UTILITY OF RESEQUENCING AND REANALYSIS FOR UNSOLVED RARE DISEASES

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

ÖMER FARUK YAZAR

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

NOVEMBER 2022

Approval of the thesis:

UTILITY OF RESEQUENCING AND REANALYSIS FOR UNSOLVED RARE DISEASES

Submitted by ÖMER FARUK YAZAR in partial fulfillment of the requirements for the degree of **Master of Science in Health Informatics, Middle East Technical University** by,

Prof. Dr. Banu Günel Kılıç
Dean, **Graduate School of Informatics** _____

Assoc. Prof. Dr. Yeşim Aydın Son
Head of Department, **Health Informatics** _____

Assoc. Prof. Dr. Yeşim Aydın Son
Supervisor, **Health Informatics, METU** _____

Prof. Dr. Fatih Süheyl Ezgü
Co-Supervisor, **Pediatric Metabolism Dept., Gazi University** _____

Examining Committee Members:

Asst. Prof. Aybar Can Acar
Health Informatics, METU _____

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

Assoc. Prof. Dr. Ceren Sucularlı
Health Informatics, Hacettepe University _____

Date: *29.11.2022*

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Ömer Faruk Yazar

Signature : _____

ABSTRACT

UTILITY OF RESEQUENCING AND REANALYSIS FOR UNSOLVED RARE DISEASES

Yazar, Ömer Faruk

MSc., Department of Health Informatics

Supervisor: Assoc. Prof. Dr. Yeşim Aydın Son

Co-Supervisor: Prof. Dr. Fatih Süheyl Ezgü

November 2022, 38 pages

Many unsolved rare diseases lack etiological information at the molecular level. In many cases mode of inheritance is may not be identified correctly, and even if a variant is identified, new functional studies are required to establish genotype-phenotype associations. Factors hamper rare disease research, such as the low number of patients, the absence of biomarkers, and the lack of effective diagnostics. The need for a joint effort between clinicians and researchers has been increasing. Enhancements in bioinformatics algorithms, new literature, and published cases in databases enable rare disease research to reveal new variants through manual curation or automated data mining. Reanalysis of exome sequencing data for unsolved rare disease cases has the potential to reveal novel gene and disease associations due to recent developments in bioinformatics tools. Additionally, recent technological advancements in sequencing technologies have increased the quality of raw exome data, increasing the success rate for variant discovery. Here, we present the importance of reanalyzing older sequencing data with recent algorithms and literature as well as resequencing DNA samples with the latest instruments for challenging rare diseases in a case study.

Keywords: rare disease, next-generation sequencing, reanalysis, resequencing

ÖZ

ÇÖZÜLMEMİŞ NADİR HASTALIK VAKALARINDA YENİDEN DİZİLEME VE ANALİZİN ÖNEMİ

Yazar, Ömer Faruk

Yüksek Lisans, Sağlık Bilişimi Bölümü

Tez Yöneticisi: Doç. Dr. Yeşim Aydın Son

Eş-danışman: Prof. Dr. Fatih Süheyl Ezgü

Kasım 2022, 38 sayfa

Çözülmemiş nadir hastalıkların çoğunlukla moleküler düzeyde etiyolojisi bilinmemektedir. Öncelikle kalıtım modelinin doğru tanımlanmasına ve sonrasında varyant tanımlansa bile, genotip-fenotip ilişkilerini kurmak için yeni fonksiyonel çalışmalara ihtiyaç duyulmaktadır. Düşük hasta sayısı, biyobelirteçlerin yokluğu ve etkili tanı araçlarının eksikliği gibi faktörler nadir hastalık araştırmalarını engelleyen faktörlerdir. Hekimler ve araştırmacılar arasındaki ortak çalışmanın gerekliliği artmaktadır. Biyoenformatik algoritmalarındaki gelişmeler, veri tabanlarında yayınlanan yeni vakalar ve gelişen literatür, manuel kürasyon veya otomatik veri madenciliği yoluyla nadir hastalık araştırmalarında yeni varyantların ortaya çıkarılmasını sağlamaktadır. Çözülmemiş nadir hastalık vakaları için ekzom dizileme verilerinin yeniden analizi, biyoenformatik araçlarındaki son gelişmeler sayesinde yeni gen ve hastalık ilişkilerini ortaya çıkarma potansiyeline sahiptir. Ek olarak, dizileme teknolojilerindeki yeni teknik gelişmeler, ham ekzom verilerinin kalitesini ve dolayısıyla varyant keşfi için başarı oranını arttırmaktadır. Bu çalışmada, dizileme verilerini son algoritmalarla yeniden analiz edilmesinin yanında DNA örneklerinin yeni teknolojilerle yeniden dizilenmesinin çözülmemiş nadir hastalık vaka çalışmasındaki önemini sunuyoruz.

Anahtar Sözcükler: nadir hastalık, yeni nesil sekanslama, tekrar dizileme, tekrar analiz

To My Family

ACKNOWLEDGMENTS

First, I would like to express my sincere gratitude to my supervisor, Assoc. Prof. Dr. Yesim Aydın Son for her patience, support, and motivation.
I am also grateful to my co-advisor, Prof. Dr. Fatih Süheyl Ezgü, for providing the necessary environment and supporting the thesis research. Thanks should also go to Ezgü Lab members for their efforts and collaborations.
Most of all I would like to thank my wife who kept me coordinated, shared my stress, and supported me with her love for all the late nights and early mornings during the thesis process.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CHIPSEQ	Chromatin immunoprecipitation sequencing
CNV	Copy Number Variation
DDI	Drug-Drug Interaction
DNA	Deoxyribonucleic Acid
D.P.I	Drug-Protein Interaction
EURORDIS	Rare Diseases Europe
GRC	Genome Reference Consortium
HRG	Human Reference Genome
INDEL	Insertion and deletion
NGS	Next Generation Sequencing
NORD	The National Organization for Rare Disorders
RNA	Ribonucleic Acid
SMA	Spinal Muscular Atrophy
SNV	Single Nucleotide Variation
SV	Structural Variation
WES	Whole Exome Sequence
WGS	Whole Genome Sequence

CHAPTER 1

INTRODUCTION

1.1 Rare Diseases

1.1.1 Definition

There are approximately 400 million individuals affected by rare diseases worldwide, 80% of which are caused by genetic backgrounds (GlobalGenes, n.d.). Even though there is no globally accepted definition of rare disease, different countries or communities define it based on the ratio of affected individuals to unaffected. In Europe and Turkey, the disease is accepted as rare if it has seen fewer than five people out of 10000. In the USA, a rare disease is defined as a condition seen in less than 200000 (Bax B. E., 2021). In Japan, it is four individuals out of 10000 people, or less than 50000 individuals in the Japanese territory (Orphanet, n.d.). It is estimated that there are between 5000 to 8000 rare diseases, and 3 to 4 new conditions are identified each year. According to a report published by TUSEB in 2019, in Turkey, 5 million rare disease patients are estimated, which equals a prevalence of 38 out of 100000 (Satman et al., 2019). Currently, only 5 percent of rare diseases have a cure. Rare diseases are mainly categorized as childhood diseases, as survivor rates are meager after five years or older (Kaufman et al., 2018).

Basic knowledge, like the causative of the disease and pathophysiology, is limited or missing for most rare diseases. Limitations such as the low number of patients, unavailability of biomarkers, and lack of efficient diagnostics are among the significant barriers to rare disease research. Even identifying the causative variants can significantly impact the patient's quality of life as that information can direct the clinicians to potential interventions or drug therapies to ease the symptoms and lead to new research on diagnostics and therapeutics.

Next-generation sequencing (NGS) technologies have increased the rate of identification of causative variants of rare diseases. Novel disease-associated genes are often identified by a functional link between a candidate gene and the patient's phenotype. Many tools

exist to examine relevant variants by referencing previously known information about their biological functions and inferring potential effects based on their genomic context.

1.1.2 Diagnostic and Therapeutic Challenges in Rare Diseases

Rare diseases involve tens of millions of patients distributed across the globe. Ultra-rare diseases are one of the most complex classes since there may be only up to ten patients in the same geographical area. The main challenge in diagnosis is the lack of information from clinicians and caregivers. Traditional medicine is mainly focused on treating the mass, yet millions of affected people with rare diseases worldwide are out of scope for some clinicians. It is estimated that 7000 rare diseases have been defined so far (Haendel et al., 2019). If we add this disease information to existing other thousands of common diseases, the human brain can be overloaded with such information. Clinicians cannot be expected to know every disease plus rare diseases. Physicians mainly increase their information about the high frequency of occurring diseases. Most rare diseases involve multiple symptoms that might be confused with other common diseases, such as reoccurring bacterial infections with a high fever. Relating symptoms in several parts of the body and resulting in a single disease diagnosis is very complex. There might be a few physicians in an area who are specialized in rare diseases. Therefore, joining a physician and a patient is challenging due to the increased number of rare disease patients.

On the other hand, most of these diseases require special tests, which are hard to find in every medical center. Besides, there is still lacking tests for metabolic and genetic disorders. Even if the test is available, special permissions and expensive methods are needed to conduct tests and studies. Some rare disorders are initially non-symptomatic; lifespan is very short once symptoms occur. This short diagnosis interval is very crucial for vitality. Matching these patients with their specialized physicians and conducting special tests in such a short period is almost impossible. Standard diagnostic procedures are not currently present. A lack of standardized diagnosis might cause misdiagnosis and mislead physicians.

In the case of rare disease studies, the limited number of patients and sources are one of the holdbacks for researchers. There are initiatives such as The National Organization for Rare Disorders (NORD) in the USA and Rare Diseases Europe (EURORDIS) in the EU region that are taking responsibility in their regions and around the world to generate communities for specific rare disorders in case of reliable and applicable diagnostic procedures and therapeutic solutions. Gathering patients and their matching physicians for an increasing number of subjects in research has a crucial role in shortening diagnostic periods, correcting misdiagnosis, raising awareness, and increasing the budget of studies.

Once the diagnosis is achieved, treatment is the next big challenge for patients and their caregivers. Although there are some treatments for rare diseases, it is hard to decide whether treatment is appropriate for that patient because each rare disease might cause different symptoms in different patients. For example, Spinal Muscular Atrophy (SMA) is a rare disease with several types, such as SMA-Type1/Type2/Type3/Type4 (ultra-rare).

Even though genetic therapy is available, it can only be applied to infants before age 2 with type 1 SMA, which is also very expensive. Lack of awareness of this issue is a conflict that is causing danger for other SMA patients seeking treatment. Most drugs for rare diseases intend to ease or strangle symptoms in rare diseases. Once the damage is done, it could be impossible to revert its consequences. Surgical options are very limited for most rare diseases.

Geographical areas, cultural biases, economic constraints, incapable health politics, and lack of awareness are also other barriers against rare disease treatments. The scarcity of experts in a geographical area for diagnosing patients is one of the main problems. Treatment options are even more scarce than experts in some geographical areas. A drug for a specific disorder could be thousands of kilometers away, and safe transportation of such a drug could be more challenging than finding it. Gender, religion, age, skin color, or economic status of the expert could be disadvantages for the caregiver's beliefs according to their culture. Lack of knowledge and access to information in remote areas also affect the treatment and diagnosis of patients.

As seen in the SMA example, a drug might cost millions of dollars, making it impossible for a caregiver to fund it quickly. The rare disease involves few patients, and drug manufacturers are businesses running for profit. The life insurance system is constructed for the mass, not the rare. Insurances cover very few rare treatments. Countries and economic communities neglect rare disease treatments, but sometimes reimbursement of a cheaper drug for an extended period costs more than intended. Once all diseases were rare until its frequent in communities. Bottle-neck effect of genetic selection among isolated communities increases the prevalence of diseases such as Familial Mediterranean Fever. Japan's rare disease status is changing since the frequency is higher now (Migita et al., 2016).

Health authorities might make different decisions when it comes to the approval of drugs. A drug could be legal in one region, whereas it could even be illegal in another. Legalization of drug usage mostly depends on the approval of its producing company or patients' request. Slow approval progression of a critical drug can cause a patient to miss their chance to be treated before the effects are permanent.

1.1.3 Case Study: Patterson-Lowry Rhizomelic Dysplasia

Caroline Patterson and R. Brian Lowry initially defined Patterson-Lowry rhizomelic dysplasia. Patterson-Lowry rhizomelic dysplasia is a rare disorder that presents clinical features such as a short upper arm (rhizomelic shortness of humeri) and upper leg bones (shortness of limb), femoral neck, deformed humeral head, some cases decrease of the angle of innominate bone (coxa vara deformity) (Patterson & Lowry, 1975), short finger bones (brachydactyly) (Williams et al., 1995), respiratory disorder (Kamoda et al., 2001), and motor and mental retardation (Damar et al., 2014).

The first ever case was an adult male. A deformed humeral head defined his disorder, along with shortness of femora, short humeri, short neck, depressed head, coxa vara deformation, dysmorphic femoral head, hollow back (lumbar lordosis), depressed skull base (platybasia). The parents of the patient had average intelligence and height. His intelligence was also as expected, but he had short stature. Family history suggested either recessive inheritance or sporadic in a word de novo inheritance (Patterson & Lowry, 1975).

The second case was a male child. In addition to the first case, the disorder of this boy was identified by inconvenient cell division of humeral metaphysis, short humeri and short finger bones, depressed spinal bones (platyspondyly), and shortness of metacarpus (brachymetacarpia). This case also demonstrates sporadic inheritance (Williams et al., 1995).

The third case was a male infant. In addition to the first two cases, this infant had respiratory distress, abnormally bigger liver size (hepatomegaly), and clinodactyly of his fifth fingers. The parents of this patient had average stature (Kamoda et al., 2001).

The fourth case included two patients, a male, and a female. The disorder of the male child was defined by shortness of upper arms, knee deformity, coxa vara deformity, shortness of metacarpals and metatarsals, depressed proximal epiphyses, and lateral thickening of the diaphysis. His parents were heterozygous thalassemia carriers, whereas he was a thalassemia patient. The disorder of a female child was defined by shortness of arms, humeral abduction, coxa vara deformity, shortness of metacarpals and metatarsals, depressed and short first toe, depressed femoral neck, and small proximal femoral epiphyses. Both cases were assumed to be sporadic (Franceschini et al., 2004).

The fifth case was our case, which included two sisters. Disorder of the older sister was identified by shortness of humeri, metaphyseal elongation and transfusing in both humeri, lateral bending and medial cortical bulging in the proximal diaphysis, coxa vara deformity, small proximal epiphyses of femora and bilateral expanded femoral neck. Clinical features of the younger sister were the same as before except for coxa vara deformity. Nevertheless, those sisters had growth and intellectual disability. The parents of the sisters were first-degree relatives, and because of that, this was estimated to be autosomal recessive inheritance (Damar et al., 2014).

1.2 Sequencing

1.2.1 Next-Generation Sequencing Era in Rare Disease Diagnosis

Before developing the Next-Generation Sequencing (NGS) technology and its adaptation, studying the genetic reason for diseases was labor-intensive, time-consuming, and expensive. Advancements in NGS and its availability across the globe have decreased the economic burden of rare disease studies. Owing to this technology, novel gene

identification and genotype-phenotype correlations are increasing hastily (Boycott et al., 2013).

Different sequencing technologies include their disadvantages. In our case, two platforms were used, Thermo Fisher Scientific Ion S5 system and Illumina NovaSeq 6000 System. The Ion S5 nucleotide detection system is developed on pH level detection. A hydrogen ion is released during each base binding and elongating, causing a pH change in the solution. According to a sensor below the well, it is detected whether DNA synthesis occurred or not. Repeating bases along the DNA might cause background noise, and homopolymer sites are often misread with this technology. This disadvantage makes variants around self-repeating nucleotide regions less reliable in Ion S5 DNAseq data (Feng et al., 2016). Illumina NovaSeq 6000 system uses a light sensor to detect nucleotide binding. During each nucleotide binding, a specific fluorescent dye attached to nucleotides emits light at a specific wavelength called sequencing by synthesis. The key disadvantage of this technology is a shorter read length, around 150 bp, compared to 200 to 600 bp with Ion S5. The shorter the reads, the higher risk of misalignment. This problem could be controlled with careful library preparation steps (Kim et al., 2021).

In the first step of rare disease diagnosis, precise phenotypic information must be gathered, and several molecular tests must be conducted. In the case of monogenic genetic disorders, whole genome sequencing (WGS) or whole exome sequencing (WES) tests have a diagnosis rate between 17-37% (Strande and Berg, 2016).

Disease-variant relations were forecasted to be completed by 2020 (Boycott et al., 2013). However, it is understood that rare diseases, such as neurodevelopmental and metabolic rare disorders, might have complex biological pathways and cannot be defined by only one variant or gene (Niemi et al., 2018), (Daoud et al., 2016). It is estimated that 20% of rare diseases still do not have an identified genetic basis (GlobalGenes, n.d.).

1.2.2 NGS Analysis

Typical NGS analysis pipeline includes sample and DNA library preparation, sequencing with NGS platforms, quality assessment and mapping to reference genome, filtering and tuning low-quality reads, variant calling and annotating, followed by phenotypical and bioinformatical filtering. Finally, clinically relevant findings are reported within a collaboration of clinicians.

Before NGS advancements, Sanger sequencing was one of the options in order to read DNA segments. Sanger sequencing was expensive and time-consuming compared to next-generation sequencing technology (Hu et al., 2021). NGS instruments provide massively parallel readings of DNA segments obtained in the sample and library preparation step. Parallel reading is a crucial point since it shortens the sequencing process and dropping costs (Metzker M. L., 2010). There are different techniques available depending on fragment length (short or long reads), detection method (optical, chemical, electrical), sample type (DNA, RNA, protein), type of sequencing (Whole Genome Sequencing

(WGS), Whole Exome Sequencing (WES), Targeted Sequencing, Methylation Sequencing, Chromatin immunoprecipitation sequencing (ChIPseq), Bisulfite sequencing) (Qin D., 2019; Levy and Boone, 2019). Medical centers decide which technology to use according to their needs and budget.

Raw sequencing data, including all the reads and their reading quality metrics, are evaluated following sequencing. Low-quality reads or their parts are trimmed out, leaving high-quality bases to be used in the reference mapping step. Raw sequence data includes all fragments regardless of their position in chromosomal structure. In order to determine the arrangement of nucleotides or amino acids, a reference genome, if it exists, is used. All reads are arranged so that they would collapse on each other at specific positions on the reference genome sequence. If a reference genome does not exist, de novo assembly is preferred, where reads are collapsed on each other at specific positions without a guarantee of correct alignment. GRC (Genome Reference Consortium) is the main human reference genome (HRG) source. Currently, two versions of HRG are used, GRCh37, initially released in February 2009, and GRCh37.p13, a widely integrated version released in June 2013 (Cunningham et al., 2015). A recent release is GRCh38.p13, assembled in May 2022, which aims to curate and join all of the information provided by NCBI and EMBL-EBI (Morales et al., 2022).

Filtering and mapping quality evaluations are taken into consideration if needed. Commonly used metrics are transition/transversion (Ti/Tv) and heterozygous/nonreference-homozygous (het/nonref-hom) ratios. A transversion mutation is the conversion of purine into pyrimidine, and the transition is a change between purines or pyrimidines by methylation, a natural biological process occurring in the human genome. Natural mechanisms bias transition mutations at high prevalence than transversions, and the estimated natural Ti/Tv ratio is around 2.0 for the whole genome and 3.0 for the whole exome. Values less than these suggest high false positive variant calls (Wang et al., 2015). On the other hand, the het/nonref-hom ratio of around 2.0 suggests a natural biological sequencing data illustration (Guo et al., 2014).

VCF (Variant Calling Format) is a standard raw file that includes basic information about each position read while sequencing and mapping. This information generally includes dbSNP identification number, base change at the position, read depth and quality, and metadata noting tools and their versions used during the analysis pipeline so far (Danecek et al., 2011). Variant calling is based on the alterations detected compared to a reference genome. These alterations may include but are not limited to, single nucleotide variations (SNVs), insertions and deletions (InDels), structural variations (SVs), and copy number variations (CNVs) (Mahmoud et al., 2019).

The final VCF file may contain thousands of variants, and assessing irrelevant and false-positive variants is critical. Further annotation of variants is mainly considered, and information such as population frequency, genomic position of variant (exonic, intronic, splice site, etc.), gene name, the biological and molecular function of gene or region, and

phenotype. All available information about that genomic position across the literature could be included in the annotation step (McLaren et al., 2016; Austin-Tse et al., 2022).

1.3 Bioinformatic Analysis

Bioinformatics is an interdisciplinary science integrating computer sciences with biology field and statistics aiming to extract desired information from raw data and analyze and develop new perspectives from biological and medical data. Besides biology, major sciences such as chemistry, physics, and engineering are widely combined for managing multi-omics (Karlin S., 2015). Clinical bioinformatics is a sub-disciplinary focusing on human diseases to find and help in treatment, diagnostics, and life quality improvement via clinical applications (Wang and Liotta, 2011).

Bioinformatic analysis is involved in somatic cancer studies, germline genetic diagnostics, drug interaction and development studies, and phylogenetic mapping (Li et al., 2020; Yohe and Thyagarajan, 2017; Chang P. L., 2005).

In oncology, associating somatic variants with drug response and tumor mutation burden is a recently developing area in bioinformatics (Holtsträter et al., 2020). The genetic basis of cancer is revealed further, and this information is integrated into oncological diagnostics and treatment design every day. Comparative studies between tumor and normal tissue enable researchers to characterize genomic and proteomic variations affecting cancer progression (Liu et al., 2015). Drug interaction and development studies are crucial in personal medication and cancer therapies. Pharmacogenetic studies merging with bioinformatic analysis are involved in drug-drug interaction (DDI) and drug-protein interaction (DPI) experiments (Wu et al., 2014; Tabei et al., 2019). Commonalities and differences between species may reveal important information about evolution, and combining results from different species into human genetic knowledge is one topic of phylogenetic mapping (Shakya et al., 2020). Variations among orthologous genes between humans and closely related species are sometimes key reasons for human-specific diseases (Wu et al., 2006). For example, recent studies found that these alterations improved our knowledge about cancer genetics and further therapy options (Somarelli et al., 2020).

There are a variety of databases and tools available in variant interpretation and prioritization processes. Both paid and publicly provided tools are accessed online or as desktop applications. Most of these databases are reached with a web interface or API services. Table 1 consists of some of the databases and Table 2 consists of the tools used for variant interpretations. Not all of them were included in this study.

Table 1: Databases for Variant Prioritization. This table includes online databases used for variant interpretation and prioritization.

Database	Detail	Reference
ACMG (American College of Medical Genetics and Genomics)	Database for guidelines and suggestions about variant prioritization	(Richards et al., 2015)

PhenomeCentral	A platform for physicians and scientists to connect and share data about rare disease cases	(Buske et al., 2015)
PhenoTips	Collecting and analyzing phenotype information of genetic diseases	(Girdea et al., 2013)
DECIPHER	Web-based database with tools for variant interpretations	(Firth et al., 2009)
DDD (Deciphering Developmental Disorders)	A platform for improving information about children with developmental disorders	(Bragin et al., 2014)
GPAP (Genome-Phenome Analysis Platform)	Platform with tools for phenotype and genome association and connecting clinicians and researchers	(Laurie et al., 2022)

Table 2: Tools for Variant Prioritization. This table includes online and local tools used for variant interpretation and prioritization.

Tools	**Detail**	**Reference**
Exomiser (hiPHIVE)	Annotating variants, prioritize variants according to phenotype from HPO and pathogenicity	(Robinson et al., 2014)
ANNOVAR	Variant annotation tool for functional information retrieved from publicly available databases and variant prioritization for Mendelian diseases	(Wang et al., 2010)
Variant Effect Predictor (VEP)	Variant annotation tool for determining variants effects by combining several publicly available databases and in silico prediction tools	(McLaren et al., 2016)
SnpEff	Variant annotation and tool for prediction of functional effects of variants on genes and proteins	(Cingolani et al., 2012)
interVAR	Tool for clinical interpretation of genetic variants and their pathogenicity according to ACMG/AMP 2015 guidelines	(Li and Wang, 2017)
eXtasy	Tool for prioritization of variants based on given phenotypic information	(Sifrim et al., 2013)
Phenomizer	Analyzing phenotypic information in HPO terms and resulting in matching candidate diseases	(Köhler et al., 2009)
FACE2GENE	Phenotype search tool for given face photo and resulting candidate disease information	(Javitt et al., 2022)
PhenIX	Variant evaluation and ranking tool based on pathogenicity and semantic similarity of patients' phenotype based on HPO terms	(Zemojtel et al., 2014)

1.3.1 Bioinformatic Analysis in Rare Diseases

Bioinformatic methods combining genome sequencing with phenotype information can reveal genetic reasons and related pathway consequences of rare diseases. Standard pipelines follow WES or WGS data generation, mapping, variant calling, and annotation. Comparison of variants against population variant frequency databases such as 1000Genomes (Fairley et al., 2020), gnomAD exome and genome (Karczewski et al.,

2020), filtering out the common variants (%1) and clinically irrelevant variations to narrow candidate variants into small subset. Rare diseases are caused by rare mutations (Nellåker et al., 2019), and it is challenging as few cases are available for bioinformatical analysis, limiting the options for analysis.

There are two main methods used to prioritize disease-causing genes or variants. The first case is chosen when a group of patients has the same clinic. In this case, filtering common variants out of individual variants in a group is an effective way to list candidate genes. The second case is when there is a single patient. In this case mode of inheritance is crucial information to be specified. In autosomal recessive inheritance, if the patient's parents are not consanguineous, variants homozygous in the patient and heterozygous in the parents are chosen, also known as loss of heterozygosity (LOH) analysis. On the other hand, if families are consanguineous, homozygous variants of patients that are heterozygous in parents and compound heterozygous in the patient are chosen for further investigation. In X-linked recessive modes, variants in affected male patients and heterozygous in carrier females are chosen.

Autosomal dominant inheritance can present two options; affected family members or healthy family members and affected patients. In the affected family members' case, heterozygous variants in affected members are filtered from the heterozygous variants in healthy members. If there are not any affected members besides the patient, de novo mutations are considered, which are unique to the patient and not found in healthy members.

Finally, advancements in computational technologies made mosaic mutation detection easier and faster. In this case, affected tissues are compared with unaffected tissues, and unique variants in affected tissues are selected and combined with network analysis (Niemi et al., 2018; Sun et al., 2015; Rahit and Tarailo-Graovac, 2020; Lee et al., 2014). Once the candidate list is constructed, variants with phenotypic information are considered, and phenotypically irrelevant variants are filtered out (Pengelly et al., 2017). Case of de novo mutations, which do not have literature information, are investigated through protein structure prediction algorithms for assessing their functional effects (Neveling et al., 2012).

1.3.2 Bioinformatic Analysis and Variant Effect Predicting Tools

Different variant effect prediction tools are available for the bioinformatic analysis of human DNA sequences Here are the tools briefly explained that were used in this study.

ION Torrent Suite assembles and maps raw sequence data generated by the ION Torrent S5 sequencing platform into BAM and VCF formats. CLC Genomics Workbench is a paid software used in NGS data analysis, such as assembling raw sequencing data, quality assessment and trimming, mapping against the reference genome, and variant calling.

The Ingenuity Variant Analysis (IVA) tool is a paid variant annotation tool provided by QIAGEN. It is easily linked with ION Torrent Suite software, and VCF files are annotated with information found online or in available databases. It also performs Trio sample analysis by merging and annotating thoroughly. This software was discontinued by QIAGEN and replaced with the QCI Interpret software platform. QIAGEN QCI Interpret platform is a variant annotation and interpretation software similar to IVA, yet this platform integrates more functional data and effect prediction tools and increases the available database range. Ensembl VEP (Variant Effect Predictor) is a public online variant annotation tool provided by EMBL-EBI. This tool collects information from variant effect prediction tools and clinical and variant frequency databases.

Ensembl Assembly Converter is a tool provided by EMBL-EBI, which is based on the CrossMap tool to convert genomic coordinates between different genome assemblies. (Zhao et al., 2013). Bcftools merge is a VCF manipulation tool that is publicly available and used for joining multiple VCF files and arranging variants according to their coordinates.

DANN is a functional prediction tool based on a deep neural network based on evolutional conservation and calculates the variant's pathogenicity for both coding and non-coding variants. DEOGEN2 predicts variant pathogenicity by combining molecular effects of variants in amino acid or domain scale and gene relations. FATHMM-MKL is a prediction tool providing a pathogenicity score integrating the conservation of sequence and domains and the effect of variation on protein's function. PredictSNP prediction tool can perform the functional effect of variants on both nuclear and amino-acid levels for disease-related mutations integrating scores from different available prediction tools.

M-CAP is a pathogenicity score prediction tool focused on rare missense variations tailored for clinical approaches combining its algorithm and scores from SIFT, Polyphen-2, and CADD tools. MVP is a prediction tool developed for missense variants integrating different types of data such as gene's type of function in pathways, gene's toleration against loss-of-function variations, and mode of action using a deep residual network. PolyPhen is an effect prediction tool for amino acid variations using sequence homology features. PolyPhen-2 is considered to be an updated and improved version of PolyPhen-1. SIFT is a functional prediction tool based on conservation scores of positions in similar protein sequences. DANN is a functional prediction tool based on a deep neural network based on evolutional conservation and calculates the variant's pathogenicity for both coding and non-coding variants. MAPP tool predicts the pathogenicity of a variation by measuring physicochemical properties incorporating impacts of all possible amino-acid variations at the position of homologous proteins. MutationAssessor is a functional pathogenicity prediction tool for evolutionary conversation of the position of homologous proteins.

MutationTaster is a prediction tool combining different levels of information about variations such as type of mutation, conservation of amino acids, functional loss of protein domains at both DNA and amino-acid levels, and information from data sources like

UniProt, ClinVar, ExAC, and Ensembl. This tool can predict the effect of missense, indel, intronic, and synonymous variations. REVEL is a prediction tool where pathogenicity scores of missense variations are generated by different prediction tools and weighed into an overall score. SNAP is a prediction tool for evaluating the protein function effect of single amino acid variations. Information is gathered from protein databases and combines pathogenicity predictions of other tools. PhDSNP tool is a prediction algorithm developed for single-point variations using support vector machines on sequence-based information for coding and non-coding variants. PrimateAI uses an algorithm based on common missense variants from humans and other primate species and their previously known effects by deep residual neural network approach. SpliceAI is a prediction tool specially developed for the splicing effect of variations along exon-intron junctions.

SDM prediction tool measures the stability and functional effects of a given amino acid variation on a protein by comparing homologous proteins of known three-dimensional structures. SwissPDB Viewer is a tool for comparing 3D structures of proteins and analyzing mutation effects on protein stability. Structure prediction tools for proteins can be used if the structural information for proteins is unavailable. The QUARK uses an ab initio method for structural predictions, while I-TASSER uses a threading approach.

1.4 Motivation

Many unsolved rare diseases lack information about causative molecular and genetic processes. Reanalysis of exome sequencing data for unsolved rare disease cases has the potential to reveal novel gene and disease associations. Advancements in bioinformatics algorithms, new literature, and published cases in databases enable rare disease research to reveal new variants. Here, we present the importance of reanalyzing older sequencing data with recent algorithms and literature and resequencing old DNA samples for challenging rare diseases. The prospect of this study is that resequencing will cover more positions and reevaluate older candidate variants. Reanalysis will increase the knowledge of variant interpretation and prioritization processes.

CHAPTER 2

MATERIALS AND METHOD

2.1 Pedigree

A patient or an individual being studied is called a proband. In our case, two siblings are our probands. They are both affected by humeri shortness, metaphyseal elongation of humeri, bilateral expanded femoral neck, shortness of limbs, growth retardation, and intellectual disability. Parents of probands are not affected and are assumed not to share any phenotypical findings with their children. The consanguinity of parents is known, and they are first-degree relatives. Siblings were diagnosed with congenital Rhizomelic Dysplasia Patterson-Lowry type. There are no known affected relatives as well.

2.2 Next Generation Sequencing Sample Preparation

DNA was extracted from the peripheral leukocytes of each patient by IprepTM PureLink® gDNA Blood Kit (Invitrogen, Carlsbad, CA) according to the manufacturer's protocol.

Patients underwent exome sequencing. Ion AmpliSeq™ Library Kit Plus (Life Technologies, Guilford, CT, South San Francisco, CA), a ready-to-go analysis kit that includes primer pairs for interested genes sequences to be analyzed with the Ion GeneStudio S5 platform (Life Technologies, Guilford, CT, South San Francisco, CA). Analyses were done using an ION Torrent 540 chip (Life Technologies, Guilford, CT, South San Francisco, CA).

2.3 Next Generation Sequencing Analysis

2.3.1 Initial NGS analysis of ION Torrent S5 data in 2017

Exome data of parents and two probands were acquired via NGS instrument Ion S5 System by Thermo Fisher Scientific in Binary Alignment Map (BAM) and Variant Call

Format (VCF) formats. VCF files were fed into Ingenuity Variant Analysis Tool by Qiagen for annotations and LOH on trio sample analyses.

VCF and BAM files were generated with Torrent Suite, and the following tools were embedded into the suite; Torrent Variant Caller version "tvc 5.8-17 (93ef10d)", parameters of "Generic – S5/S5XL (540) – Germ Line – Low Stringency", "TS version: 5.8", basecaller version was "5.2-25/46145e1", Torrent Mapping Alignment Program (TMAP) version was "5.2.25 (46145e1) (201609011819)", human reference genome assembly "hg19".

2.3.2 Recent NGS analysis of ION Torrent S5 data in 2022

Exome data (VCF files) of each parent and two probands acquired with ION Torrent S5 data in 2017) were merged with bcftools merge tool (v1.15.1). The merged VCF file was reannotated with the online Ensembl VEP tool in 2022 using default parameters provided by Ensembl.

2.3.3 Recent NGS analysis of NovaSeq 6000 data in 2022

Exome data of the same DNA samples of two probands and two parents (initially sequenced with ION S5) were acquired via Illumina NovaSeq 6000 with QIAseq Human Exome Kit and bcl to fastq conversion, quality assessments, reference genome mapping, and low-quality reads and variants filtering. VCF file generation was achieved with Qiagen CLC Genomics Workbench 12.0.3.

2.4 Converting GRCh37/hg19 into GRCh38/hg38

VCF files, initially built on genome assembly version hg19, were used as input in the Assembly Converter tool to convert hg19 build genomic coordinates into recent build hg38 (Zhao et al., 2013). Once assembly conversion was achieved, resulting files were annotated with the Ensembl VEP tool individually, and LOH on trio sample analysis was held manually in Excel software.

2.5 Variant Interpretation

In addition to variants below thresholds of Q20, 5X coverage, p-value, and MAF 0.05, intronic and synonymous variants were filtered out, remaining with only splicing (-2/+2 - +10 bp) and exonic regions. Following the prior filtering steps, pedigree analysis of the phenotype Patterson-Lowry Rhizomelic Dysplasia according to family history revealed a recessive or de novo heritage.

Two probands and both parents were considered during the LOH on trio sample analysis. TRIO analyses were achieved separately for both probands. Results were gathered in Excel format. QIAGEN Ingenuity Variant Analysis Tool was used in the 2017 analysis of

ION Torrent S5 data, and QIAGEN QCI Interpret Tool was used in the 2022 analysis of NovaSeq 6000 data. Loss of Heterozygosity analysis was held, and heterozygote variants of parents and homozygous variants of probands and de novo variants were kept. Common homozygous and de novo variants between siblings were collected.

2.6 Bioinformatic Analysis

After the abovementioned actions, the final variant list was narrowed with homozygous variants, and there were no de novo variants. Variants in the final list were taken into further bioinformatic analysis in the means of clinics and mutation impacts. In silico variant effect algorithms of DANN, FATHMM-MKL, M-CAP, MutationAssessor, MutationTaster, SIFT, PredictSNP, MAPP, PhDSNP, PolyPhen-1, PolyPhen-2, SNAP, DEOGEN2, M.V.P., PrimateAI, and REVEL were tested on each variant.

Chromosomal loci of final candidate variants were assessed. One of the in-silico algorithms, PredictSNP, was focused on the change and effect at the amino acid level which was a closer approach to translational effect predictions. Protein sequences of genes were retrieved from the UniProt database. Sequences were inserted on PredictSNP "Consensus classifiers for prediction of disease-related amino acid mutations" algorithm, variations were tested separately, and results were recorded. The results were investigated with Integrated Genomic Viewer (Robinson, 2011). ACMG guidelines were used for variant interpretation (Richards, 2015).

2.7 Phenotypic and Literature Prioritization

The next step was variant prioritization and assessment according to the phenotypic relations found in the available as such Pubmed, Google Scholar, OMIM, ClinVar, Mastermind, and gnomAD. Pubmed and Google Scholar were used for searching literature information found in articles, journals, or books. OMIM database is used for genotype-phenotype relations investigations. ClinVar is a database of information on the variant level and variant-phenotype relations provided by researches. GnomAD database holds variant frequencies found in healthy individuals. Mastermind is a platform providing variant-article matches by digging internet resources manually and automatically.

The *OBSCURIN* gene has very little information linked with phenotypes. In the OMIM database, there is not any disease correlation published yet. In PDB, 3D protein structures were not covering all amino acids yet. Therefore, ab initio modeling of the *OBSCURIN* gene and the functional effects of these two mutations were unpredictable. Since OBSCURIN is a musculoskeletal protein, its relation with bone deformities is considered. Both variants are located in Ig-like domains in different exons in UniProt. Both variants were not detected in the homozygous genotype in publicly available population databases.

COQ8A gene has more information compared to *OBSCN.* In OMIM, it is related to COENZYME Q10 deficiency and has diverse clinical features, and the severity of its

progress is variable. The 3D protein model in PDB covers all its amino acids with experimental validations. It was located at the end of a protein kinase domain in UniProt. There was not any homozygous individual detected in publicly available population databases.

CHAPTER 3

RESULTS

This study aims to find the genetic basis of a rare disease observed in a family in Turkey and compares methods that can be used during analysis steps. To investigate this rare unsolved condition, we followed three approaches. The first approach is reanalyzing ION Torrent Sequence data with the hg19 genome built in recent literature. The second approach is reanalyzing ION Torrent sequences with the recent (updated) genome build, and the third is resequencing samples with recent technologies. Sequencing the same old DNA samples with NovaSeq 6000 and analyzing the data with both hg19 and hg38 assemblies was achieved in 2022. All candidate variants listed in the Ion S5 study were covered within NovaSeq 6000 study.

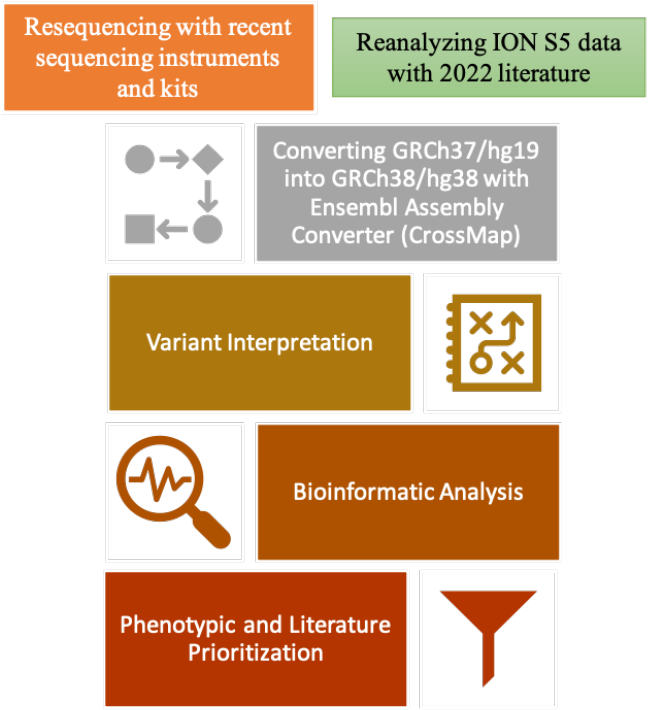

Figure 1: Workflow of processes taken during thesis study.

Tools used in different steps are included below in Table 5. All variant effect prediction tools were not included in this table. Some tools were meant to operate for their belonging sequencing platforms. Paid tools were out of access in the analysis of the hg38 build for both ION Torrent S5 and NovaSeq 6000 data, publicly available tools used in this part of the study are mentioned in Table 4.

Table 3: Variant interpretation tools and databases used in different steps for ION Torrent S5 and Illumina NovaSeq 6000 instruments.

	ION TORRENT S5	**ILLUMINA NOVASEQ 6000**
Seq. Instrument → VCF	Torrent Suite (v5.8-17)	Qiagen CLC Genomics Workbench
LOH analysis of TRIO	QIAGEN Ingenuity Variant Analysis Tool (Paid)	QIAGEN QCI Interpret (Paid)
Phenotypic & Literature Prioritization	Pubmed, Google Scholar, Mastermind	Pubmed, Google Scholar, Mastermind

Table 4: Bioinformatic tools used in 2022 hg38 build analysis.

Variant Interpretation	Ensembl VEP (free)
LOH analysis of TRIO	Merge (bcftools), Filtration (Manually)
Phenotypic & Literature Prioritization	Pubmed, Google Scholar, Mastermind (free version), Franklin (free version)

3.1 Analysis of ION Torrent sequencing data with the hg19 genome build with 2017 literature

Data of two siblings (probands) and two parents were sequenced by an ION S5 instrument assembled, and variants were called with ION Torrent Suite. After each analysis step, the number of prospective variants decreased (see Table 5. Initially, there were over 25000 variants. After quality trimming and intronic variants exclusion-more than ten bases, splice sites, and merging all samples, almost 1000 variants were obtained. The next step was to filter out irrelevant variants with the mode of expected inheritance and clinical futures, higher allele frequency, reported benignly in literature which revealed 13 variants. Out of them, pathogenically affordable five candidate variants are listed.

Table 5: Number of variants found after each step during the analysis of ION Torrent data of two probands and two parents

Steps	**# of Variants**
Initial	>25000
Variant Interpretation	1000
Merging 4 samples; Phenotypic and Literature Prioritization	13
Clinical Assessment	5
Final Literature and Bioinformatic Evaluation (2022)	3

The first analysis in 2017 resulted in 5 candidate variants shown in Table 6 below, and the same genetic basis of the sister's condition was unresolved at that time. Due to a lack of information and publicly available tools, analysis was held up.

Table 6: Initial candidate variants. This table includes canonical transcripts, HGNC gene symbols, HGVS nucleotide coding, and HGVS amino acid, rsID, and exon location information.

Transcript	**Gene**	**HGVS coding**	**HGVS amino acid**	**Exon**	**rsID**
NM_001271223.2/ ENST00000570156.2	*OBSCN*	c.4489C>T	p.R1589W	16 of 116	rs367856512
NM_001271223.2/ ENST00000570156.2	*OBSCN*	c.9754_9755delAGinsTT	p.R2823L	36 of 116	rs386640014
NM_020247.4/ ENST00000366777.3	*COQ8A (ADCK3)*	c.1534C>T	p.R512W	13 of 15	rs149682899
NM_001075.4/ ENST00000265403.7	*UGT2B10*	c.1420C>T	p.H474Y	6 of 6	rs200109225
NM_001242729.1/ ENST00000420470.2	*ARHGEF38*	c.904G>A	p.E302K	7 of 14	-

In 2017, there was not any research mentioning *COQ8A*:c.1534C>T variant, and the *COQ8A* gene was annotated formally as *ADCK3*. Gene was only related to COENZYME Q10 deficiency, encephalopathy followed by muscle weakness (Lagier-Tourenne et al., 2008; Liu et al., 2013; Jacobsen et al., 2017).

At that time, even less information was available about the gene *OBSCN*. Gene was related to cardiomyopathies (Cirino et al., 2008; Marston, 2017) and cancers (Manring et al., 2017; Rajendran et al., 2017; Perry et al., 2013).

3.2 Analysis of ION Torrent Sequencing data with the hg19 genome build with 2022 literature

While analysis of recent literature revealed new information, the final candidate list was not changed. However, this time with the help of multiple bioinformatic variants predicted effect tools and recent literature, variants in *UGT2B10* and *ARHGEF38* were filtered out. These two variants out of 5 were predicted to be tolerated or benign according to tools or unrelated to the case mentioned in the method section and excluded. Two variants out of the remaining three possible pathogenic variants were at the same gene (i.e., *OBSCN*:c.4489C>T and *OBSCN*:c.8467_8468delAGinsTT), a third variant (*COQ8A*:c.1534C>T) were at the same locus (1q42.13). In FASTA format, amino acid sequences of OBSCURIN and COQ8A were taken from UniProt, Q5VST9 *(OBSCN_HUMAN),* and Q8NI60 *(COQ8A_HUMAN*). Sequences were inserted into the PredictSNP algorithm, and these three variations were considered pathogenic according to PredictSNP results. In Table 7, although ACMG classifications were not changed, changes in the pathogenicity prediction are observed.

Table 7: Pathogenicity predictions compared before and after PredictSNP tool, 2022 analysis

	ACMG Classification	Prior Prediction	PredictSNP Prediction
COQ8A:p.R512W	VUS	Damaging	Damaging
UGT2B10:p.H474Y	VUS	Damaging	Mildly damaging-neutral
ARHGEF38:p.E302K	VUS	Damaging/Tolerated	Benign
OBSCN:p.R1589W	VUS	Damaging	Damaging
OBSCN:p.R2823L	VUS	Damaging	Damaging

According to the research held by Nair et al. (2019), encephalopathy with congenital hip luxation, cardiac involvement, short stature, and facial dysmorphic phenotypes was observed in a patient with homozygous *COQ8A*:c.1534C>T variant. Hip luxation and short stature were similar clinical features in our case. This research was the first report mentioning bone malformations associated with the *COQ8A* gene. However, the variant in other reported genes *MED25* was not expected in our case. This was the first literature information associating *COQ8A* with any bone deformation. Analysis of the 2022 literature revealed a relationship between candidate variants and one of our clinical findings.

Recent research found new clinical synopsis relations between the *OBSCN* gene. Cabrera-Serrano et al. (2021) identified ten biallelic variants in rhabdomyolysis patients. Skeletal muscle atrophy was one of the results related to *OBSCN* gene variations. Qiu et al. (2020) revealed that *OBSCN* was highly expressed after the denervation of the rat model. Even though *OBSCN* was previously related to cardiomyopathies, Cabrera-Serrano et al. (2021) patients were not diagnosed with cardiac findings. These results imply the complexity and lack of knowledge of the *OBSCN* gene, which could also be associated with other phenotypes in the future. The number of articles published between the early 2000s and 2017 about the *OBSCN* function and its causes was almost the same between 2017 and mid-2022 (retrieved August 2022, from https://pubmed.ncbi.nlm.nih.gov/?term=obscn).

3.3 Analysis of ION Torrent Sequencing data with the hg38 genome build with 2022 literature

Human reference genome builds also updated with developing literature (Schneider et al., 2017). Samples were already sequenced prior to this research to diagnose patients in 2016. The diagnosis was achieved; however, a precise genetic basis could not have been identified, and data were taken into research. During the initial analysis, the available human reference genome build version was hg19 at the laboratory. As Schneider et al. (2017) mentioned, hg19 was not accurate enough, and the newest human genome build, hg38, needed to be studied. Morales et al. (2022) constructed the latest assembly GRCh38.p13 and merged NCBI and EMBL-ENI transcripts to make a universal infrastructure.

Table 8: Candidate variants after converting genome build hg19 to hg38. Differences and similarities between candidate variants identified after genome converter

	GRCh37/hg19	GRCh38/hg38

Gene HGVS nucleotide Amino acid rsID	**Transcript**	**Chromosomal Position**	**Transcript**	**Chromosomal Position**
OBSCN c.4489C>T p.R1589W rs367856512	NM_001271223.2/ ENST00000570156.2	1:228444531 C>T	NM_001386125.1/ ENST00000680850.1	chr1:228256830 C>T
OBSCN c.9754_9755 delAGinsTT p.R2823L rs386640014	NM_001271223.2/ ENST00000570156.2	1:228469903 AG>TT	NM_001386125.1/ ENST00000680850.1	chr1:228282202 AG>TT
COQ8A (*ADCK3*) c.1534C>T p.R512W rs149682899	NM_020247.4/ ENST00000366777.3	1:227172604 C>T	NM_020247.5/ ENST00000366777.4	chr1:226984903 C>T
UGT2B10 c.1420C>T p.H474Y rs200109225	NM_001075.4/ ENST00000265403.7	4:69696430 C>T	NM_001075.6/ ENST00000265403.12	chr4:68830712 C>T
ARHGEF38 c.904G>A p.E302K -	NM_001242729.1/ ENST00000420470.2	4:106569735 G>A	NM_001242729.2/ ENST00000420470.3	chr4:105648578 G>A

Before converting the reference genome assembly of data, we had to normalize multiallelic variants and remove duplicated calls using the bcftools norm tool to have matching annotations of the same variants between family members. With the help of the Ensembl Assembly Converter tool, our VCF files with the hg19 build were converted into hg38, meaning all chromosomal positions of variants were transformed into their corresponding hg38 positions. Later converted VCF files were fed into the online Ensembl VEP service, and manually LOH on the trio sample was held with Excel software.

In our case, updating the human reference genome assembly of our data did not make any improvements in diagnosing the exact causative variant. The 13 variant list was narrowed to 10 due to recent pathogenicity prediction updates. After all filtering steps were achieved, the final candidate gene list was not changed. However, transcript ids and chromosomal positions were changed, and the rest of the information was the same, as seen in Table 8. Nucleotide changes and their positions on exons, distances from splice sites, rsIDs, and cumulative pathogenicity predictions were the same.

So far most useful step was to reanalyze all data with recent publicly available sources and literature information. This step helped us narrow the final five candidate variants list into three and increase one variant's pathogenicity potential due to its relation with similar phenotypic features in another study (Nair et al. 2019).

Table 9: Candidate variants after resequencing. This table includes canonical transcripts, HGNC gene symbols, HGVS nucleotide coding and HGVS amino acid, rsID and exon location information.

Transcript	Gene	HGVS coding	HGVS amino acid	Exon	rsID
NM_001271223.2/ ENST00000570156.2	*OBSCN*	c.4489C>T	p.R1589W	16 of 116	rs367856512
NM_001271223.2/ ENST00000570156.2	*OBSCN*	c.9754_9755delAGinsTT	p.R2823L	36 of 116	rs386640014
NM_020247.4/ ENST00000366777.3	*COQ8A* (*ADCK3*)	c.1534C>T	p.R512W	13 of 15	rs149682899
NM_001075.4/ ENST00000265403.7	*UGT2B10*	c.1420C>T	p.H474Y	6 of 6	rs200109225
NM_001242729.1/ ENST00000420470.2	*ARHGEF38*	c.904G>A	p.E302K	7 of 14	-

3.4 Analysis of Illumina Sequencing of two probands and two parents with the hg19 genome build with 2022 literature

Sequencing the same samples with different sequencing platforms and exome kits helped us to compare variants at complex regions and regions with low coverages. All candidate variants listed in the Ion S5 study were covered within NovaSeq 6000 study. All reads at candidate variant positions seem to have high qualities (>Q30); the reliability of candidate variants is therefore increased. The final candidate list was not changed, as seen in Table 9. As a result of all data gathered by reanalyzing and resequencing efforts, the final three prospective variants are decided, see Table 10.

Table 10: Final candidate variant list in conclusion. This table includes canonical transcripts, HGNC gene symbols, HGVS nucleotide coding, and HGVS amino acid, rsID, and exon location information.

Transcript	Gene	HGVS coding	HGVS amino acid	Exon	rsID
NM_001271223.2/ ENST00000570156.2	*OBSCN*	c.4489C>T	p.R1589W	16 of 116	rs367856512
NM_001271223.2/ ENST00000570156.2	*OBSCN*	c.9754_9755delAGinsTT	p.R2823L	36 of 116	rs386640014
NM_020247.4/ ENST00000366777.3	*COQ8A* (ADCK3)	c.1534C>T	p.R512W	13 of 15	rs149682899

These three variants are not found in population databases in the homozygous state, do fit with the estimated inheritance pattern, and are predicted pathogenic bioinformatically. Finally, converting the hg19 reference genome assembly of resequenced data into the hg38 version did not reveal any additional information.

3.5 Analysis of Illumina Sequencing of two probands and two parents with the hg38 genome build with 2022 literature

In order to compare different genome builds with Illumina Sequencing, genome assemblies of data of four samples sequenced with Illumina NovaSeq 6000 are converted into hg38 build and analyzed as mentioned in the methods. In this case, genome build conversion and further investigations did not reveal additional information, and changes were the same as seen in Table 8.

3.6 Protein stability effect prediction of two variations on relative domains of OBSCURIN protein

Energy alterations in mutated proteins might indicate an effect of variation on their function. Besides the PredictSNP algorithm's pathogenic prediction on both variations in OBSCURIN protein, two protein stability effect prediction tools, SDM and SwissPDB Viewer, were used. These tools calculate the energy difference between two 3D protein structure models. The difference in the stability of proteins can be used to detect the impact of a variation. The Gibbs Free Energy difference between mutated protein (ΔGw) and wild type (ΔGm), ΔΔG = ΔGm - ΔGw, is measured to guess the mutation effects on protein stability. There was a 3D protein model in PDB at the position of variation p.R2823L, yet no models at the position of p.R1589W. In this case, the PDB structure with PDBid 2ENY was selected for domain Ig-like 27, where R2823L is located. For the 3D structure prediction of domain Ig-like 16 p.R1589W I-TASSER tool. The amino acid sequence of the domain containing the first mutation is gathered from UniProt with Q5VST9 in FASTA format. The sequence is given to the I-TASSER prediction server. I-TASSER builds a model on prior knowledge using the threading method in its structure prediction algorithm. 3D models of both variations were fed into SDM, and SwissPDB Viewer tools and mutations were introduced. According to the results shown in Table 11, SDM prediction of both variants shows a pathogenic effect in the domain's function. While SwissPDB Viewer prediction for R1589W is insignificant, the R2823L prediction reports a pathogenic effect.

Table 11: ΔΔG(Gibbs Free Energy) (kcal/mol) calculations of both variations introduced on their domain structure models by SDM and SwissPDB Viewer tools on OBSCURIN protein.

Variation	SDM	SwissPDB Viewer
R1589W	-0.48	-401.14
R2823L	-0.49	-0.59

3.7 Quality metrics of sequencing data from Illumina NovaSeq 6000 And ION Torrent S5

After sequencing the same samples with different sequencing platforms, data quality was assessed. All data from both ION Torrent S5 and Illumina NovaSeq 6000 had enough

quality to be analyzed in further steps. The average read depth was over 100 in all samples, and read numbers with high qualities were abundant.

Table 12: Quality control metrics for raw sequencing data and aligned data gathered from ION Torrent S5 device

Sample	**Mapped Reads**	**On Target**	**Mean Depth**	**Bases**	**≥ Q20**	**Reads**	**Mean Read Length**
Proband1	38,916,356	91.43%	116.5	7,477,089,447	6,504,409,612	39,201,848	190 bp
Proband2	46,632,429	91.96%	141.2	9,029,111,300	7,884,307,328	46,930,814	192 bp
Father	34,310,029	92.15%	103.8	6,636,415,190	5,806,137,038	34,522,017	192 bp
Mother	49,604,397	91.46%	148.7	9,553,417,116	8,328,127,028	49,941,740	191 bp

Table 13: Quality control metrics for raw sequencing data and aligned data gathered from Illumina NovaSeq 6000

Sample	**Mapped Reads**	**On Target**	**Mean Depth**	**Coverage ≥ 20X**	**Bases**	**≥ Q25**	**Mean Read Length**
Proband1	103,340,994	78.44%	128	89.15%	14.257.396.640	99,46%	137bp
Proband2	120,476,465	77.93%	153	95,51%	16.596.265.596	99,37%	137bp
Father	101,423,522	78.97%	131	95,56%	13.945.597.762	99,43%	137bp
Mother	90,146,207	77.17%	113	95,41%	12.399.408.413	99,40%	137bp

CHAPTER 4

DISCUSSION AND CONCLUSION

This thesis aims to find the genetic basis of a rare disease, Patterson-Lowry rhizomelic dysplasia, in a family in Turkey with two affected sisters and compares methods that can be used in the analysis. Patterson-Lowry rhizomelic dysplasia is a disorder mainly diagnosed by humeral abduction, coxa vara deformity, shortness of metacarpals and metatarsals, and in some cases, mental and motor retardation. The genetic basis is still unknown (OMIM, 2022).

First, we have reanalyzed the ION Torrent Sequence variant calls built with the hg19, in recent literature. In order to apply this approach, the literature search for the candidate variants is repeated four years apart, in 2018 and 2022. Five candidate variants were narrowed down to 3 after a recent literature search. Three candidate variants were unrelated to the case due to high population frequencies, unrelated phenotypic features, and benign pathogenicity predictions reported recently.

One of the candidates in *COQ8A* was mentioned by Nair et al. (2019) after the first analysis, and clinical findings partially matched our case. According to the clinical report, probands with hip luxation and short stature were similar, but cardiac involvement was not observed in our case.

The initial analysis focused mainly on *OBSCN* variants since two separate homozygous variants and parents were carriers. Two homozygous variants in the same gene are expected to affect function loss. However, lacking information about the *OBSCN* gene and mostly related phenotypes of cardiomyopathies, cancer, and muscle atrophies were not enough for direct association with our case. Emerging literature about *OBSCN* might reveal new information related to bone deformation might increase the chance of it.

Since there were publications and more information about COQ8A, the pathogenicity, and its phenotype-genotype relation were known more than the OBSCURIN protein. Further stability analysis of variations on OBSCURIN protein was conducted. The aim was to predict the impact of two variations on the three-dimensional structure and function of the protein. OBSCURIN protein consists of more than 8000 amino acids, and most of its

domain structures are not available yet. In our case, the structural model for the domain Ig-like 27 of R2823L was available in the PDB database with accession id 2ENY. There was no structural model for the domain Ig-like 16 of R1589W, and a model prediction was achieved with the I-TASSER tool using the amino acid sequence of domain Ig-like 16. Once models were available, mutations were introduced with SDM and SwissPDB Viewer. The effects of variations on the two domains' OBSCURIN 3D structure and function were calculated. Results showed that two variations decrease the stability of domain structures and thus might cause loss-of-function on the protein. The pathogenicity possibility of these variations was increased.

The following approach was to reanalyze the ION Torrent Sequence with the recent (up-to-date) genome build to seek new variation possibilities mapping to different transcripts and chromosomal positions in the hg38 build. Once chromosomal positions of variants in each sample were transformed from hg19 to hg38, data was annotated with Ensembl VEP online tool, and LOH on trio sample analysis was carried out manually. The final candidate list of variants was the same, whereas chromosomal positions, gene transcript ids, and population frequencies were different. High-frequency variants were still the same; finally, the same three variants were assessed as pathogenic.

Genome conversion might reveal new information and ease reanalysis jobs in the future for new literature inputs with hg38 annotations. However, this approach could not help us reach our aim to detect the exact genetic basis of unsolved rare disease cases. The difference it could make in our case is to have ready-to-analyze data in the future with hg38 annotations. We only had reliable variant call annotations aligned against the manually curated reference genome.

Lastly, resequencing samples with recent technologies might have benefits over other approaches. Even the same kit and sequencing platform could have been improved over time. In our case, we had a chance to resequence our same old DNA samples with another platform, Illumina NovaSeq 6000. Two sequencing platforms have significantly different approaches. Ion S5 uses Semiconductor Sequencing technology, whereas NovaSeq 6000 Sequencing-by-Synthesis (Feng et al., 2016; Kim et al., 2021). Both technologies have advantages and disadvantages of their own; therefore, combining both results from these sequencing platforms might reveal additional variants or increase the reliability of previously detected candidate variants even though it raises costs. In our case, joining two exome sequencing data generated via two different technologies assisted us in overcoming the beforementioned disadvantages platforms bring along.

There were additional variants captured with the NovaSeq instrument. Once the same filters we used before were applied and variants compared with the literature, the final candidate list was the same. Data sequenced with NovaSeq had deeper intronic variants, higher coverage, and average read depths than Ion S5. Since our *OBSCN* variant p.R3252L was located near splice sites, read quality and reliability were essential parameters. Read lengths were 200 bp average in Ion S5 data and 150 bp average in NovaSeq 6000. Read fragments were distributed evenly in NovaSeq 6000 data; therefore, high GC percentage

regions were also covered genuinely. ION Torrent Sequence data involved approximately 13000 genes; Illumina Sequence data involved approximately 19000 genes. Thus, resequencing spanned broader coverage, and uncovered genes were ready to analyze. Considerably higher gene coverage could unhide potential variants.

In our case, germline variations were considered with autosomal recessive inheritance mode. Common filters were read quality, quality by depth, allele frequency, ACMG pathogenicity, and pathogen identification via bioinformatic tools. Homozygous or compound heterozygous variants in siblings and heterozygous variants that parents do not carry are considered. Parents were known to be consanguineous and did not share any phenotypic features of the sibling's condition. Allele frequency of variants in public databases played a vital role in this case. The reason for using a stricter allele frequency filter is that rare diseases are also expected to be caused by rare variants. Otherwise, we would have seen those diseases more frequently than now.

Not all diseases follow Mendelian inheritance or are caused by single nucleotide mutations in germline genetic material. Multiple genes could be involved in conditions' progress, and multiple variations could produce the final phenotype. Germline mosaicism could be another reason, clinical information about parents and the prognosis of siblings should be investigated and reevaluated to avoid missing critical phenotypical data. Other pathways in the organism might cause this genetic condition. Besides all assumptions, the exact reason for this rare disease remains unsolved. Genomic analysis is only one way to search for disease diagnosis. Different layers in protein function pathways might have been altered and caused consequences as the disease.

After all, approaches proposed in this thesis, reanalyzing with updated reference assembly and resequencing with another NGS platform supported the final list of variants and narrowed five into three main variants. Two variants on a large protein called OBSCURIN, and another variant on COENZYME Q8A protein previously related with similar phenotypic features are proposed as prospective variants related to our case.

One of the motivations for this thesis study was to lower the durations and costs of following functional studies. It is hoped that work achieved in this study could help further investigations on this rare disease, Rhizomelic dysplasia Patterson Lowry type. Additional molecular genetic investigations are needed to solve this case.

Since our analyses were mainly focused on variants located in exonic regions and splice sites, deep intronic variations and transcriptome level changes were uncovered. In future studies, WGS and RNAseq analyses will be conducted to cover wider genetic regions and other types of variations.

REFERENCES

Austin-Tse, C. A., Jobanputra, V., Perry, D. L., Bick, D., Taft, R. J., Venner, E., Gibbs, R. A., Young, T., Barnett, S., Belmont, J. W., Boczek, N., Chowdhury, S., Ellsworth, K. A., Guha, S., Kulkarni, S., Marcou, C., Meng, L., Murdock, D. R., Rehman, A. U., Spiteri, E., … Medical Genome Initiative* (2022). Best practices for the interpretation and reporting of clinical whole genome sequencing. *N.P.J. genomic medicine*, *7*(1), 27. https://doi.org/10.1038/s41525-022-00295-z

Bax, B. E. (2021). Biomarkers in Rare Diseases. *International Journal of Molecular Sciences*, *22*(2), 673. https://doi.org/10.3390/ijms22020673

Boycott, K. M., Vanstone, M. R., Bulman, D. E., & MacKenzie, A. E. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature reviews. Genetics*, *14*(10), 681–691. https://doi.org/10.1038/nrg3555

Bragin, E., Chatzimichali, E. A., Wright, C. F., Hurles, M. E., Firth, H. V., Bevan, A. P., & Swaminathan, G. J. (2014). DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic acids research, 42*(Database issue), D993–D1000. https://doi.org/10.1093/nar/gkt937

Buske, O. J., Girdea, M., Dumitriu, S., Gallinger, B., Hartley, T., Trang, H., Misyura, A., Friedman, T., Beaulieu, C., Bone, W. P., Links, A. E., Washington, N. L., Haendel, M. A., Robinson, P. N., Boerkoel, C. F., Adams, D., Gahl, W. A., Boycott, K. M., & Brudno, M. (2015). PhenomeCentral: a portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases. *Human mutation, 36*(10), 931–940. https://doi.org/10.1002/humu.22851

Cabrera-Serrano, M., Caccavelli, L., Savarese, M., Vihola, A., Jokela, M., Johari, M., Capiod, T., Madrange, M., Bugiardini, E., Brady, S., Quinlivan, R., Merve, A., Scalco, R., Hilton-Jones, D., Houlden, H., Ibrahim Aydin, H., Ceylaner, S., Vockley, J., Taylor, R. L., Folland, C., … Ravenscroft, G. (2021). Bi-allelic loss-of-function OBSCN variants predispose individuals to severe recurrent

rhabdomyolysis. Brain: a journal of neurology, awab484. Advance online publication. https://doi.org/10.1093/brain/awab484

Chang P. L. (2005). Clinical bioinformatics. *Chang Gung medical journal*, *28*(4), 201–211

Cingolani, P., Platts, A., Wang, l., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Cirino, A. L., & Ho, C. (2008). Hypertrophic Cardiomyopathy Overview. In M. P. Adam (Eds.) et. al., *GeneReviews*®. University of Washington, Seattle.

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Kähäri, A. K., Keenan, S., … Flicek, P. (2015). Ensembl 2015. *Nucleic acids research*, *43*(Database issue), D662–D669. https://doi.org/10.1093/nar/gku1010

Damar, Ç., Boyunağa, Ö., Derinkuyu, B. E., Battaloğlu, N., & Ezgü, F. S. (2014). Radiologic findings of Patterson-Lowry rhizomelic dysplasia in two sisters. *Skeletal radiology*, *43*(11), 1651–1654. https://doi.org/10.1007/s00256-014-1957-8

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Daoud, H., Luco, S. M., Li, R., Bareke, E., Beaulieu, C., Jarinova, O., Carson, N., Nikkel, S. M., Graham, G. E., Richer, J., Armour, C., Bulman, D. E., Chakraborty, P., Geraghty, M., Lines, M. A., Lacaze-Masmonteil, T., Majewski, J., Boycott, K. M., & Dyment, D. A. (2016). Next-generation sequencing for diagnosis of rare diseases in the neonatal intensive care unit. *CMAJ: Canadian Medical Association journal = journal de l'Association medicale canadienne*, *188*(11), E254–E260. https://doi.org/10.1503/cmaj.150823

Fairley, S., Lowy-Gallego, E., Perry, E., & Flicek, P. (2020). The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic acids research, 48*(D1), D941–D947. https://doi.org/10.1093/nar/gkz836

Feng, W., Zhao, S., Xue, D., Song, F., Li, Z., Chen, D., He, B., Hao, Y., Wang, Y., & Liu, Y. (2016). Improving alignment accuracy on homopolymer regions for semiconductor-based sequencing technologies. *B.M.C. Genomics*, *17*(S7). https://doi.org/10.1186/s12864-016-2894-9

Firth, H. V., Richards, S. M., Bevan, A. P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R. M., & Carter, N. P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American journal of human genetics, 84*(4), 524–533. https://doi.org/10.1016/j.ajhg.2009.03.010

Franceschini, P., Licata, D., Guala, A., Ingrosso, G., Di Cara, G., & Franceschini, D. (2004). Patterson-Lowry rhizomelic dysplasia: report of two new patients. *American journal of medical genetics. Part A*, *127A*(1), 86–92. https://doi.org/10.1002/ajmg.a.20638

Girdea, M., Dumitriu, S., Fiume, M., Bowdin, S., Boycott, K. M., Chénier, S., Chitayat, D., Faghfoury, H., Meyn, M. S., Ray, P. N., So, J., Stavropoulos, D. J., & Brudno, M. (2013). PhenoTips: patient phenotyping software for clinical and research use. *Human mutation, 34*(8), 1057–1065. https://doi.org/10.1002/humu.22347

GlobalGenes. (n.d.). *RARE Disease Facts*. Retrieved July 20, 2022, from https://globalgenes.org/rare-disease-facts/

Guo, Y., Ye, F., Sheng, Q., Clark, T., & Samuels, D. C. (2014). Three-stage quality control strategies for DNA resequencing data. *Briefings in bioinformatics*, *15*(6), 879–889. https://doi.org/10.1093/bib/bbt069

Haendel, M., Vasilevsky, N., Unni, D., Bologa, C., Harris, N., Rehm, H., Hamosh, A., Baynam, G., Groza, T., McMurry, J., Dawkins, H., Rath, A., Thaxon, C., Bocci, G., Joachimiak, M. P., Köhler, S., Robinson, P. N., Mungall, C., & Oprea, T. I. (2020). How many rare diseases are there?. *Nature reviews. Drug discovery, 19*(2), 77–78. https://doi.org/10.1038/d41573-019-00180-y

Holtsträter, C., Schrörs, B., Bukur, T., & Löwer, M. (2020). Bioinformatics for Cancer Immunotherapy. *Methods in molecular biology (Clifton, N.J.)*, *2120*, 1–9. https://doi.org/10.1007/978-1-0716-0327-7_1

Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Human immunology, 82(*11), 801–811. https://doi.org/10.1016/j.humimm.2021.02.012

Jacobsen, J. C., Whitford, W., Swan, B., Taylor, J., Love, D. R., Hill, R., Molyneux, S., George, P. M., Mackay, R., Robertson, S. P., Snell, R. G., & Lehnert, K. (2017).

Compound Heterozygous Inheritance of Mutations in Coenzyme Q8A Results in Autosomal Recessive Cerebellar Ataxia and Coenzyme Q10 Deficiency in a Female Sib-Pair. *JIMD Reports*, 31–36. https://doi.org/10.1007/8904_2017_73

Javitt, M. J., Vanner, E. A., Grajewski, A. L., & Chang, T. C. (2022). Evaluation of a computer-based facial dysmorphology analysis algorithm (Face2Gene) using standardized textbook photos. *Eye (London, England)*, *36*(4), 859–861. https://doi.org/10.1038/s41433-021-01563-5

Kamoda, T., Nakajima, R., Matsui, A., & Nishimura, G. (2001). Patterson-Lowry rhizomelic dysplasia: a potentially lethal bone dysplasia?. *Pediatric radiology*, *31*(2), 81–83. https://doi.org/10.1007/s002470000401

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., Walters, R. K., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Karlin S. (2005). Statistical signals in bioinformatics. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(38), 13355–13362. https://doi.org/10.1073/pnas.0501804102

Kaufmann, P., Pariser, A. R., & Austin, C. (2018). From scientific discovery to treatments for rare diseases – the view from the National Center for Advancing Translational Sciences – Office of Rare Diseases Research. *Orphanet Journal of Rare Diseases, 13(1).* doi:10.1186/s13023-018-0936-x

Kim, H. M., Jeon, S., Chung, O., Jun, J. H., Kim, H. S., Blazyte, A., Lee, H. Y., Yu, Y., Cho, Y. S., Bolser, D. M., & Bhak, J. (2021). Comparative analysis of 7 short-read sequencing platforms using the Korean Reference Genome: MGI and Illumina sequencing benchmark for whole-genome sequencing. *GigaScience*, *10*(3), giab014. https://doi.org/10.1093/gigascience/giab014

Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., Mundlos, C., Horn, D., Mundlos, S., & Robinson, P. N. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *American journal of human genetics*, *85*(4), 457–464. https://doi.org/10.1016/j.ajhg.2009.09.003

Lagier-Tourenne, C., Tazir, M., López, L. C., Quinzii, C. M., Assoum, M., Drouot, N., Busso, C., Makri, S., Ali-Pacha, L., Benhassine, T., Anheim, M., Lynch, D. R., Thibault, C., Plewniak, F., Bianchetti, L., Tranchant, C., Poch, O., DiMauro, S., Mandel, J. L., . . . Koenig, M. (2008). ADCK3, an Ancestral Kinase, Is Mutated in a Form of Recessive Ataxia Associated with Coenzyme Q10 Deficiency. *The*

American Journal of Human Genetics, *82*(3), 661–672. https://doi.org/10.1016/j.ajhg.2007.12.024

Laurie, S., Piscia, D., Matalonga, L., Corvó, A., Fernández-Callejo, M., Garcia-Linares, C., Hernandez-Ferrer, C., Luengo, C., Martínez, I., Papakonstantinou, A., Picó-Amador, D., Protasio, J., Thompson, R., Tonda, R., Bayés, M., Bullich, G., Camps-Puchadas, J., Paramonov, I., Trotta, J. R., Alonso, A., … Beltran, S. (2022). The RD-Connect Genome-Phenome Analysis Platform: Accelerating diagnosis, research, and gene discovery for rare diseases. *Human mutation, 43*(6), 717–733. https://doi.org/10.1002/humu.24353

Lee, H., Deignan, J. L., Dorrani, N., Strom, S. P., Kantarci, S., Quintero-Rivera, F., Das, K., Toy, T., Harry, B., Yourshaw, M., Fox, M., Fogel, B. L., Martinez-Agosto, J. A., Wong, D. A., Chang, V. Y., Shieh, P. B., Palmer, C. G., Dipple, K. M., Grody, W. W., Vilain, E., … Nelson, S. F. (2014). Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA*, *312*(18), 1880–1887. https://doi.org/10.1001/jama.2014.14604

Levy, S. E., & Boone, B. E. (2019). Next-Generation Sequencing Strategies. *Cold Spring Harbor perspectives in medicine*, *9*(7), a025791. https://doi.org/10.1101/cshperspect.a025791

Li, K., Du, Y., Li, L., & Wei, D. Q. (2020). Bioinformatics Approaches for Anti-cancer Drug Discovery. *Current drug targets*, *21*(1), 3–17. https://doi.org/10.2174/1389450120666190923162203

Li, Q., & Wang, K. (2017). InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *American journal of human genetics*, *100*(2), 267–280. https://doi.org/10.1016/j.ajhg.2017.01.004

Liu, Y. T., Hersheson, J., Plagnol, V., Fawcett, K., Duberley, K. E. C., Preza, E., Hargreaves, I. P., Chalasani, A., Laura, M., Wood, N. W., Reilly, M. M., & Houlden, H. (2013). Autosomal-recessive cerebellar ataxia caused by a novel ADCK3 mutation that elongates the protein: clinical, genetic and biochemical characterisation. *Journal of Neurology, Neurosurgery & Psychiatry*, *85*(5), 493–498. https://doi.org/10.1136/jnnp-2013-306483

Liu, Y., Hu, X., Han, C., Wang, L., Zhang, X., He, X., & Lu, X. (2015). Targeting tumor suppressor genes for cancer therapy. *BioEssays: news and reviews in molecular, cellular and developmental biology*, *37*(12), 1277–1286. https://doi.org/10.1002/bies.201500093

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. *Genome biology*, *20*(1), 246. https://doi.org/10.1186/s13059-019-1828-7

Manring, H. R., Carter, O. A., & Ackermann, M. A. (2017). Obscure functions: the location-function relationship of obscurins. *Biophysical reviews*, *9*(3), 245–258. https://doi.org/10.1007/s12551-017-0254-x

Marston, S. (2017). Obscurin variants and inherited cardiomyopathies. *Biophysical reviews*, *9*(3), 239–243. https://doi.org/10.1007/s12551-017-0264-8

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome biology*, *17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

Metzker M. L. (2010). Sequencing technologies - the next generation. *Nature reviews. Genetics*, *11*(1), 31–46. https://doi.org/10.1038/nrg2626

Migita, K., Izumi, Y., Jiuchi, Y. et al. Familial Mediterranean fever is no longer a rare disease in Japan. *Arthritis Res Ther 18, 175* (2016). https://doi.org/10.1186/s13075-016-1071-5

Morales, J., Pujar, S., Loveland, J. E., Astashyn, A., Bennett, R., Berry, A., Cox, E., Davidson, C., Ermolaeva, O., Farrell, C. M., Fatima, R., Gil, L., Goldfarb, T., Gonzalez, J. M., Haddad, D., Hardy, M., Hunt, T., Jackson, J., Joardar, V. S., Kay, M., … Murphy, T. D. (2022). A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature*, *604*(7905), 310–315. https://doi.org/10.1038/s41586-022-04558-8

Entry - 601438 - RHIZOMELIC DYSPLASIA, PATTERSON-LOWRY TYPE - *OMIM*. Omim.org. (2022). Retrieved August 2022, from https://www.omim.org/entry/601438.

Nair, P., Lama, M., El-Hayek, S., Abou Sleymane, G., Stora, S., Obeid, M., Al-Ali, M. T., Delague, V., & Mégarbané, A. (2019). *COQ8A* and *MED25* Mutations in a Child with Intellectual Disability, Microcephaly, Seizures, and Spastic Ataxia: Synergistic Effect of Digenic Variants?. *Molecular syndromology*, *9*(6), 319–323. https://doi.org/10.1159/000494465

Nellåker, C., Alkuraya, F. S., Baynam, G., Bernier, R. A., Bernier, F., Boulanger, V., Brudno, M., Brunner, H. G., Clayton-Smith, J., Cogné, B., Dawkins, H., deVries, B., Douzgou, S., Dudding-Byth, T., Eichler, E. E., Ferlaino, M., Fieggen, K., Firth, H. V., FitzPatrick, D. R., Gration, D., … Minerva Consortium (2019). Enabling Global Clinical Collaborations on Identifiable Patient Data: The Minerva Initiative. *Frontiers in genetics, 10,* 611. https://doi.org/10.3389/fgene.2019.00611

Neveling, K., Collin, R. W., Gilissen, C., van Huet, R. A., Visser, L., Kwint, M. P., Gijsen, S. J., Zonneveld, M. N., Wieskamp, N., de Ligt, J., Siemiatkowska, A. M., Hoefsloot, L. H., Buckley, M. F., Kellner, U., Branham, K. E., den Hollander, A.

I., Hoischen, A., Hoyng, C., Klevering, B. J., van den Born, L. I., … Scheffer, H. (2012). Next-generation genetic testing for retinitis pigmentosa. *Human mutation*, *33*(6), 963–972. https://doi.org/10.1002/humu.22045

Niemi, M., Martin, H. C., Rice, D. L., Gallone, G., Gordon, S., Kelemen, M., McAloney, K., McRae, J., Radford, E. J., Yu, S., Gecz, J., Martin, N. G., Wright, C. F., Fitzpatrick, D. R., Firth, H. V., Hurles, M. E., & Barrett, J. C. (2018). Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. *Nature*, *562*(7726), 268–271. https://doi.org/10.1038/s41586-018-0566-4

Orphanet. (n.d.). *Orphan drugs in Japan.* Orphanet: About Orphan Drugs. Retrieved July 20, 2022, from https://www.orpha.net/consor/cgi-bin/Education_AboutOrphanDrugs.php?lng=EN&stapage=ST_EDUCATION_EDUCATION_ABOUTORPHANDRUGS_JAP

Patterson, C., & Lowry, R. B. (1975). A New Dwarfing Syndrome with Extreme Shortening of Humeri and Severe Coxa Vara. *Radiology, 114(2), 341–342.* doi:10.1148/114.2.341

Pengelly, R. J., Alom, T., Zhang, Z., Hunt, D., Ennis, S., & Collins, A. (2017). Evaluating phenotype-driven approaches for genetic diagnoses from exomes in a clinical setting. *Scientific reports*, *7*(1), 13509. https://doi.org/10.1038/s41598-017-13841-y

Perry, N. A., Ackermann, M. A., Shriver, M., Hu, L. Y., & Kontrogianni-Konstantopoulos, A. (2013). Obscurins: unassuming giants enter the spotlight. *IUBMB life*, *65*(6), 479–486. https://doi.org/10.1002/iub.1157

Qin D. Next-generation sequencing and its clinical application. *Cancer Biol Med. 2019 Feb;16*(1):4-10. doi: 10.20892/j.issn.2095-3941.2018.0055.

Qiu, J., Wu, L., Chang, Y., Sun, H., & Sun, J. (2021). Alternative splicing transitions associate with emerging atrophy phenotype during denervation-induced skeletal muscle atrophy. *Journal of cellular physiology*, *236*(6), 4496–4514. https://doi.org/10.1002/jcp.30167

Rahit, K., & Tarailo-Graovac, M. (2020). Genetic Modifiers and Rare Mendelian Disease. *Genes*, *11*(3), 239. https://doi.org/10.3390/genes11030239

Rajendran, B. K., & Deng, C. X. (2017). A comprehensive genomic meta-analysis identifies confirmatory role of *OBSCN*gene in breast tumorigenesis. *Oncotarget*, *8*(60), 102263–102276. https://doi.org/10.18632/oncotarget.20404

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H. L., & ACMG Laboratory Quality Assurance Committee (2015). 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. *Genetics in medicine: official journal of the American College of Medical Genetics, 17*(5), 405–424. https://doi.org/10.1038/gim.2015.30

Robinson, P. N., Köhler, S., Oellrich, A., Sanger Mouse Genetics Project, Wang, K., Mungall, C. J., Lewis, S. E., Washington, N., Bauer, S., Seelow, D., Krawitz, P., Gilissen, C., Haendel, M., & Smedley, D. (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome research*, *24*(2), 340–348. https://doi.org/10.1101/gr.160325.113

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology, 29*(1), 24–26. https://doi.org/10.1038/nbt

Satman, İ., Güdük, Ö., Yemenci, M., & Ertürk, N. (2019, September). *Nadir Hastalıklar Raporu.* TÜSEB. https://www.tuseb.gov.tr/tuhke/uploads/genel/files/haberler/nadir_hastaliklar_raporu.pdf

Shakya, M., Ahmed, S. A., Davenport, K. W., Flynn, M. C., Lo, C. C., & Chain, P. (2020). Standardized phylogenetic and molecular evolutionary analysis applied to species across the microbial tree of life. *Scientific reports*, *10*(1), 1723. https://doi.org/10.1038/s41598-020-58356-1

Sifrim, A., Popovic, D., Tranchevent, L. C., Ardeshirdavani, A., Sakai, R., Konings, P., Vermeesch, J. R., Aerts, J., De Moor, B., & Moreau, Y. (2013). eXtasy: variant prioritization by genomic data fusion. *Nature methods*, *10*(11), 1083–1084. https://doi.org/10.1038/nmeth.2656

Somarelli, J. A., Gardner, H., Cannataro, V. L., Gunady, E. F., Boddy, A. M., Johnson, N. A., Fisk, J. N., Gaffney, S. G., Chuang, J. H., Li, S., Ciccarelli, F. D., Panchenko, A. R., Megquier, K., Kumar, S., Dornburg, A., DeGregori, J., & Townsend, J. P. (2020). Molecular Biology and Evolution of Cancer: From Discovery to Action. *Molecular biology and evolution*, *37*(2), 320–326. https://doi.org/10.1093/molbev/msz242

Strande, N. T., & Berg, J. S. (2016). Defining the Clinical Value of a Genomic Diagnosis in the Era of Next-Generation Sequencing. *Annual review of genomics and human genetics*, *17*, 303–332. https://doi.org/10.1146/annurev-genom-083115-022348

Sun, Y., Ruivenkamp, C. A., Hoffer, M. J., Vrijenhoek, T., Kriek, M., van Asperen, C. J., den Dunnen, J. T., & Santen, G. W. (2015). Next-generation diagnostics: gene panel, exome, or whole genome?. *Human mutation*, *36*(6), 648–655. https://doi.org/10.1002/humu.22783

Tabei, Y., Kotera, M., Sawada, R., & Yamanishi, Y. (2019). Network-based characterization of drug-protein interaction signatures with a space-efficient approach. *BMC systems biology*, *13*(Suppl 2), 39. https://doi.org/10.1186/s12918-019-0691-1

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H. C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., Fulton, R. S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K. M., Auger, K., Chow, W., Collins, J., Harden, G., … Church, D. M. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome research*, *27*(5), 849–864. https://doi.org/10.1101/gr.213611.116

Wang, J., Raskin, L., Samuels, D. C., Shyr, Y., & Guo, Y. (2015). Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics (Oxford, England)*, *31*(3), 318–323. https://doi.org/10.1093/bioinformatics/btu668

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, *38*(16), e164. https://doi.org/10.1093/nar/gkq603

Wang, X., & Liotta, L. (2011). Clinical bioinformatics: a new emerging science. *Journal of clinical bioinformatics*, *1*(1), 1. https://doi.org/10.1186/2043-9113-1-1

Williams, M. S., Josephson, K. D., & Pauli, R. M. (1995). Patterson-Lowry rhizomelic dysplasia: a possible second example. *Clinical dysmorphology*, *4*(3), 216–221.

Wu, F., Mueller, L. A., Crouzillat, D., Pétiard, V., & Tanksley, S. D. (2006). Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. *Genetics*, *174*(3), 1407–1420. https://doi.org/10.1534/genetics.106.062455

Wu, H. Y., Chiang, C. W., & Li, L. (2014). Text mining for drug-drug interaction. *Methods in molecular biology (Clifton, N.J.)*, *1159*, 47–75. https://doi.org/10.1007/978-1-4939-0709-0_4

Yohe, S., & Thyagarajan, B. (2017). Review of Clinical Next-Generation Sequencing. *Archives of pathology & laboratory medicine*, *141*(11), 1544–1557. https://doi.org/10.5858/arpa.2016-0501-RA

Zemojtel, T., Köhler, S., Mackenroth, L., Jäger, M., Hecht, J., Krawitz, P., Graul-Neumann, L., Doelken, S., Ehmke, N., Spielmann, M., Oien, N. C., Schweiger, M. R., Krüger, U., Frommer, G., Fischer, B., Kornak, U., Flöttmann, R., Ardeshirdavani, A., Moreau, Y., Lewis, S. E., … Robinson, P. N. (2014). Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Science translational medicine*, *6*(252), 252ra123. https://doi.org/10.1126/scitranslmed.3009262

Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J. P., & Wang, L. (2013). CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, *30*(7), 1006–1007. https://doi.org/10.1093/bioinformatics/btt730

TEZ İZİN FORMU / THESIS PERMISSION FORM

ENSTİTÜ / INSTITUTE

Fen Bilimleri Enstitüsü / Graduate School of Natural and Applied Sciences ☐

Sosyal Bilimler Enstitüsü / Graduate School of Social Sciences ☐

Uygulamalı Matematik Enstitüsü / Graduate School of Applied Mathematics ☐

Enformatik Enstitüsü / Graduate School of Informatics ☑

Deniz Bilimleri Enstitüsü / Graduate School of Marine Sciences ☐

YAZARIN / AUTHOR

Soyadı / Surname : YAZAR
Adı / Name : ÖMER FARUK
Bölümü / Department : BİYOENFORMATİK

TEZİN ADI / TITLE OF THE THESIS (**İngilizce** / English) : ..... UTILITY OF RESEQUENCING AND REANALYSIS FOR UNSOLVED RARE DISEASES……
.....
.....

TEZİN TÜRÜ / DEGREE: Yüksek Lisans / Master ☑ **Doktora** / PhD ☐

1. **Tezin tamamı dünya çapında erişime açılacaktır. /** Release the entire work immediately for access worldwide. ☑
2. **Tez iki yıl süreyle erişime kapalı olacaktır.** / Secure the entire work for patent and/or proprietary purposes for a period of **two year. *** ☐
3. **Tez altı ay süreyle erişime kapalı olacaktır.** / Secure the entire work for period of **six months**. ***** ☐

***** *Enstitü Yönetim Kurulu Kararının basılı kopyası tezle birlikte kütüphaneye teslim edilecektir.*
A copy of the Decision of the Institute Administrative Committee will be delivered to the library together with the printed thesis.

Yazarın imzası / Signature ..... **Tarih** / Date ……..