variant could not be made, although Sanger sequencing of the proband was repeated several times. When looking at the mother and father, heterozygous c.1520C>G was observed in the father, while homozygous c.1520C>A variant was observed in the mother. In line with these results, the most appropriate scenario was compound heterozygosity for the variants in the *MYB* gene. WES001 participated in this thesis as the 4th center analysed and this was the only variant that could be confirmed and the molecular diagnosis of the patient remained unsolved.

WES002, multiple ulcerative colitis and psoriasiform dermatitis phenotypes are seen among the proband's mother and her relatives. The paternal inheritance of the

candidate variant *CARD14* :c.1675C>T made this variant unable to explain the clinic. There was no information about the disease history of the father and his relatives.

Patient WES003 was directed without a medical geneticist's control. The pedigree was drawn with difficulty, subtracting from the patient's son's description, who was also a physician . Considering the phenotype and patient history, it is noteworthy that the patient may have minor epileptic seizures, but no neurological examination information could be obtained. The fact that the patient's grandmother also had similar complaints brought to mind incomplete penetrance. The *CACNA1H* gene is one of the calcium voltage-gated channel genes. The relationship of Voltage-Dependent T-Type Calcium Channel genes with epilepsy has been described many times in the literature (122). In addition, it shows significance in terms of the phenotype of the patient with Hyperaldosteronism, a familial, type IV disease caused by the pathogenic variants of *CACNA1H* (123). After the variant was confirmed by Sanger sequencing, the clinician who directed the patient was informed about the findings and it was recommended to control the aldosterone levels. After this stage, the patient could not be followed up, as the contact with the clinician could not be established.

Patient WES005 participated in this thesis after reanalysis with a preliminary diagnosis of osteochondrodysplasia. This data, which was re-analysed more than once before participating in this study, was repeatedly reported as negative. Finally, NM_001323544:c.338C>T variant was detected in the heterozygous *GALNS* gene from this data analysed with our pipeline in this study, and it was diagnosed as Mucopolysaccharidosis IVA. The *GALNS* gene variant was eliminated in this study but further analysis in another center showed uniparental disomy and the diagnosis was confimed.

Patient WES007 showed phenotypes consistent with Spondylocostal dysostosis 3 disease caused by pathogenic *LFNG* variants (124). Although the variant providing evidence of PP3 strong and PM2 moderate was classified as likely pathogenic with ACMG classification, as a result of the Sanger sequencing analysis performed on the parents and the proband, it was decided that this variant could not explain the clinic

because the mother also had the variant as homozygous (52). Besides this variant, no other variant that could explain the suspected phenotype was determined.

In the analysis performed at an external center for patient WES008, detected variants that were not clinically compatible, so the patient was included in this study by his clinician for re-analysis. The two *PTPN23* variants obtained as a result of the re-analysis were discussed with the clinician, and Sanger sequencing was recommended for co-segregation analysis after the patient's samples were requested. It was stated that these samples were found in the center where sequencing was performed before, and it was found more appropriate to perform the sequencing for the variants found in that center. It was confirmed that one of the alleles with variants came from the mother and the other from the father after co-segregation analysis was performed in an external center where patient samples were located.

WGS001, the first WGS data included in the study, a very large pedigree could be obtained. The family history reveals the multiple consanguineous marriages of this Iranian heritage family. As a result, this patient's homozygous variations were thoroughly investigated. The prevalence of many hypercalcurias in the family history, as well as the fact that the proband had a more severe phenotype than them, bolstered the possibility. However, no candidate could be obtained because all homozygous variants found were previously reported as benign. When all other variants were examined, the obtained *CLDN16* and *COL4A3* gene variants were accepted as candidates, and the *CLDN16* variant was eliminated in Sanger sequencing with samples taken from the parents and the proband. The primers were designed with the given transcript version by the pipeline and the variant could not be confirmed. Further analysis revealed that there was a version update of the transcript NM_006580 transcript and 200 nucleotides were deleted from the coding sequence start region (125,126). For this reason, the region we work on was shifted to the 5' UTR region. When the *COL4A3* gene variant was examined, it was classified as a VUS variant according to the ACMG classification, and it was noteworthy that the frameshift variants in its position were reported as pathogenic (127). For this reason, it was presented as the strongest candidate and confirmed by Sanger sequencing. Considering

that the monoallelic and biallelic forms of *COL4A3* pathogenic variants are involved in the kidneys, it has been reported to the clinician as the strongest candidate (128).

Patient WES021 was analysed with his parents. Result of the co-segregation analysis, the father was wildtype for the variant but the mother was heterozygous for *CUX2* variant. Therefore, the pathogenicity of the variant we obtained for this case was weakened. The clinician was informed for the possibility of incomplete penetrance.

Patient WGS003 was transferred by the clinician for re-analysis with deep phenotype information. The *KMT2B* variant obtained as a result of the re-analysis was also reported by the previously analyzed center. In addition, as a result of Sanger sequencing, it was stated that the patient's mother also carried the same variant. After filtering for Structural and SNP-INDEL variants, no other candidate variants were found. The data of this patient were archived for re-analysis 6 months later. Trio WGS was suggested for future studies.

Patient WES022, the *ABCA4* variant discovered in the re-analysis results of patient was thought to be the cause of the patient's vision problems. The data of the patient whose mother and father were related were first analyzed for homozygous variants, but no variant that could explain the clinic was found. As a result of filtering for heterozygous and compound heterozygous variants, no variant that could explain kidney problems was found, while *ABCA4* variant, which was thought to explain the eye problems of the patient, was shared with the clinician. Sanger sequencing had to be performed twice in the parents and proband because the mother and father were heterozygous carrying the variant, while the proband appeared as wildtype as a result of the first sequencing. The father and proband were sequenced again after we warned the sequencing center, and it was observed that both father and proband had the variant as heterozygous.

It was noteworthy that no structural variants were found as a result of these analyzes. The working principle of the CNV analysis tools for WES data is designed

to find the copy number changes over the increase and decrease in the mean reading depths by distinguishing them from the healthy control data (102). However, sequencing all samples at the same depth and with the same sequencing kits will make the result more meaningful and reliable so that this data can be interpreted and analyzed correctly. Most of the data collected for re-analysis was sequenced at different centers using different kits. However, information about the sequencing of healthy control data could not be obtained. This makes the reliability of CNV data debatable from WES data. Since no CNV data were available in this study, it is not possible to make an accurate discussion. When we looked at the WGS data, we did not find a clinical relationship in the variants that were found in the data that we made a structural variant call with parliament2 and annotated with AnnotSV. Due to the large size of the HTML files obtained from AnnotSV, we had difficulties in running these files on personal computers, so we had to open them on computers with high RAM on the server. We also had problems opening the SNP-INDEL variant lists we obtained from WGS on PCs. Since the amount of RAM required for processing big data is only available on ACURARE servers, analyzes were performed on workstation computers. In order to use such tools on personal computers or standard non-HPC business computers, remote systems should be developed or computers with high RAM should be provided to those who will do this job, where they can view and analyze such files.

Medical and experimental limitations as well as technical limitations were frequently encountered in this study. Deep phenotyping is especially important for the analysis of patients with negative results. Incomplete phenotyping is one of the most important causes of misdiagnosis and undiagnosed patients. We did not include the data, which did not meet our standards, which would be waste of time and power. It would also be no different from previous analyses that were made in other centres. It has been observed that deep phenotyping knowledge is better made especially by medical geneticists who are experts in their field. For this reason, we think that it is important that the patients who will undergo genetic analysis are also examined by medical geneticists. Analysis reports given earlier were also an element that we paid attention to in this study. When reporting the variants, they should be written according to the HGVS nomenclature and transcript IDs should be added strictly. In this way, we

have made a preliminary analysis against the missing variant notification in our pipeline by checking that the previously reported variants also come out of the ACUGEN pipeline, even if they are not clinically relevant. Experimentally, the samples we submitted for Sanger sequencing were not often mixed with each other, but it was always a problem to be considered. For this reason, it would be appropriate to repeat the confirmation analyses when necessary, and to check again if it does not explain the appropriate inheritance pattern. While making the primary designs, the wrong design caused by the version changes was another problem in this study. It will be in the interest of the researcher to examine the version differences of the primers that will be designed especially for regions close to the exon or coding boundaries.

6. CONCLUSION

Molecular Diagnosis of Undiagnosed and Rare Disease Patients with Whole-Exome Re-Analysis study has two different types of outcomes. Firstly, a bioinformatics outcome, a genomic NGS pipeline was developed for the process and analysis of undiagnosed patient data. ACUGEN is still under development and updated by supervision of Özkan Özdemir, PhD to solve bioinformatics bottle-necks in molecular diagnosis of genomic NGS data. Secondly, clinic outcomes, clinically related variant explorations, interpretations, and a novel *PTPN23*:NM_015466:c.4124A>C variant was reported. Eight percent of the total patients were clinically diagnosed in this study and negative or patients with clinically unexplained variant results will be re-analysed in ACURARE after a year. Besides these, this study shows that the expanding filtering methods may be applied to the undiagnosed patients in standard to improve the diagnosis rates of WES and WGS.

As a conclusion of this study, we hope that re-analysis of undiagnosed patients and using of expanding filtering methods will become widespread and ACUGEN will become one of the state-of-the-art steps in this field.

In summary with this thesis;

- A pipeline for NGS data analysis for undiagnosed cases has been developed.
- The importance and success of re-analysis has been demonstrated.
- The methods and experiences used in this study were provided as a resource for future rare disease studies.

7. REFERENCES

1. Wright CF, FitzPatrick DR, Firth H V. Paediatric genomics: diagnosing rare disease in children. Nat Rev Genet [Internet]. 2018 May 1 [cited 2022 Mar 7];19(5):253–68. Available from: https://pubmed.ncbi.nlm.nih.gov/29398702/
2. Boulanger V, Schlemmer M, Rossov S, Seebald A, Gavin P. Establishing Patient Registries for Rare Diseases: Rationale and Challenges. Pharmaceut Med [Internet]. 123AD;34:185–90. Available from: https://doi.org/10.1007/s40290-020-00332-1
3. Ferreira CR. The burden of rare diseases. Am J Med Genet Part A [Internet]. 2019 Jun 1 [cited 2022 Mar 7];179(6):885–92. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/ajmg.a.61124
4. Mccombie WR, Mcpherson JD, Mardis ER. Next-Generation Sequencing Technologies. 2019; Available from: http://perspectivesinmedicine.cshlp.org/
5. Best S, Wou K, Vora N, Van Der Veyver IB, Wapner R, Chitty LS. SPECIAL TOPIC ISSUE ON ADVANCES IN THE DIAGNOSIS OF SINGLE GENE DISORDERS Promises, pitfalls and practicalities of prenatal whole exome sequencing. 2017;
6. Petersen B-S, Fredrich B, Hoeppner MP, Ellinghaus D, Franke A. Opportunities and challenges of whole-genome and-exome sequencing.
7. Roessler HI, Knoers NVAM, van Haelst MM, van Haaften G. Drug Repurposing for Rare Diseases. Trends Pharmacol Sci [Internet]. 2021 Apr 1 [cited 2022 Mar 3];42(4):255–67. Available from: https://doi.org/10.1016/j.tips.2021.01.003
8. Rahit KMTH, Tarailo-Graovac M. Genetic Modifiers and Rare Mendelian Disease. Genes (Basel) [Internet]. 2020 Feb 25 [cited 2022 Mar 7];11(3). Available from: https://pubmed.ncbi.nlm.nih.gov/32106447/
9. Orphanet: About rare diseases [Internet]. [cited 2022 Mar 7]. Available from: https://www.orpha.net/consor/cgi-bin/Education_AboutRareDiseases.php?lng=EN
10. Haendel M, Vasilevsky N, Unni D, Bologa C, Harris N, Rehm H, et al. How many rare diseases are there? Nat Rev Drug Discov [Internet]. 2020 Feb 1 [cited 2022 Mar 7];19(2):77–8. Available from: https://pubmed.ncbi.nlm.nih.gov/32020066/
11. Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, et al. A brief history of human disease genetics. Nature [Internet]. 2020 Jan 9 [cited 2022 Mar 7];577(7789):179–89. Available from: https://pubmed.ncbi.nlm.nih.gov/31915397/
12. Hartley T, Lemire G, Kernohan KD, Howley HE, Adams DR, Boycott KM. New Diagnostic Approaches for Undiagnosed Rare Genetic Diseases. Annu Rev Genomics Hum Genet [Internet]. 2020 Aug 31 [cited 2022 Mar 7];21:351–72. Available from: https://pubmed.ncbi.nlm.nih.gov/32283948/
13. Kulski JK. Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications. Next Gener Seq - Adv Appl Challenges [Internet]. 2016 Jan 14 [cited 2022 Feb 18]; Available from: https://www.intechopen.com/chapters/49602
14. Pijuan J, Rodríguez-Sanz M, Natera-de Benito D, Ortez C, Altimir A, Osuna-López M, et al. Translational Diagnostics: An In-House Pipeline to Validate Genetic Variants in Children with Undiagnosed and Rare Diseases. J Mol Diagnostics. 2021 Jan 1;23(1):71–90.
15. Wise AL, Manolio TA, Mensah GA, Peterson JF, Roden DM, Tamburro C, et al. Genomic Medicine for Undiagnosed Diseases. Lancet (London, England) [Internet]. 2019 Aug 10 [cited 2021 Sep 14];394(10197):533. Available from: /pmc/articles/PMC6709871/
16. General description [Internet]. [cited 2022 Apr 25]. Available from: https://www.udninternational.org/schede-8-general_description
17. Taruscio D, Groft SC, Cederroth H, Melegh B, Lasko P, Kosaki K, et al. Undiagnosed Diseases Network International (UDNI): White paper for global actions to meet patient needs. Mol Genet Metab. 2015 Dec 1;116(4):223–5.
18. Tifft CJ, Adams DR. The National Institutes of Health undiagnosed diseases program.
19. Human Genome Project Results [Internet]. [cited 2022 May 9]. Available from: https://www.genome.gov/human-genome-project/results
20. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. 2017; Available from:

http://www.genome.org/cgi/doi/10.1101/gr.213611.116.
21. Lander S, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome International Human Genome Sequencing Consortium* The Sanger Centre: Beijing Genomics Institute/Human Genome Center [Internet]. Vol. 409, NATURE. 2001. Available from: www.nature.com
22. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze A V, Mikheenko A, et al. The complete sequence of a human genome [Internet]. Available from: https://www.science.org
23. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature. 2015 Sep 30;526(7571):68–74.
24. Introduction to Patches - Genome Reference Consortium [Internet]. [cited 2022 Apr 29]. Available from: https://www.ncbi.nlm.nih.gov/grc/help/patches/
25. Alberts , B. Molecular biology of the cell. 2015.
26. Sequence Variant Nomenclature [Internet]. [cited 2022 Mar 7]. Available from: https://varnomen.hgvs.org/bg-material/glossary/
27. Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, et al. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Hum Mol Genet [Internet]. 2014 Nov 15 [cited 2022 Mar 7];23(22):5866–78. Available from: https://pubmed.ncbi.nlm.nih.gov/24939910/
28. Ule J, Blencowe BJ. Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. Mol Cell [Internet]. 2019 Oct 17 [cited 2022 Mar 7];76(2):329–45. Available from: https://pubmed.ncbi.nlm.nih.gov/31626751/
29. Ward AJ, Cooper TA. The pathobiology of splicing. J Pathol [Internet]. 2010 Jan [cited 2022 Mar 7];220(2):152–63. Available from: https://pubmed.ncbi.nlm.nih.gov/19918805/
30. Zien A, Schölkopf B, Tsuda K, Vert JP. A primer on molecular biology. Kernel Methods Comput Biol 3-34. 2004;
31. Shafee T, Lowe R. Eukaryotic and prokaryotic gene structure. 2017;4(1).
32. Genetics vs. Genomics Fact Sheet [Internet]. [cited 2022 Mar 7]. Available from: https://www.genome.gov/about-genomics/fact-sheets/Genetics-vs-Genomics
33. Nussbaum , McInnes, Roderick R.,, Willard, Huntington F.,, Hamosh, Ada,, Nussbaum, Robert L.,, RL. Thompson & Thompson genetics in medicine [Internet]. 2016. Available from: https://www.clinicalkey.com/dura/browse/bookChapter/3-s2.0-C2009059798X
34. Ku CS, Loy EY, Salim A, Pawitan Y, Chia KS. The discovery of human genetic variations and their use as disease markers: past, present and future. J Hum Genet 2010 557 [Internet]. 2010 May 20 [cited 2022 Apr 3];55(7):403–15. Available from: https://www.nature.com/articles/jhg201055
35. Collins FS, Brooks LD, Chakravarti A. A DNA Polymorphism Discovery Resource for Research on Human Genetic Variation. Genome Res [Internet]. 1998 Dec 1 [cited 2022 Apr 15];8(12):1229–31. Available from: https://genome.cshlp.org/content/8/12/1229.full
36. Klug, W. S., Cummings MR. Concepts of Genetics. Prentice Hall; 2006.
37. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar S V., et al. A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science [Internet]. 2007 Jan 26 [cited 2022 Apr 27];315(5811):525–8. Available from: https://pubmed.ncbi.nlm.nih.gov/17185560/
38. Strachan T, Read A. Human Molecular Genetics [Internet]. CRC Press; 2018. Available from: https://books.google.com.tr/books?id=dSwWBAAAQBAJ
39. Watson JD. Molecular Biology of the Gene [Internet]. Pearson Education; 2004. Available from: https://books.google.com.tr/books?id=VMm3SYa4k%5C_oC
40. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res [Internet]. 2006 [cited 2022 Apr 18];16(9):1182–90. Available from: https://pubmed.ncbi.nlm.nih.gov/16902084/
41. Väli Ü, Brandström M, Johansson M, Ellegren H. Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. 2008; Available from: http://www.biomedcentral.com/1471-2156/9/8
42. Bai H, Cao Y, Quan J, Dong L, Li Z, Zhu Y, et al. Identifying the genome-wide sequence variations and developing new molecular markers for genetics research by re-sequencing a Landrace cultivar of foxtail millet. PLoS One [Internet]. 2013 Sep 10 [cited 2022 Apr 18];8(9). Available from: https://pubmed.ncbi.nlm.nih.gov/24039970/

43. Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. A map of human genome variation from population-scale sequencing. Nature [Internet]. 2010 Oct 28 [cited 2022 Apr 18];467(7319):1061–73. Available from: https://pubmed.ncbi.nlm.nih.gov/20981092/
44. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Stephen Pittard W, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. Available from: www.genome.org/cgi/doi/10.1101/gr.4565806.
45. Lupski JR. Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet. 1998 Oct 1;14(10):417–22.
46. Id JH, Saif R, Jagannathan V, Schmocker C, Zeindler F, Bangerter E, et al. Selection signatures in goats reveal copy number variants underlying breed-defining coat color phenotypes. 2019; Available from: https://doi.org/10.1371/journal.pgen.1008536
47. Wright D, Boije H, Meadows JRS, Bed'hom B, Gourichon D. Copy Number Variation in Intron 1 of SOX5 Causes the Pea-comb Phenotype in Chickens. PLoS Genet [Internet]. 2009;5(6):1000512. Available from: www.plosgenetics.org
48. Sharp AJ, Cheng Z, Eichler EE. Structural Variation of the Human Genome. 2006; Available from: www.annualreviews.org
49. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, et al. Copy number variation: New insights in genome diversity. Genome Res. 2006;16(8):949–61.
50. Richards CS, Bale S, Bellissimo DB, Das S, Grody WW, Hegde MR, et al. ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. Genet Med [Internet]. 2008 Apr [cited 2022 Mar 7];10(4):294–300. Available from: https://pubmed.ncbi.nlm.nih.gov/18414213/
51. Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, et al. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Hum Mutat [Internet]. 2008 Nov [cited 2022 Mar 7];29(11):1282–91. Available from: https://pubmed.ncbi.nlm.nih.gov/18951446/
52. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, 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 8;17(5):405–24.
53. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A [Internet]. 1977 [cited 2022 Mar 7];74(12):5463–7. Available from: https://pubmed.ncbi.nlm.nih.gov/271968/
54. Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P. Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. Anal Biochem [Internet]. 1996;242(1):84–9. Available from: https://www.sciencedirect.com/science/article/pii/S0003269796904327
55. Tucker T, Marra M, Friedman JM. Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine. Am J Hum Genet. 2009 Aug 14;85(2):142–54.
56. Illumina. Technology Spotlight: Illumina ® Sequencing.
57. Gulilat M, Lamb T, Teft WA, Wang J, Dron JS, Robinson JF, et al. Targeted next generation sequencing as a tool for precision medicine. BMC Med Genomics [Internet]. 2019 Jun 3 [cited 2022 Mar 15];12(1):1–17. Available from: https://bmcmedgenomics.biomedcentral.com/articles/10.1186/s12920-019-0527-2
58. Kanzi AM, San JE, Chimukangara B, Wilkinson E, Fish M, Ramsuran V, et al. Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance. Front Genet. 2020 Oct 23;11:1250.
59. Alfaro MP, Sepulveda JL, Lyon E. Molecular testing for targeted therapies and pharmacogenomics. Accurate Results Clin Lab. 2019 Jan 1;349–63.
60. Mertes F, ElSharawy A, Sauer S, van Helvoort JMLM, van der Zaag PJ, Franke A, et al. Targeted enrichment of genomic DNA regions for next-generation sequencing. Brief Funct Genomics [Internet]. 2011 Nov [cited 2022 Mar 15];10(6):374. Available from: /pmc/articles/PMC3245553/
61. Gahl WA, Adams DR, Markello TC, Toro C, Tifft CJ. Chapter 101 - Genetic Approaches to Rare and Undiagnosed Diseases. In: Nelson Textbook of Pediatrics. 2020. p. 683–7.
62. Yang Y, Muzny DM, Xia F, Niu Z, Person ; Richard, Ding Y, et al. Molecular Findings Among Patients Referred for Clinical Whole-Exome Sequencing. JAMA [Internet]. 2014;312(18):1870–9. Available from: https://jamanetwork.com/

63. Bertoli-Avella AM, Beetz C, Ameziane N, Eugenia Rocha M, Guatibonza P, Pereira C, et al. Successful application of genome sequencing in a diagnostic setting: 1007 index cases from a clinically heterogeneous cohort. Eur J Hum Genet [Internet]. 2021;29:141–53. Available from: https://doi.org/10.1038/s41431-020-00713-9
64. Whole-Genome Sequencing [Internet]. [cited 2022 Mar 10]. Available from: https://emea.illumina.com/techniques/sequencing/dna-sequencing/whole-genome-sequencing.html
65. Hawkins GA. Analysis of Human Genetic Variations Using DNA Sequencing. Basic Sci Methods Clin Res. 2017 Jan 1;77–98.
66. Mizuguchi T, Suzuki T, Abe C, Umemura A, Tokunaga K, Kawai Y, et al. A 12-kb structural variation in progressive myoclonic epilepsy was newly identified by long-read whole-genome sequencing. J Hum Genet 2019 645 [Internet]. 2019 Feb 13 [cited 2022 Mar 10];64(5):359–68. Available from: https://www.nature.com/articles/s10038-019-0569-5
67. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet [Internet]. 2011 Jan [cited 2022 Mar 10];13(1):36–46. Available from: https://pubmed.ncbi.nlm.nih.gov/22124482/
68. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The Third Revolution in Sequencing Technology. Trends Genet [Internet]. 2018 Sep 1 [cited 2022 Mar 10];34(9):666–81. Available from: https://pubmed.ncbi.nlm.nih.gov/29941292/
69. Xiao T, Zhou W. The third generation sequencing: the advanced approach to genetic diseases. Transl Pediatr [Internet]. 2020 Apr 1 [cited 2022 Mar 10];9(2):163. Available from: /pmc/articles/PMC7237973/
70. High Performance Long Read Assay Enables Contiguous Data up to 10Kb on Existing Illumina Platforms [Internet]. [cited 2022 Mar 10]. Available from: https://emea.illumina.com/science/genomics-research/articles/infinity-high-performance-long-read-assay.html?scid=2022-303SMP5438&catt=JPM22 - Social Posts
71. Sequence File Formats | FASTQ & BCL formats for Illumina sequencing [Internet]. [cited 2022 Mar 10]. Available from: https://emea.illumina.com/informatics/sequencing-data-analysis/sequence-file-formats.html
72. 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. Surv Summ [Internet]. Available from: https://academic.oup.com/nar/article/38/6/1767/3112533
73. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinforma Appl NOTE. 2009;25(16):2078–9.
74. samtools/hts-specs: Specifications of SAM/BAM and related high-throughput sequencing file formats [Internet]. [cited 2022 Mar 14]. Available from: https://github.com/samtools/hts-specs
75. Germline short variant discovery (SNPs + Indels) – GATK [Internet]. [cited 2022 Mar 15]. Available from: https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-
76. Depristo MA, Banks E, Poplin R, Garimella K V., Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011 435 [Internet]. 2011 Apr 10 [cited 2022 Mar 14];43(5):491–8. Available from: https://www.nature.com/articles/ng.806
77. de Sá PHCG, Guimarães LC, das Graças DA, de Oliveira Veras AA, Barh D, Azevedo V, et al. Next-Generation Sequencing and Data Analysis: Strategies, Tools, Pipelines and Protocols. Omi Technol Bio-engineering Towar Improv Qual Life. 2018 Jan 1;1:191–207.
78. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct;467(7319):1061–73.
79. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nat 2020 5817809 [Internet]. 2020 May 27 [cited 2022 Mar 14];581(7809):434–43. Available from: https://www.nature.com/articles/s41586-020-2308-7
80. Fattahi Z, Beheshtian M, Mohseni M, Poustchi H, Sellars E, Nezhadi SH, et al. Iranome: A catalog of genomic variations in the Iranian population. Hum Mutat. 2019 Nov;40(11):1968–84.
81. Scott EM, Halees A, Itan Y, Spencer EG, He Y, Azab MA, et al. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery [Internet]. Available

from: http://www.nature.com/authors/editorial_policies/license.html#terms
82. Masica DL, Karchin R. Towards Increasing the Clinical Relevance of In Silico Methods to Predict Pathogenic Missense Variants. PLOS Comput Biol [Internet]. 2016 May 1 [cited 2022 Mar 14];12(5):e1004725. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004725
83. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med [Internet]. 2020 Dec 1 [cited 2022 Mar 14];12(1):1–8. Available from: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-00803-9
84. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, et al. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004 Jan;32(Database issue):D258-61.
85. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. 2012.
86. Li H, Barrett J. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. 2011;27(21):2987–93. Available from: http://samtools.sourceforge.net
87. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol [Internet]. 2018 Nov 1 [cited 2022 Apr 1];36(10):983. Available from: https://pubmed.ncbi.nlm.nih.gov/30247488/
88. Hwang KB, Lee IH, Li H, Won DG, Hernandez-Ferrer C, Negron JA, et al. Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings. Sci Rep [Internet]. 2019 Dec 1 [cited 2022 Apr 1];9(1). Available from: https://pubmed.ncbi.nlm.nih.gov/30824715/
89. Zarate S, Carroll A, Mahmoud M, Krasheninina O, Jun G, Salerno WJ, et al. Parliament2: Accurate structural variant calling at scale. Gigascience [Internet]. 2020 Nov 30 [cited 2022 Apr 3];9(12):1–9. Available from: https://academic.oup.com/gigascience/article/9/12/giaa145/6042728
90. Smith DR. The design of divide and conquer algorithms. Sci Comput Program [Internet]. 1985;5:37–58. Available from: https://www.sciencedirect.com/science/article/pii/0167642385900036
91. Blahut RE. Fast Algorithms for Signal Processing. Cambridge University Press; 2010.
92. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data [Internet]. [cited 2022 Apr 23]. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
93. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. 2010;26(5):589–95. Available from: https://academic.oup.com/bioinformatics/article/26/5/589/211735
94. broadinstitute/picard: A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF. [Internet]. [cited 2022 Apr 23]. Available from: https://github.com/broadinstitute/picard
95. Van der Auwera G, O'Connor B, Safari an OMC. Using Docker, GATK, and WDL in Terra. In: Genomics in the Cloud [Internet]. 2020 [cited 2022 Mar 15]. p. 300. Available from: https://www.oreilly.com/library/view/genomics-in-the/9781491975183/
96. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal snp and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018 Nov 1;36(10):983.
97. Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, et al. Best practices for benchmarking germline small-variant calls in human genomes. Nat Biotechnol [Internet]. 2019 May 1 [cited 2022 Mar 15];37(5):555–60. Available from: https://pubmed.ncbi.nlm.nih.gov/30858580/
98. Illumina/hap.py: Haplotype VCF comparison tools [Internet]. [cited 2022 Mar 15]. Available from: https://github.com/Illumina/hap.py
99. Desvignes JP, Bartoli M, Delague V, Krahn M, Miltgen M, Béroud C, et al. VarAFT: a variant annotation and filtration system for human next generation sequencing data. Nucleic Acids Res [Internet]. 2018 Jul 2 [cited 2022 Mar 15];46(W1):W545–53. Available from: https://academic.oup.com/nar/article/46/W1/W545/5025894
100. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res [Internet]. 2010 Jul 3 [cited 2022 Mar

15];38(16). Available from: https://pubmed.ncbi.nlm.nih.gov/20601685/
101. Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc. 2015 Oct 29;10(10):1556–66.
102. Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, et al. Copy number variation detection and genotyping from exome sequence data. Genome Res [Internet]. 2012 Aug [cited 2022 Mar 15];22(8):1525–32. Available from: https://pubmed.ncbi.nlm.nih.gov/22585873/
103. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods [Internet]. 2008 Jul [cited 2022 Mar 15];5(7):621–8. Available from: https://pubmed.ncbi.nlm.nih.gov/18516045/
104. Table Browser [Internet]. [cited 2022 Mar 15]. Available from: https://genome.ucsc.edu/cgi-bin/hgTables
105. Ronique Geoffroy V, Herenger Y, Kress A, Stoetzel C, Lie Piton A, Lè Ne Dollfus H, et al. AnnotSV: an integrated tool for structural variations annotation. Available from: http://lbgi.fr/AnnotSV/.
106. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM. Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res [Internet]. 2007 Jul 1 [cited 2022 Mar 15];35(suppl_2):W71–4. Available from: https://academic.oup.com/nar/article/35/suppl_2/W71/2922185
107. OligoAnalyzer Tool - primer analysis | IDT [Internet]. [cited 2022 Mar 15]. Available from: https://www.idtdna.com/pages/tools/oligoanalyzer
108. UCSC In-Silico PCR [Internet]. [cited 2022 Mar 15]. Available from: https://genome.ucsc.edu/cgi-bin/hgPcr
109. Geoffroy V, Guignard T, Kress A, Gaillard JB, Solli-Nowlan T, Schalk A, et al. AnnotSV and knotAnnotSV: a web server for human structural variations annotations, ranking and analysis. Nucleic Acids Res [Internet]. 2021 Jul 2 [cited 2022 Apr 3];49(W1):W21–8. Available from: https://academic.oup.com/nar/article/49/W1/W21/6281473
110. Ewans LJ, Schofield D, Shrestha R, Zhu Y, Gayevskiy V, Ying K, et al. Whole-exome sequencing reanalysis at 12 months boosts diagnosis and is cost-effective when applied early in Mendelian disorders. Genet Med [Internet]. 2018;20:1564–74. Available from: http://www.ensembl.org/info/docs/tools/vep/
111. Clark MM, Stark Z, Farnaes L, Tan TY, White SM, Dimmock D, et al. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. npj Genomic Med [Internet]. 2018;3:16. Available from: www.nature.com/npjgenmed
112. Van der 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. Curr Protoc Bioinforma [Internet]. 2013 Oct 1 [cited 2022 Apr 21];43(1):11.10.1-11.10.33. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/0471250953.bi1110s43
113. Hintzsche JD, Robinson WA, Tan AC. A Survey of Computational Tools to Analyze and Interpret Whole Exome Sequencing Data. Int J Genomics. 2016;2016.
114. Yang R, Van Etten JL, Dehm SM. Indel detection from DNA and RNA sequencing data with transIndel. BMC Genomics [Internet]. 2018 Apr 19 [cited 2022 Apr 1];19(1). Available from: https://pubmed.ncbi.nlm.nih.gov/29673323/
115. Glusman G, Rodrigues Alves Margarido G, Aganezov S, Mainzer LS, Kendig KI, Baheti S, et al. Sentieon DNASeq Variant Calling Workflow Demonstrates Strong Computational Performance and Accuracy. 2019;10. Available from: www.frontiersin.org
116. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci data [Internet]. 2016 Jun 7 [cited 2022 Apr 1];3. Available from: https://pubmed.ncbi.nlm.nih.gov/27271295/
117. Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. An integrated map of genetic variation from 1,092 human genomes. Nature [Internet]. 2012 Nov 1 [cited 2022 Apr 1];491(7422):56–65. Available from: https://pubmed.ncbi.nlm.nih.gov/23128226/
118. Langmore J, De F, Vega L, Hilbush B, Trigg L, Littin R, et al. Quality Control and Pre-Qualification of NGS Libraries Made from Clinical Samples Towards Clinical Grade Genomes with Joint Bayesian Variant Identification Which RNA-Seq Processing Algorithm Should I Pick? A Comparison of RNA-Seq Pipelines Based on Experimental Design, Sample

Properties and Sequencing Technology Different Sorts for Different Folks: The Importance of Technological Diversity in a Cell Sorting Facility. Vol. 24, ABRF 2013 POSTER ABSTRACTS S46 JOURNAL OF BIOMOLECULAR TECHNIQUES. 2013.
119. Lee S, Won S, Kim YJ, Kim Y, Kim BJ, Park T. Rare variant association test with multiple phenotypes. Genet Epidemiol [Internet]. 2017 Apr 1 [cited 2022 Apr 1];41(3):198–209. Available from: https://pubmed.ncbi.nlm.nih.gov/28039885/
120. Guo Y, Ye F, Sheng Q, Clark T, Samuels DC. Three-stage quality control strategies for DNA re-sequencing data. Brief Bioinform [Internet]. 2014 Aug 2 [cited 2022 Apr 1];15(6):879–89. Available from: https://pubmed.ncbi.nlm.nih.gov/24067931/
121. Meynert AM, Bicknell LS, Hurles ME, Jackson AP, Taylor MS. Quantifying single nucleotide variant detection sensitivity in exome sequencing. BMC Bioinformatics [Internet]. 2013 Jun 18 [cited 2022 Apr 1];14(1). Available from: https://pubmed.ncbi.nlm.nih.gov/23773188/
122. Becker F, Reid CA, Hallmann K, Tae HS, Phillips AM, Teodorescu G, et al. Functional variants in HCN4 and CACNA1H may contribute to genetic generalized epilepsy. Epilepsia Open [Internet]. 2017 Sep 1 [cited 2022 Apr 1];2(3):334. Available from: /pmc/articles/PMC5862120/
123. Scholl UI, Stölting G, Nelson-Williams C, Vichot AA, Choi M, Loring E, et al. Recurrent gain of function mutation in calcium channel CACNA1H causes early-onset hypertension with primary aldosteronism. Elife [Internet]. 2015 Apr 24 [cited 2022 Apr 3];4:e06315. Available from: https://pubmed.ncbi.nlm.nih.gov/25907736/
124. Otomo N, Mizumoto S, Lu HF, Takeda K, Campos-Xavier B, Mittaz-Crettol L, et al. Identification of novel LFNG mutations in spondylocostal dysostosis. J Hum Genet [Internet]. 2019 Mar 1 [cited 2022 Apr 3];64(3):261–4. Available from: https://pubmed.ncbi.nlm.nih.gov/30531807/
125. Homo sapiens claudin 16 (CLDN16), mRNA - Nucleotide - NCBI [Internet]. [cited 2022 Apr 3]. Available from: https://www.ncbi.nlm.nih.gov/nuccore/NM_006580.3
126. Homo sapiens claudin 16 (CLDN16), transcript variant 1, mRNA - Nucleotide - NCBI [Internet]. [cited 2022 Apr 3]. Available from: https://www.ncbi.nlm.nih.gov/nuccore/NM_006580.4
127. Storey H, Savige J, Sivakumar V, Abbs S, Flinter FA. COL4A3/COL4A4 mutations and features in individuals with autosomal recessive alport syndrome. J Am Soc Nephrol. 2013;24(12):1945–54.
128. Butkowski RJ, Wieslander J, Kleppel M, Michael AF, Fish AJ. Basement membrane collagen in the kidney: regional localization of novel chains related to collagen IV. Kidney Int [Internet]. 1989 [cited 2022 Apr 3];35(5). Available from: https://pubmed.ncbi.nlm.nih.gov/2671463/

Appendix 2. Voluntary Consent Form

AYDINLATILMIŞ ONAM FORMU

I-Yapılan Araştırmayla İlgili Özet Bilgilendirme Metni:

Bu çalışmanın adı **"Tanısız ve Nadir Hastalıklarda Yeniden Ekzom Dizileme Analizi ile Moleküler Tanı"**dır. Amacı tüm ekzom dizi analizi yapılmış fakat hastalığa uygun varyant bulunmamış hasta verilerinin tekrar incelenerek moleküler tanının koyulabilmesidir.

Yapılacak işlem: Çocuğunuzun veya yasal velisi olduğunuz hastanızın tıbbi genetik veya pediyatrik tıbbi genetik kliniğince takibi devam etmektedir. Bu araştırma için daha önce yapılmış olan tüm ekzom dizileme verisinin bizlerle paylaşılması gerekli.

Çalışmaya dahil edilmek için klinik özellikler ve aile öykünüz göz önünde bulundurulacaktır. Bu çalışmada genetik olarak, DNA'mızda yer alan şifrenin protein kodladığı bilinen kısmının tamamı incelenmektedir. Çalışma, genlerimizde yer alan varyantların tespiti, tiplendirilmesi ve istatistik çalışmalarla elde edilen bulguların anlamlandırılmasına dayanmaktadır. Çalışma Acıbadem Mehmet Ali Aydınlar Üniversitesi Kerem Aydınlar Kampüsü bünyesinde gerçekleştirilecektir.

Varyant bulunması halinde doğruluğunun tespiti için hastadan en az 5 mililitre kan alınacaktır. Bulunan varyantın ailesel kalıtılmış olma ihtimalinde ebeveynlerden ve kardeşlerden de aynı miktarda kan alınması gerekebilir. Bu işlemin bilinen herhangi bir tıbbi riski yoktur. Kan alınması için iğne batırılan bölgede biraz acı olabilir, 1-2 gün sürecek morarma, şişme görülebilir. Tamamı tıbbi genetik hekiminizin gözetiminde izleme alınacaktır.

Araştırma toplam 12 ay olarak planlanmış olup, size bu süre içerisinde herhangi bir sonuç raporu düzenlenmeyecektir. Araştırma sonuçlarınız da isminizle değil, anonim olarak numaralandırılarak arşivlenecektir. Tarafınıza ait hiçbir kişisel veya tıbbi bilgi 3. şahıslara aktarılmayacaktır. ANCAK;

Tıbbi Genetik polikliniğindeki aile ağacı incelemeniz ve klinik değerlendirmeniz sırasında;

- Ailenizde taşınıyor olması olası farklı hastalıklar konusunda da ön tanılar oluşabilir. Bunlar açısından talep ederseniz test edilen bireylerde -tespit edilirse- taşıyıcılık durumları paylaşılır ve genetik danışması verilir.
- Tesadüfi olarak bu çalışmanın ayrıntılı analiz kapasitesinin doğası gereği bugüne kadar ailenizde hiç görülmeyen farklı hastalıklar için mutasyonlar tespit edilebilir, yasal, etik kurallar çerçevesinde bunlarla ilgili bilgi paylaşılabilir.
- Çalışma tamamlanıp, kesin veriler elde edilip literatüre kazandırıldıktan sonra çocuğunuzda tespit edilen veriler konusunda tıbbi veriler paylaşılıp, genetik danışmanız verilecektir.

Çalışmaya katılmayı kabul etmezseniz, çocuğunuz herhangi bir nedenle kan vermek ve katılmak istemezse, takip ve tedavinize aynen devam edeceksiniz. Katılmayı kabul etmiş olsanız bile, istediğiniz zaman ayrılmayı talep edebilirsiniz, bu kararınız da tıbbi takip ve tedavinizi hiçbir şekilde etkilemeyecektir.

Aynı şekilde olur vermiş olsanız bile bilimsel nedenlerle çalışma dışı bırakılma olasılığınız mevcuttur.

Sağlıklı Kontrol Gerekliliği:

Çalışmada elde edilecek verilerin sağlıklı değerlendirilmesi ve bilimsel niteliklerinin artırılması için sağlıklı kontroller eşliğinde çalışılmalı ve bulgular karşılaştırılmalıdır.ç Bununla birlikte anne ve babanın da analizlere dahil edilmesi kalıtımsal özelliklerin test edilmesine olanak sağlayıp, daha kesin veriler elde edilmesine yardımcı olacaktır.

Çalışmaya katılmış olmakla herhangi bir ödül vb maddi kazanç elde etmeyeceksiniz. Aynı zamanda araştırmayı yapan hiçbir araştırıcının da çalışmadan maddi kazancı yoktur.

Appendix 3. Quality control results of all data.

Patient ID	Sequencing Method	Data Type	Is the raw data has high-quality?	Is the phenotype information of patient enough?
WES001	WES	FASTQ	Yes	Yes
WES002	WES	FASTQ	Yes	Yes
WES003	WES	BAM	Yes	Yes
WES004	WES	FASTQ	Yes	Yes
WES005	WES	FASTQ	Yes	Yes
WES006	WES	FASTQ	Yes	Yes
WES007	WES	FASTQ	Yes	Yes
WES008	WES	FASTQ	Yes	Yes
WES009	WES	BAM	Yes	Yes
WES010	WES	FASTQ	Yes	Yes
WES011	WES	FASTQ	Yes	Yes
WES012	WES	FASTQ	Yes	Yes
WES013	WES	FASTQ	Yes	Yes
WES014	WES	FASTQ	Yes	Yes
WES015	WES	FASTQ	Yes	Yes
WES016	WES	FASTQ	Yes	Yes
WES017	WES	FASTQ	Yes	Yes
WES018	WES	FASTQ	Yes	Yes
WES019	WES	FASTQ	Yes	Yes
WES020	WES	FASTQ	Yes	Yes
WES021	WES	FASTQ	Yes	Yes
WES022	WES	FASTQ	Yes	Yes
WES023	WES	FASTQ	Yes	No
WES024	WES	VCF	NA	Yes
WES025	WES	FASTQ	Yes	No
WES026	WES	FASTQ	Yes	No
WES027	WES	VCF	NA	Yes
WES028	WES	BAM	NA	Yes
WES029	WES	VCF	NA	Yes
WES030	WES	FASTQ	Yes	No
WES031	WES	VCF	NA	Yes
WES032	WES	BAM	NA	Yes
WES033	WES	FASTQ	No	Yes
WES034	WES	VCF	NA	Yes
WES035	WES	BAM	NA	Yes
WES036	WES	FASTQ	Yes	No
WES037	WES	BAM	NA	Yes
WES038	WES	VCF	NA	Yes
WES039	WES	VCF	NA	Yes
WGS001	WGS	FASTQ	Yes	Yes
WGS002	WGS	FASTQ	Yes	Yes
WGS003	WGS	FASTQ	Yes	Yes
WGS004	WGS	XLSX	NA	Yes
WGS005	WGS	FASTQ	Yes	No
WGS006	WGS	VCF	NA	Yes

Appendix 4. List of variants that found patient WES002.

Chíomosom	Position	Refeíence	Alteínative	Gene	Exonic Ïunc.	Zygosity
chí17	80198415	C	**I’**	*CARD14*	nonsynonymous SNV	het
chí7	72728896	C	**I’**	*I’YW1B*	stop_gained	het
chí11	47642411	G	A	*MI’CH2*	stop_gained	het
chí14	31483548	G	A	*GPR33*	stop_gained	hom
chí7	1.43E+08	G	**I’**	*PRSS1*	stop_gained	het
chí21	10569701	C	**I’**	*I’PI’E*	stop_gained	het
chí21	10569462	C	**I’**	*I’PI’E*	stop_gained	het
chí11	47642399	A	G	*MI’CH2*	nonsynonymous SNV	het
chí13	1.13E+08	G	A	*ADPRHL1*	stop_gained	het
chí18	63712604	G	**I’**	*SERPINB11*	stop_gained	hom
chíX	1.53E+08	G	C	*CSAG1*	stop_gained	hom
chí17	2700818	G	A	*CLUH*	nonsynonymous SNV	het
chí11	47642396	C	**I’**	*MI’CH2*	nonsynonymous SNV	het

Appendix 5. List of variants that found patient WES005.

Chíom	Position	Ref	Alt	Gene	Exonic Func.	Zygosity
chí7	24702830	A	G	*GSDME*	nonsynonymous SNV	Het
chí4	71440709	G	A	*SLC4A4*	nonsynonymous SNV	Het
chí5	158731100	C	Ľ	*EBF1*	nonsynonymous SNV	Het
chí3	127619139	C	Ľ	*MCM2*	nonsynonymous SNV	Het
chí2	140534066	C	Ľ	*LRP1B*	nonsynonymous SNV	Het
chí12	20880898	Ľ	G	*SLCO1B3-SLCO1B7*	nonsynonymous SNV	Het
chí1	99288109	C	Ľ	*PLPPR4*	nonsynonymous SNV	Het
chí16	88841094	G	A	*GALNS*	nonsynonymous SNV	Het
chí6	46835512	G	A	*MEP1A*	nonsynonymous SNV	Het
chíX	119470922	C	Ľ	*SLC25A5*	nonsynonymous SNV	Het

Appendix 5. List of variants that found patient WES018.

Chí	Staít	Ref	Alt	Gene	Exonic Func.	Zygosity
17	74866471	L'	C	*FDXR*	nonsynonymous SNV	hom
17	74866908	C	L'	*FDXR*	nonsynonymous SNV	hom
15	51478261	L'	-	*DMXL2*	.	hom
X	10213675	L'	C	*CLCN4*	.	hom
17	74862956	G	A	*FDXR*	.	hom
X	1.11E+08	L'	C	*PAK3*	.	hom
18	10677885	A	-	*PIEZO2*	.	hom
19	53882668	A	L'	*PRKCG*	.	het
19	53882673	A	L'	*PRKCG*	.	het
1	1.51E+08	A	G	*PRUNE1*	nonsynonymous SNV	het
7	56019681	C	L'	*PSPH*	nonsynonymous SNV	het
7	56019599	C	-	*PSPH*	.	het
7	56021238	-	A	*PSPH*	.	het
3	47410304	C	L'	*PI'PN23*	stopgain	het
3	47412144	A	C	*PI'PN23*	nonsynonymous SNV	het

Appendix 6. List of variants that found patient WES021.

Chíomosom	Position	Refeíence	Alteínative	Gene	Exonic Ĩunc.	Zygosity
4	183705029	C	I’	I’RAPPC11	nonsynonymous SNV	het
14	45148886	G	A	*FANCM*	nonsynonymous SNV	het
17	21703417	C	I’	*KCNJ18*	nonsynonymous SNV	het
12	111347567	G	A	*CUX2*	nonsynonymous SNV	het
17	21703340	C	I’	*KCNJ18*	nonsynonymous SNV	het
1	20655747	C	I’	*DDOS/’*	nonsynonymous SNV	het
11	71444008	-	A	*DHCR7*	fíameshift inseítion	het
17	49973822	I’	A	*DLX4*	nonsynonymous SNV	het
14	45148886	G	A	*FANCM*	nonsynonymous SNV	het
20	35434589	C	A	*GDF5*	nonsynonymous SNV	hom
20	3221849	C	I’	*II’PA*	nonsynonymous SNV	het
15	40407806	I’I’	-	*IVD*	fíameshift deletion	het
8	142664633	I’	C	*JRK*	nonsynonymous SNV	het
17	21702953	C	A	*KCNJ18*	nonsynonymous SNV	hom
17	21703340	C	I’	*KCNJ18*	nonsynonymous SNV	het
17	21703362	G	C	*KCNJ18*	nonsynonymous SNV	het
17	21703417	C	I’	*KCNJ18*	nonsynonymous SNV	het
7	2513247	-	GAI’G	*LFNG*	fíameshift inseítion	het
5	56882802	C	I’	*MAP3K1*	nonsynonymous SNV	het
9	35793912	G	A	*NPR2*	nonsynonymous SNV	het
11	31793332	A	-	*PAX6*	fíameshift deletion	het
2	231738054	C	I’	*PDE6D*	nonsynonymous SNV	het
X	78125810	G	A	*PGK1*	nonsynonymous SNV	hom
19	39423417	I’	C	*PLEKHG2*	nonsynonymous SNV	het
1	11960766	G	A	*PLOD1*	nonsynonymous SNV	het
10	120355038	I’	A	*RPL21*	nonsynonymous SNV	hom
3	38714049	A	I’	*SCN10A*	nonsynonymous SNV	het
3	189886509	G	A	*I’P63*	nonsynonymous SNV	het
4	183705029	C	I’	*I’RAPPC11*	nonsynonymous SNV	het
Y	9337985	A	C	*I’SPY1;I’SPY10;I’SPY3;I’SPY4;/’SPY8*	nonsynonymous SNV	het
2	178769939	G	A	*I’I’N*	nonsynonymous SNV	het
2	178775079	C	I’	*I’I’N*	nonsynonymous SNV	het

9. CURRICULUM VITAE