Ziya Furkan ZORLUER

A Master's Thesis

AGU 2022

DISCOVERING NEW PATHOGENIC VARIANTS BY IN SILICO ANALYSIS

A THESIS
SUBMITTED TO THE DEPARTMENT OF BIOENGINEERING
AND THE GRADUATE SCHOOL OF ENGINEERING AND
SCIENCE OF ABDULLAH GUL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER

By
Ziya Furkan ZORLUER
September, 2022

DISCOVERING NEW PATHOGENIC VARIANTS BY IN SILICO ANALYSIS

A THESIS
SUBMITTED TO THE DEPARTMENT OF BIOENGINEERING AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE OF ABDULLAH GUL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER

By
Ziya Furkan ZORLUER
September, 2022

SCIENTIFIC ETHICS COMPLIANCE

I hereby declare that all information in this document has been obtained 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 materials and results that are not original to this work.

Name-Surname: Ziya Furkan ZORLUER

REGULATORY COMPLIANCE

M.Sc. thesis titled Discovering new pathogenic variants by in silico analysis has been prepared in accordance with the Thesis Writing Guidelines of the Abdullah Gül University, GraduateSchool of Engineering & Science.

Prepared by	Advisor
Ziya Furkan ZORLUER	Assist. Prof. Dr. Oktay I. KAPLAN

Head of the Bioengineering Program

Prof. Dr. Sevil Dinçer İşoğlu

ACCEPTANCE AND APPROVAL

M.Sc. thesis titled Discovering new pathogenic variants by in silico analysis and prepared by Ziya Furkan ZORLUER has been accepted by the jury in the Bioengineering Graduate Program at Abdullah Gül University, Graduate School of Engineering & Science.

26/09/2022

(Thesis Defense Exam Date)

JURY:

Advisor: Assist. Prof. Dr. Oktay İsmail KAPLAN ……..

Member: Assoc. Prof. Dr. M. Duygu Saçar Demirci……

Member: Assoc. Prof. Dr. Osman DOLUCA ……..

APPROVAL:

The acceptance of this M.Sc. thesis has been approved by the decision of the Abdullah Gül University, Graduate School of Engineering & Science, Executive Board dated /2022 and numbered …..……. .

…….. /…….. / ……..

Graduate School Dean
Prof. Dr. Irfan ALAN

ABSTRACT

DISCOVERING NEW PATHOGENIC VARIANTS BY IN SILICO ANALYSIS

Ziya Furkan ZORLUER
MSc in Bioengineering
Advisor: Assist. Prof. Dr Oktay İsmail KAPLAN

September, 2022

Inherited diseases are health problems caused by one or more abnormalities in the genome. It can be caused by changes in a single gene (monogenic) or multiple genes (polygenic), or by a damage on chromosomes. Genetic variation is the differences in the DNA sequences that can be observed within a species or in alleles. Evaluation of genetic variants, together with reported phenotypic or pathogenic annotations from non-human organisms, facilitates the comparison of these variants with their human counterparts. In this work, we combined pathogenic and phenotypic annotations with variants, and these phenotypic orthologous variants from seven organisms can provide clues to the functional consequences of human genetic variants.

Keywords Pathogenic variants, In silico analysis, New candidate pathogenic variants

ÖZET

İN SİLİKO YÖNTEMLERLE YENİ PATOJENİK VARYANTLAR BULMAK

Ziya Furkan ZORLUER
Biyomühendislik Anabilim Dalı Yüksek Lisans
Tez Yöneticisi: Dr.Öğr. Üyesi. Oktay İ. KAPLAN
Haziran, 2022

Kalıtsal hastalıklar, genomdaki bir veya daha fazla anormalliğin neden olduğu sağlık sorunlarıdır. Tek bir gendeki (monogenik) veya çoklu genlerdeki (poligenik) değişikliklerden veya kromozomlardaki bir hasardan kaynaklanabilir. Genetik varyasyonlar ise bir tür içinde veya alellerde gözlemlenebilen DNA dizilerindeki farklılıklardır ve bu varyasyonlar patojenik veya fenotipik etkilere neden olabilir. İnsan olmayan organizmalardan alınan ve literatürde fenotipik veya patojenik olduğu bildirilen varyant verilerinin tüm varyantlarla eşleştirilip birlikte değerlendirilmesi, bu varyantların da insan eşdeğerleri ile karşılaştırılması birçok analiz ve tahmin için yardımcı olabilir. Bu çalışmada, patojenik ve fenotipik varyant verilerini varyantlarla birleştirdik ve yedi organizma için muadil insan varyantları ile eşleştirilmelerini sağladık. Elde edilen eşleşmiş fenotipik varyantların, insan genetik varyantlarının işlevsel sonuçlarına dair ipuçları sağlayabiliceği sonucuna vardık.

Anahtar kelimeler Patojenik varyanlar, in siliko analiz, yeni aday patojenik varyanlar

Acknowledgements

I would like to express my sincere gratitude to my supervisor Assist.Prof.Dr. Oktay İsmail KAPLAN for his understanding and supportive guidance during my M.Sc.

I also would like to thank Assist.Prof.Dr. Sebiha Cevik KAPLAN and all other Kaplan lab members for their supports.

I would like to thank Mustafa Samet PIR to his support and solutions whenever I need help.

Finally, I need to thank my family members Atiyye ZORLUER, İsmail ZORLUER, M. Esad ZORLUER and Özgül ZORLUER for their patience, support and encouragement during all these challenging times. And also want to thank my beautiful daughter Zeynep Itır ZORLUER for her adorable smiles.

Ziya Furkan ZORLUER

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATONS

CNV	Copy Number Variation
DNA	Deoxyribonucleic Acid
GWAS	Genome wide association
Kb	Kilobase
MSA	Multiple Sequence Alignment
NGS	Next Generation Sequencing
RNA	Ribonucleic Acid
RFLP	Restriction Fragment Length Polymorphism
SNP	Single Nucleotide Polymorphism

Chapter 1

1.Introduction

The demonstration of DNA as a genetic material by the experiment carried out by Oswald Avery, Colin MacLeod, and Maclyn McCarty in 1944 opened a novel avenue of research [1]. Many scientists attempted to understand how DNA stores genetic information. The excitement increased with the discovery of the structure of DNA in 1953 by James Watson and Francis Crick [2]. With the development of sequencing technologies in 1976, scientists aimed at generating the sequence of mysterious genetic material DNA. Frederick Sanger published the DNA sequencing method, later bringing him Nobel Prizes in 1977[3]. Sanger sequencing is one of the most widely used sequencing methods in which only a particular region of the DNA can be sequenced in a well. Thanks to new developments in sequencing techniques, the idea of sequencing the human genome started in 1984, and 6 years later, the Human Genome Project was initiated by the National Institutes of Health (NIH), aiming to sequence the human genome. Even though Sanger sequencing has been accelerated and performed in a practical manner with the help of developed automated systems, sequencing the entire genome of an organism with this system is both time-consuming and costly. This has been clearly understood thanks to the human genome project, which was completed by using the sanger method [4].

As a faster and more practical sequencing method, an entire genome or exome can be sequenced in a single well with the next-generation sequencing method (NGS). This allows us to read the whole genome or exome simultaneously. The analysis phase of a large amount of data that emerges from DNA readings is also critical. The cost of sequencing a whole genome is decreasing day by day, which means the proliferation of genome or exome data being sequenced and analyzed every day [5]. Next-generation sequencing can be performed not only on DNA but also with RNA, chromosome, or epigenetic targets such as methylation.

Thanks to all new sequencing tech, many projects were carried out, and our knowledge about the genetic material of different organisms and their similarity and even differences between each individual of the same species.

Every living organism carries their genetic material similar to their relatives. For example, the DNA of an individual Homo Sapiens is 99.9% similar to another Homo Sapiens individual. However, every individual can still be distinguished and differentiated from each other, thanks to the minor differences and changes in their DNA. A very small part of these changes are pathogenic, but almost all genetic diseases are caused by pathogenic changes that have taken place at the DNA level. It is essential to diagnose and understand these changes in DNA which can cause anomalies.

1.1 Variations

Genetic variation is the differences in the DNA sequences that can be observed within a species or in alleles. Nevertheless, the definition and categorization of genetic variation do not have a consensus to explain them clearly. Variations among the genomes can occur for many reasons. Such as single nucleotide changes in DNA, additions or deletions of different pieces in DNA, location change or duplication of existing pieces, etc. All these variations can extend from single nucleotide polymorphism to thousand base pair changes which can lead to diversity and also can cause phenotypic effects.

Our knowledge about human genetic variation and gene mapping of the human genome has dramatically increased in the last two decades. RFLP studies in the 1980s, Tandem repeat studies in the 1990s, SNP, indels, CNV, and loss of heterozygosity/homozygosity studies in the millennium and post-millennium era provided us massive genetic data which illuminate the relationship between many complex and monogenic human disease with specific variations.

1.1.1 Single Nucleotide Changes

As the name implies, single nucleotide changes are seen in a single base pair of the DNA sequences. Single nucleotide change can be seen in three forms. One nucleotide can be deleted or inserted into the wild-type sequence. They are called single nucleotide deletion or single nucleotide insertion, respectively (Figure 1.1.1).

Single Nucleotide Polymorphisms (SNP) are nucleotide changes on a single base pair in the DNA sequences and are used as genetic markers in many genetic studies (Figure 1.1.1) [6]. SNPs, seen every 200 to 300 base pairs throughout the genome, are used effectively in determining some genes related to specific diseases [7]. SNPs provide significant benefits in illuminating the susceptibility of individuals to the diseases,

differences in their response to treatments, and clinical dimensions of diseases. A broad spectrum of diseases has been identified by SNPs, such as diabetes, cancer, psychiatric diseases, etc. [8].

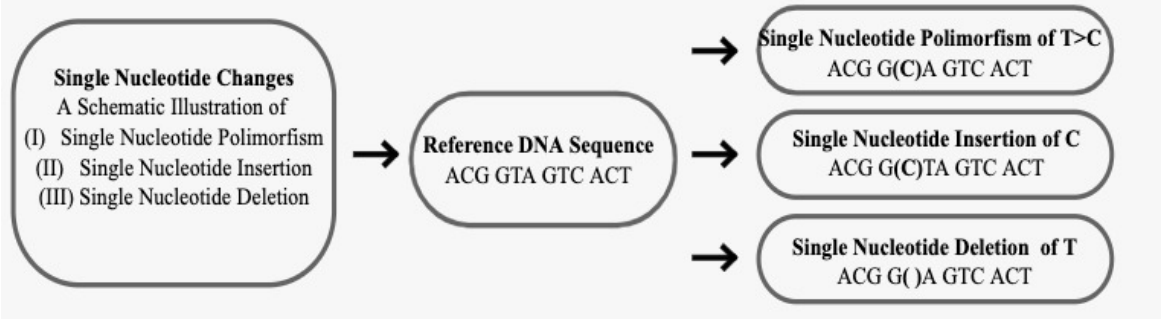

Figure 1.1.1 Types of single nucleotide changes

1.1.2 Structural variations

Differences in DNA segments that are 1 kb (kilobases) or more in size compared to the reference genome are called copy number variation (CNV) or copy neutral variations (Figure 1.1.2) [9]. CNVs can be observed in the form of deletion or duplication in the DNA sequence, and Various studies postulate that copy number variation may be responsible for part of the susceptibility to disease [10]. Copy neutral variations are observed in the form of inversion or translocation in a segment of DNA.

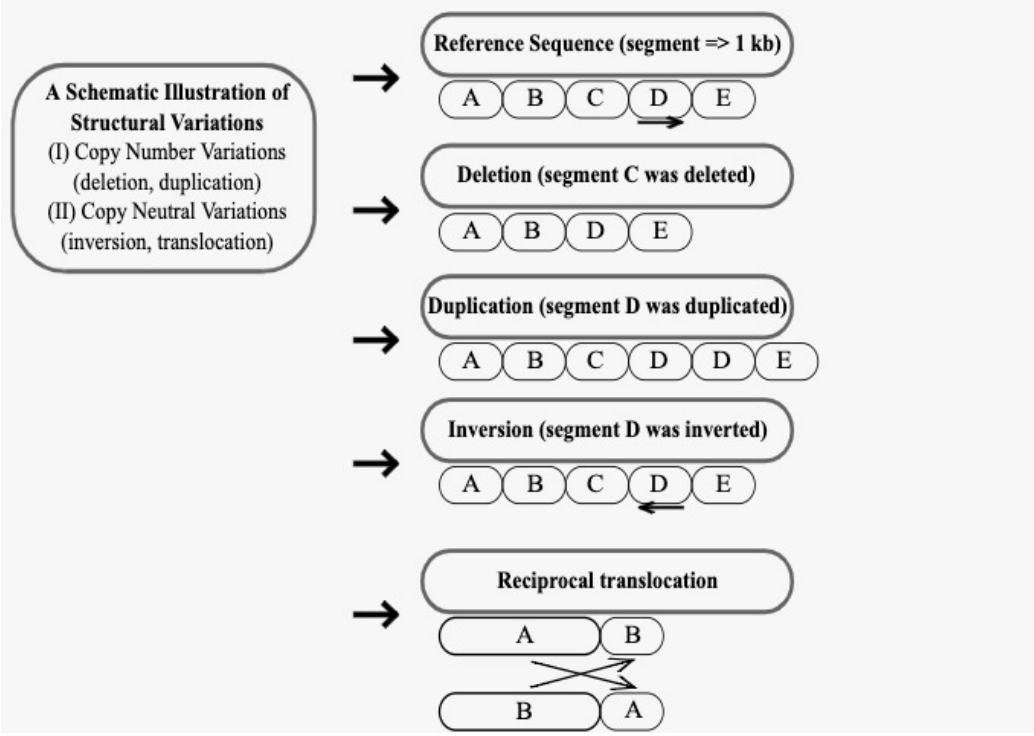

Figure 1.1.2 Types of structural variations

1.1.3 Short indels

Deletions and insertions are frequently seen as changes in the human genome which are called together as indels. DNA nucleotides' deletions and insertions are generally less than 1 kb in length [11]. Those deletions and insertions are classified as short indels (Figure 1.1.2).

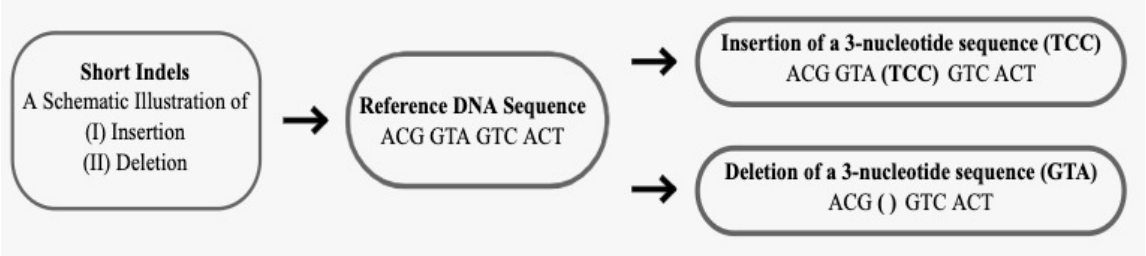

Figure 1.1.3 Types of Short indels

1.2 Obtaining Variant Data

Today, along with the many developments in DNA sequencing techniques and the increasing number of genome projects, studies to reveal DNA sequences are increasing and accelerating. To interpret genomic data, many genome projects of model organisms were carried out after the human genome project, and important information about the genome's organizational structure and evolutionary development was obtained.
Projects like; the 1000 genome project, HapMap, and ENCODE can be excellent examples of the effort to obtain more genetic variants of human beings.

1.2.1 HapMap Project

After completing the human genome project, The International HapMap project was initiated to understand the patterns of genetic variation in the human genome [12]. Two hundred seventy people are involved in the HapMap project by donating DNA samples. One-third of people from Nigeria, one-third of people from the east part of the World (Tokyo/Japan and Beijing/China), one-third of people from Northern and Western European originated U.S. residents [13]. The HapMap project was performed by using common SNPs (Single Nucleotide Polymorphism). By knowing the SNP of an individual, it is often possible to predict the alleles of SNPs that are close to each other [12]. Thus, based on the data obtained, it will be possible for researchers to select SNPs for disease-causing genes and to make their studies more effective [13]. The HapMap project has created great hope in the biomedical field around the world. In addition to biomedical research, it has contributed to identifying genes that respond to treatment and drugs,

especially in the diseased population [14]. Project data is available on the HapMap website for unlimited public use. Also included are interactive data browsing and analysis results not found elsewhere [12]. The HapMap will be used as a resource to facilitate future studies of health, disease, drug response, and genetic diversity [14].

1.2.2 ENCODE Project

The ENCODE project started in 2003 to identify all functional elements in the human genome sequence [15]. This international collaborative collaboration of research groups was funded by the National Human Genome Institute (NHGRI). Initially, the pilot project focused on 1% of the genome, then expanded to the whole genome in 2007 [16]. Encyclopedia of Model Organisms (modENCODE) was also created by NHGRI in 2007, and it was designed to characterize the genomes of Drosophila melanogaster (fruit fly), Caenorhabditis elegans (roundworm), and Mus musculus (mouse) [17], [18].

Thanks to the genetic and molecular biology experiments on mentioned species, important data on how the human genome works have been obtained [18]. It was revealed that 80% of the human genome was functional in at least one cell type in 1640 datasets in 147 different cell types using 24 standard experimental types. ENCODE data have been combined with SNPs identified by genome-wide association studies (GWAS) against attacks by much more complex diseases [18]. Typical disease analyzes can be performed by GWAS in larger regions thanks to the SNP. It can help identify disease-causing variants with good mapping techniques in functional or non-functional areas. The information combined with ENCODE data allows us to get an idea about the effect of a genetic variant on the genome sequence [19].

1.2.3 1000 Genome Project

The 1000 genomes project was performed by whole genome and exome sequencing of 1092 individuals from 14 populations [20]. The last phase of this project in June 2014 includes 26 populations of more than 2500 individuals. In particular, variants with more than 95% of the genomic regions and an allele frequency (polymorphism) of 1% or more were used.

1000 genome projects; represent a step in the right direction for the complete elucidation of the variation of the polymorphic human DNA sequence. The large dataset provided by the 1000 genomes project will enable more accurate localization of disease-associated

variants within the GWAS. A template will be provided for studies using the genome sequence data from the project. These data and the methods and applications developed to produce them will contribute to a much more comprehensive understanding of human history, evolution, disease, and the role of hereditary DNA variation [20].

1.3 Related Bioinformatics Tools and databases

Thanks to the breakthroughs in DNA sequencing technology, there is now a huge pile of information in diagnostic processes based on variant screening. If all the protein-coding regions (exome) of a human are sequenced, it is possible to detect approximately twenty thousand meaningful variants. While the majority of these variants are called benign or polymorphisms, a small number of them may be pathogenic. It is essential to use bioinformatics algorithms and bioinformatics databases in which clinical data on variants are compiled in order to detect these few variants that may be pathogenic among thousands of variants in the diagnostic processes.

1.3.1 Historical background of bioinformatics

One of the very early bioinformatics studies started with a computer program called comprotein, written in Fortran language, which identifies the primary structures of proteins by combining the short amino acid sequences obtained from Edman reactions in the article published by Margaret Dayhoff in 1962 and can run on the IBM 7090 [21]. Thus, the first de novo assembler algorithm in history began to be used. After the publication of the comprotein algorithm, the complete sequence of 65 proteins was published as a book (Dayhoff, 1965) [21]. This book is considered to be the first bioinformatics database in history [22]. This database was converted into a digital database within UniProt in 1983. Today, UniProtKB/Swiss-Prot contains records of more than 565,000 proteins.

After that algorithm was known by researchers, the power of computers was understood and used widely. As a consequence of the wide usage of computers and algorithms, the number of sequenced proteins began to increase. Hemoglobin proteins were isolated and sequenced from different species and published between 1962 and 1965. Thanks to this study, the rate of evolutionary change of a protein was measured [23].

Although computers and algorithms make sequencing easier, evaluation of the proteins and calculating the evolutionary value of the differences was not easy. Needleman-

Wunsch developed a dynamic programming algorithm that aligns two protein sequences globally as a solution to this problem [24]. For more than two sequences, multiple sequence alignment (MSA) was developed in 1985 [25]. BLAST, Gapped BLAST, and PSI-BLAST algorithms were developed between 1990-1997 [26].

1.3.2 Databases

World Wide Web technology, which was introduced to the world at the same time as the Human Genome Project, enabled the development of online databases that would enable access to the obtained biological data. During this period, databases such as NCBI (1994), Genomes (1995), PubMed (1997), and Human Genomes (1999) came online. In the 2000s, in line with the experience gained in the Human Genome Project, studies were carried out to reduce both the time and cost of genome-level sequencing studies.

The most critical databases in this field are EMBL-Bank, maintained by EBI (European Bioinformatics Institute, U.K.), GenBank, operated by the US-based NCBI (National Center for Biotechnology Information), and DDJB, based in Japan (Japan). DNA Data Bank, DNA Data Bank of Japan). These three databases collaborate by sharing the data they which contain. In addition, UCSC and Ensembl are also frequently used databases for both annotation and nucleotide and amino acid sequences. In addition to these, databases such as TrEMBL, Entrez Protein, and UniProt also contain data on amino acid sequences. Some of the bioinformatics databases are frequently used in studies. Although sequence databases are frequently used in the field of bioinformatics, there are structural databases containing data on the three-dimensional structures and models of biomolecules, as well as databases containing functional data such as genome, variant or protein annotation, pathway information, protein-drug interactions, alphafold which is also maintained by EBI can be an excellent example for this type of databases.

1.3.2.1 ClinVar

In addition to NCBI and PubMed, NIH supports many other programs like TOPMed and ClinVar. ClinVar is an open-access archive that provides human variations, human phenotypes, and their associations with supporting evidences. [27] Launched in 2013 at the National Center for Biotechnology Information for clinical geneticist and researchers. The database contains almost 600,000 submitted records from thousands of senders representing 430,000 unique variants. Submissions are collected and made downloadable

files via FTP on the ClinVar website (https://www.ncbi.nlm.nih.gov/clinvar/) [28].

1.3.2.2 TOPMed

Trans-Omics for Precision Medicine program (TOPMed) is another NIH-supported project which provides whole genome sequencing (WGS) and other omics data from existing studies in the literature, with the primary purpose of which is to advance scientific understanding of the fundamental biological processes underlying heart, lung, blood, and sleep (HLBS) disorders. TOPMed data are available as a series of "data freezes". TOPMed comprises approximately 180,000 participants from over 85 different studies of varying designs, and 60% are non-European ancestry [29]. In the TOPMed database, 811 million SNV and 66 million short insertion/deletion variants were identified [29]

1.3.2.3 gnomAD

The Genome Aggregation Database (gnomAD) was founded as Exome Aggregation Consortium (ExAC) by the scientist who wanted to aggregate and harmonize genome and exome sequencing data from different sequencing projects and make them available as a summary for the scientific community. [30] The database has different versions. For example, v2 release is composed of 125,748 exomes and 15,708 genomes (GRCh37) and v3.1 has 76,156 genomes (GRCh38). [30] All these data can be downloaded from their website (https://gnomad.broadinstitute.org/downloads).

1.3.2.4 COSMIC

COSMIC – the Catalogue of Somatic Mutations in Cancer – was launched in 2004 to collect and display information on somatic mutations in cancer. At the beginning, there were only data from four genes, HRAS, KRAS2, NRAS, and BRAF. [31] Over the years, its expansion continued rapidly and became the world's largest somatic mutation information database for human cancers. Currently, there are over 27,000 peer-reviewed papers that are curated manually by experts, and Over 37,000 genomes are available on their website (https://cancer.sanger.ac.uk/cosmic/download) [31]

1.3.2.5 Databases for non-human organisms

For the big data of model organisms, there should also be databases. To address this problem, there are many model organism databases for frequently used model organisms.

Such as the Saccharomyces Genome Database, which is a database that contains a lot of data about molecular biology and genetics of yeast. The Zebrafish Information Network (ZFIN) which about the zebrafish (Danio rerio) [32], FlyBase, which is about the insect family Drosophilidae [33] and Wormbase, which is about the Caenorhabditis elegans other related nematodes [34] can be another database example for widespread model organisms. For the other animals, there is a database of inherited disorders, other traits, and associated genes and variants, which is known as OMIA [35]. This helpful database includes 346 animal species.

1.3.3 Variant Prediction Tools

These projects and developments cause huge amounts of variant data, which relatively increases the need for prediction tools. Since most variant data don't have reported phenotypic outcomes, identifying deleterious variants or designing a wet lab study for every variant data in the literature is quite hard. Thanks to variant effect prediction tools, identification of genes and protein function would be easier, which can help the treatment of many human genetic diseases, and also researchers can conduct target-driven experiments by using variant effect estimation tools. To achieve this goal, many databases and prediction tools have been developed by scientists over recent years. Widely used variant effect prediction methods include SIFT [36], PolyPhen (v2) [37, 38], GERP++ [39, 40], Condel [41], CADD [42], fathmm [43] etc. Each tool uses different methods and algorithms for prediction and use datasets such as HumDiv [44], HumVar [45], Humsavar [46] and dbSNP [47]. Computational predictions generally make variant classifications, which may cause poor results. Disease risk inflation has been observed in the ClinVar and HumVar databases, whereby considerably fewer individuals in the general population are afflicted with given diseases than would be expected based on pathogenicity classifications within these clinical databases [48].

To be able to eliminate poor results, comparing non-human and human equivalent variants might be helpful. But this kind of need should have been fulfilled by a proper tool and database. Pir et al (2021) have developed a tool named ConVarT which can allow users to search and compare their variants between human and non-human species [49]. Thanks to this tool; orthologous variants between humans, mice, and *C. elegans* can be visualized easily and give elucidate the functional results of human and non-human genetic variations.

Chapter 2

2. MATERIAL AND METHODS

As the first step, necessary files and documents which are reported in the tables below, were downloaded.

Table 2.1 Links of pep fasta files

DownloadFile	Download Link	Download Date
Human pep fasta	*http://ftp.Ensembl.org/pub/release-105/fasta/homo_sapiens/pep/Homo_sapiens.GRCh38.pep.all.fa.gz*	12-Dec-2021
Mouse pepfasta	*http://ftp.Ensembl.org/pub/release-105/fasta/mus_musculus/pep/Mus_musculus.GRCm39.pep.all.fa.gz*	12-Dec-2021
Fly pep fasta	*http://ftp.Ensembl.org/pub/release-105/fasta/drosophila_melanogaster/pep/Drosophila_melanogaster. BDGP6.32.pep.all.fa.gz*	12-Dec-2021
Dog pepfasta	*http://ftp.Ensembl.org/pub/release-105/fasta/canis_lupus_familiaris/pep/Canis_lupus_familiaris.ROS _Cfam_1.0.pep.all.fa.gz*	12-Dec-2021
Cow pepfasta	*http://ftp.Ensembl.org/pub/release-105/fasta/bos_taurus/pep/Bos_taurus.ARS-UCD1.2.pep.all.fa.gz*	12-Dec-2021
Worm pepfasta	*http://ftp.Ensembl.org/pub/release-105/fasta/caenorhabditis_elegans/pep/Caenorhabditis_elegans.WB cel235.pep.all.fa.gz*	12-Dec-2021
Zebrafish pep fasta	*http://ftp.Ensembl.org/pub/release-105/fasta/danio_rerio/pep/Danio_rerio.GRCz11.pep.all.fa .gz*	12-Dec-2021
Chimp pepfasta	*http://ftp.Ensembl.org/pub/release-105/fasta/pan_troglodytes/pep/Pan_troglodytes.Pan_tro_3.0.pep.a ll.fa.gz*	12-Dec-2021

After downloading fasta files of each organisms' necessary software was downloaded and set up.

Table 2.2 Links of setup files of necessary software' and orthology list

Ensembl VEP program	*https://github.com/Ensembl/Ensembl-vep/archive/release/105.zip*	20-February-2022
Blast+ program	*https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.11.0/ncbi-blast-2.11.0+-x64-win64.tar.gz*	17-April-2021
orthology list	*https://fms.alliancegenome.org/download/ORTHOLOGY-ALLIANCE_COMBINED.tsv.gz*	11-Dec-2021

Lastly variation and phenotypic variants files were downloaded.

Table 2.3 Links of variant data

Cow Ensembl variants	*http://ftp.Ensembl.org/pub/release-105/variation/vcf/bos_taurus/bos_taurus.vcf.gz*	20-February-2022
Dog Ensembl variants	*http://ftp.Ensembl.org/pub/release-105/variation/vcf/canis_lupus_familiaris/canis_lupus_familiaris.vc f.gz*	20-February-2022
Zebrafish Ensembl variants	*http://ftp.Ensembl.org/pub/release-105/variation/vcf/danio_rerio/danio_rerio.vcf.gz*	20-February-2022
Mouse Ensembl variants	*http://ftp.Ensembl.org/pub/release-105/variation/vcf/mus_musculus/mus_musculus.vcf.gz*	20-February-2022
Chimp Ensembl variants	*http://ftp.Ensembl.org/pub/release-105/variation/vcf/pan_troglodytes/pan_troglodytes.vcf.gz*	20-February-2022
Rat Ensembl variants	*http://ftp.Ensembl.org/pub/release-105/variation/vcf/rattus_norvegicus/rattus_norvegicus.vcf.gz*	20-February-2022
Human Ensembl variants	*http://ftp.Ensembl.org/pub/release-105/variation/vcf/homo_sapiens/homo_sapiens-chr1.vcf.gz (/homo_sapiens-chr2-3-4-5-6-7-8-9-10-11-12-13-14-15-16-17-18-19-20-21-22.vcf*	20-February-2022
Human cosmic variants	*https://cancer.sanger.ac.uk/cosmic/vcf/CosmicCodingMuts.vcf.gz*	13-February-2022

Human gnomAD variants	*https://storage.googleapis.com/gcp-public-data--gnomAD/release/2.1.1/liftover_grch38/vcf/exomes/gnomAD.exomes. r2.1.1.sites.liftover_grch38.vcf.bgz*	7-January-2022
Human ClinVar variants	*https://ftp.ncbi.nlm.nih.gov/pub/ClinVar/tab_delimited/hgvs4variati on.txt.gz*	7-January-2022

Table 2.4 Links of phenotypic variants data

C elegans phenotypic variants	*https://fms.alliancegenome.org/download/VARIANT-ALLELE_NCBITaxon6239.tsv.gz*	30-March-2022
Mouse phenotypic variants	*https://fms.alliancegenome.org/download/VARIANT-ALLELE_NCBITaxon10090.tsv.gz*	30-March-2022
Zebrafish phenotypic variants	*https://fms.alliancegenome.org/download/VARIANT-ALLELE_NCBITaxon7955.tsv.gz*	30-March-2022
Rat phenotypic variants	*https://fms.alliancegenome.org/download/VARIANT-ALLELE_NCBITaxon10116.tsv.gz*	30-March-2022
Fly phenotypic variants	*https://fms.alliancegenome.org/download/VARIANT-ALLELE_NCBITaxon7227.tsv.gz*	30-March-2022
Dog phenotypic variants	*https://www.omia.org/results/?gb_species_id=9615&search_type=advanced&model=yes*	30-March-2022
Cow phenotypic variants	*https://www.omia.org/results/?gb_species_id=9913&search_type=advanced&model=yes*	30-March-2022
Chimp phenotypic variants	*https://www.omia.org/results/?exclude_gb_species_id=9615%2C9913%2C9685%2C9823%2C9940%2C9796%2C9031%2C9986%2C9925%2C93934%2C10036&search_type=advanced&result_type=variant*	30-March-2022

2.1 The gene homology list for humans and non-humanspecies

Homology lists for Danio rerio, Drosophila melanogaster, Caenorhabditis elegans, Mus musculus, and Rattus norvegicus were downloaded from the alliance genome website, and necessary data manipulations were conducted on the R programming language. To be able to form other organism's homology list, an orthology table was made by using the blast+ program of NCBI.

```
blastp -query human.faa      #amino acid seq of first organism
       -db chimp.faa         #amino acid seq of second organism
       -out homo_chimp.csv   #name of output file
       -outfmt 6             #output format
       -evalue 0.00001       #max e-value
       -max_target_seqs 1    #max target sequence
       -num_threads 15       #selection of used tread to decide speed
```

Figure 2.1.1 The gene homology list for humans and non-humanspecies

2.1.1 Download of genetic variations for 7 organisms and their amino acid conversion on Ensembl-VEP

GnomAD, cosmic, TOPMed, and ClinVar databases were used for human variation data. For other organisms, the Ensembl website and the alliance genome website were used, and all documents were downloaded from the links in Table 3.

Since downloaded files have nucleotide change data, we needed to find their amino acid change equivalent. To be able to find amino acid changes in protein variants, Ensembl-VEP software was downloaded and installed on our Linux computer. After the installation process is finished, target files and our purposes were indicated in our codes on the terminal part of the Linux system;

```
./vep --cache --force -i chimp_variants.txt
                      -o chimp_variants_output.txt
                      -species Pan troglodytes
                      -fork 15

```

Figure 2.1.1.1 Download of genetic variations for 7 organisms and their aminoacid conversion on Ensembl-VEP

Since VEP have its own command codes starts with "./vep" function. "- - cache" flag enables use of cache on local disk. "- - force" function/flag force the overwrite of the existing file. While "- i" indicates input "-o" indicates output file. After "- species" flag, binomial name of organism was written and to be able to arrange the use of processor and speed; thread the number was written after "-fork" flag.

2.1.2 Selecting only amino acid changes from Ensembl-VEP results

Since our code on Ensembl-VEP gives us all variation, we needed to eliminate and remove all unwanted data to get only amino acid changes by using the following code on the Linux terminal.

```
./filter_vep -i chimp_variants_output.txt
             -o chimp_filtered.txt
             -filter "Amino acids"
```

Figure 2.1.2.1 Selecting only amino acid changes from Ensembl-VEP results

./filter_vep function provides a filtration for VEP results. The result file was written after "-i" flag as input and the name of output decided by writing its name following to "-o" flag. Among the all variation results, only amino acid change containing variants selected by using "-filter" flag of ./filter_vep function.

2.2 Data manipulation

Filtered results should be arranged as a table that contains "Ensembl protein I.D"., "the amino acid position of variations", "From" and "To" columns, which indicates ID, changed amino acid position, the wild type and changed version of the amino acid, respectively.

As the first step; the filtered VEP result was transformed into the data table on R studio, and all necessary data was placed in a specific column for each organism. Examples of codes can be seen below for chimp.

```
#Necessary libraries should be installed
library(stringi)
library(tidyr)
library(dplyr)
library(stringr)
library(data.table)
chimp <- read.table("C:/Users/Msi/Desktop/deneme 1/chimp/chimpflt.txt",
                                                header=TRUE, quote="\"")

df1 = chimp [,14]                   #separating last column
ensp1= "(ENSPTRP[0-9]{11})"         #write pattern of Ensembl ID
ensp = str_extract(df1, ensp1)      #extracting Ensembl ID
ENSP = data.frame(ensp)             #data frame list converted into data frame

#remove unnecessary columns and add sift and polyphen columns
chimp = chimp %>% select (1, 4 , 5 , 7 , 8 , 11)
chimp$ENSP = ENSP                   #merging Ensembl ID and necessary columns
colnames(chimp)<-                   #renaming columns
c("rs_id","ensm_gene_id","ensm_transcript_id","variant_type",
                                "position","variation","ensp")
```

Figure 2.2.1 Filtered VEP result

After having a data table which contain necessary columns, amino acids changes split into from and to columns separately;

```
chimp<-as.data.table(chimp)                  #force to convert data table format
chimp<-chimp[nchar(chimp$variation)<4,]       #eliminate rows if character legth of
                                             variation column smaller than 4
chimp$position<-as.numeric(chimp$position)    #define all position column as numeric
chimp<-chimp[!is.na(chimp$position),]         #remove NAs
chimp<-chimp[stri_sub(chimp$variation, 1, 1) != "*",]   #remove rows that contain "*"
chimp<-unique(chimp)                         #remove repationg rows
chimp<-chimp[!duplicated(chimp[,c(1,3,6)])]   #similar procedure for only specific rows
chimp$from<-str_split(chimp$variation, "/",   #write the expression before "/" character on
                      simplify = TRUE)[,1]    variation column to "from" column
chimp$to<-str_split(chimp$variation, "/",     #write the expression after "/" character on
                      simplify = TRUE)[,2]    variation column to "to" column
chimp$to<-ifelse(chimp$to == "", chimp$from,  #make "from" and "to" column same
                                 chimp$to)    if "to" column empty
```

Figure 2.2.2 Separate the columns which contain necessary in table

At the end we need to have a data table which contain "rs id","Ensembl gene id","Ensembl transcript id", "Variant type"," Amino acid position"," Ensembl protein id ","From" and "To" columns. We obtained a data table of 22066-row length for the chimp example (Figure 2.2.3).

	rs_id	ensm_gene_id	ensm_transcript_id	variant_type	position	ensp	from	to
1	rs26734344	ENSPTRG00000000017	ENSPTRT00000061949	synonymous_variant	1869	ENSPTRP00000054495	A	A
2	rs26614143	ENSPTRG00000000047	ENSPTRT00000000100	synonymous_variant	600	ENSPTRP00000000080	A	A
3	rs26136121	ENSPTRG00000000049	ENSPTRT00000108734	missense_variant	325	ENSPTRP00000076645	R	W
4	rs24904357	ENSPTRG00000000057	ENSPTRT00000000123	missense_variant	376	ENSPTRP00000000103	V	I
5	rs24919072	ENSPTRG00000000057	ENSPTRT00000000123	synonymous_variant	393	ENSPTRP00000000103	F	F
6	rs26745298	ENSPTRG00000044415	ENSPTRT00000084573	synonymous_variant	78	ENSPTRP00000086205	P	P
7	rs26718831	ENSPTRG00000000059	ENSPTRT00000080699	missense_variant	3608	ENSPTRP00000077930	A	E
8	rs26732289	ENSPTRG00000045016	ENSPTRT00000080207	synonymous_variant	2370	ENSPTRP00000068580	Y	Y
9	rs26732289	ENSPTRG00000045016	ENSPTRT00000085195	synonymous_variant	2571	ENSPTRP00000070125	Y	Y
10	rs26723791	ENSPTRG00000052696	ENSPTRT00000083194	synonymous_variant	63	ENSPTRP00000082150	P	P
11	rs26668603	ENSPTRG00000000064	ENSPTRT00000000145	synonymous_variant	1400	ENSPTRP00000000121	S	S
12	rs26167957	ENSPTRG00000000064	ENSPTRT00000000145	missense_variant	864	ENSPTRP00000000121	H	N
13	rs26630217	ENSPTRG00000000068	ENSPTRT00000000149	synonymous_variant	135	ENSPTRP00000052915	L	L
14	rs26747714	ENSPTRG00000024207	ENSPTRT00000042678	missense_variant	459	ENSPTRP00000044076	G	S

Figure 2.2.3 First rows of tidied amino acid variation table for chimp

The same procedure was done for each organism to have similar tables. ClinVar, COSMIC, TOPMed, and gnomAD variation data was also manipulated in the same way to get human amino acid variation table.

2.3 Multiple sequence alignments

MSA of each organism was done by using OrthoMSA function of OrthoVar package which are available on R programming language.

```
44   library(orthoVar)
45   msa <- orthoMSA(species1 = "Homo sapiens", “Pan troglodytes”,
46                                              humanSeqFile = NA,
47                                              seqFiles = NA,
48                                              customOrt = NA,
49                                              annot = "Ensembl")
```

Figure 2.3.1 Multiple sequence alignments

Since downloading and file selection is done automatically no need to select humanSeqFile, Seqfiles, customOrt only writing species name and annot part is enough to obtain MSA result from OrthoVar package.

2.4 Matching variants

Matching variants of each organism by human were found by using OrthoFind function of OrthoVar package which are available on R programming language.

```
50   Clinvar_chimp_MatchVar = orthoFind (Clinvar,   #data table which contain clinvar variation
51                                       chimp,        #data table which contain chimp variation
52                                       Homo sapiens, #organism1 binomial name
53                                       Pan troglodytes, #organism2 binomial name
54                                       msa,            #obtained msa data table
55                                       ort = TRUE)
```

Figure 2.4.1 Matching variants

Chapter 3

3.RESULTS

In this study, genetic variations of 8 different organisms were matched to human genetic variations. To be able to match variants, multiple sequence alignment was done between Human-zebrafish, Human-rat, Human-chimp, Human-cow, Human-mouse, Human-C elegans, and Human-dog separately on R studio. As is expected, more similar organisms are more likely to have a higher number of aligned protein sequences (Table 3.1).

Table 3.1 Numbers of MSA, Variant numbers and matching variant numbers

	Number of MSA	Number of Total Variants	Number of Matching Variants
Zebrafish	232647	418200	11215
Fly	190576	18771	574
Rat	270727	81707	2285
Chimp	320298	22066	1253
Cow	212785	6304776	197226
Mouse	478986	1157418	40979
C elegans	140688	382037	86879
Dog	264702	207217	7684

Different databases were used to collect variation data for each organism (Table 2.1). For example; variant data of zebrafish were taken from the Ensembl and alliance genome website and merged on R studio. Since many of the variant data were given as nucleotide variation, the amino acid equivalent of each variation was calculated by Ensembl VEP software, and at the end, the total variant number of each organism was obtained (Table 3.1). The same procedure was done to get human variation data, and 31988340 gnomAD variations, 19054175 COSMIC variations, 2755294 ClinVar variations, and 30439686 TOPMed variations were obtained.

Human equivalents of variations (Matching variants) were found by comparing each organism's variation with human variations, which are taken from gnomAD, TOPMed, COSMIC, and ClinVar (Table 3.2).

Table 3.2 Numbers of Matching Variants of different databases

	GnomAD	COSMIC	TOPMed	ClinVar	Total Matching Variants
Fly	140	198	141	95	574
Zebrafish	1973	5226	1866	2150	11215
Rat	617	945	392	331	2285
Chimp	480	449	245	79	1253
Cow	61268	44965	72877	18116	197226
Mouse	13589	17220	10170	38270	79979
C elegans	6285	74722	5837	35	86879
Dog	2915	2602	1592	575	7684

3.1 Zebrafish

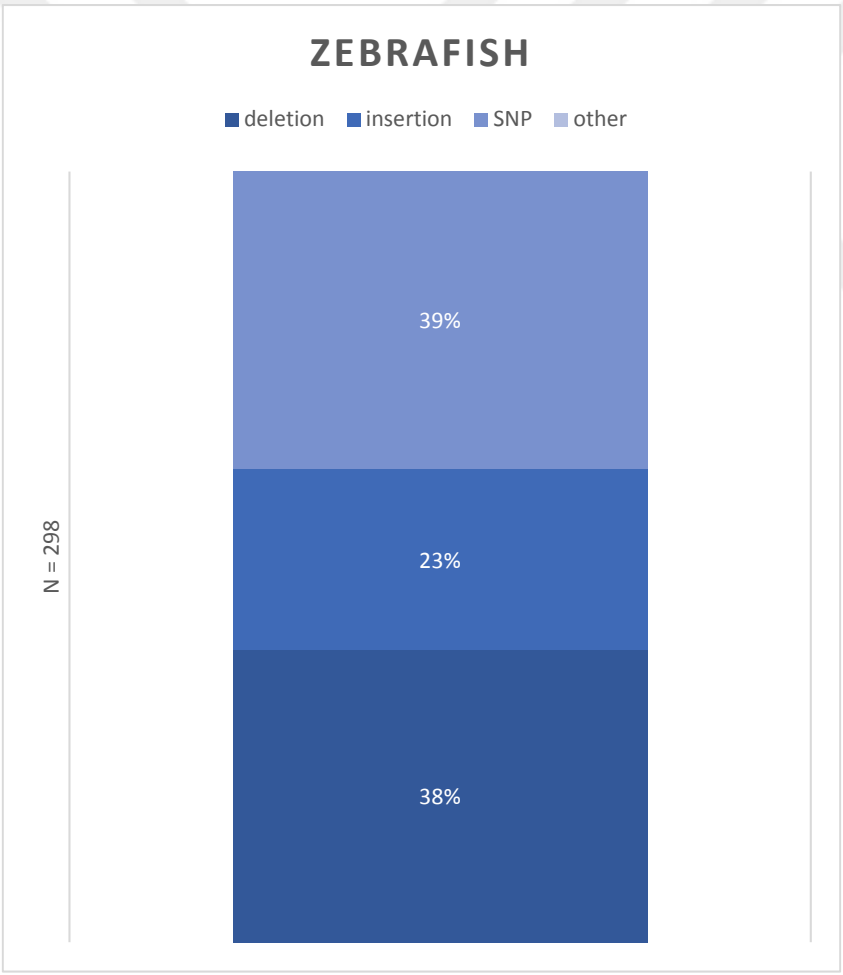

Figure 3.1.1 Types of Pathogenic Zebrafish Variants

Although there were 418200 zebrafish variants in our data, the number of reported pathogenic variants was only 298 in Alliance genome database. About one-third of the phenotypic zebrafish variants are point or SNP, another one-third is deletion, and the remaining variants are insertions mutation, delins, etc. (Figure 3.1.1.1)

57 of 298 pathogenic variants were matched with human variations.

3.2 Drosopila

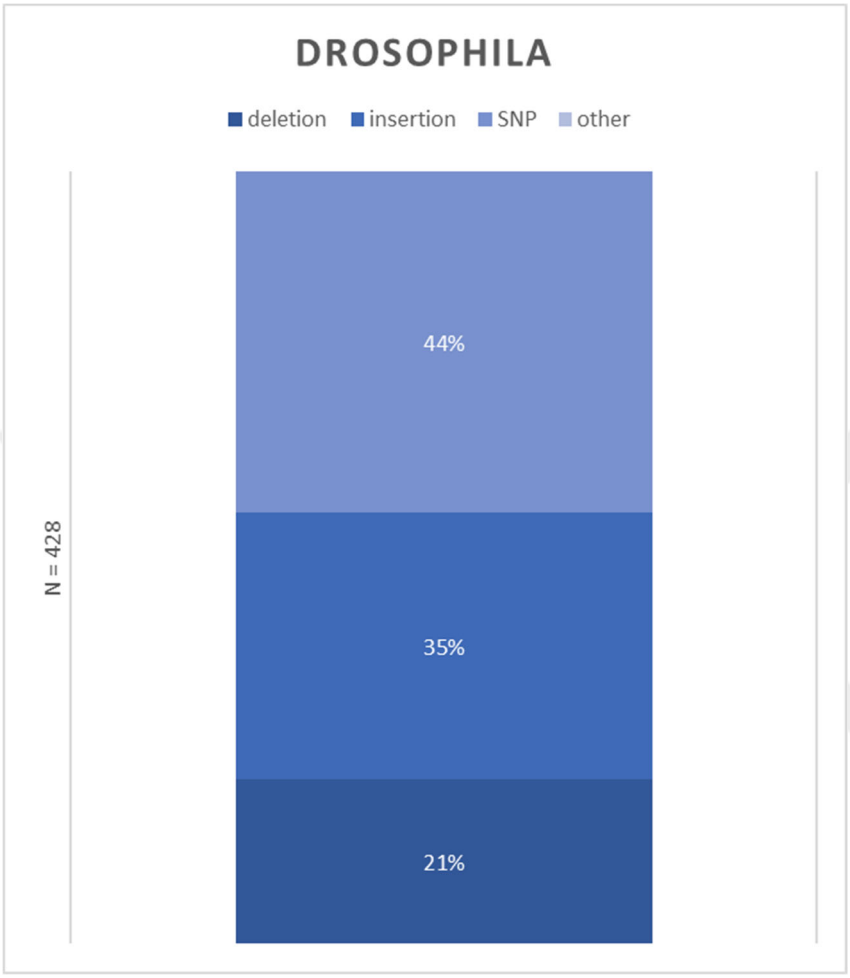

Figure 3.2.1 Types of Pathogenic Fly Variants

Among 18771 *Drosophila melanogaster*(fly) variants in our data, the number of reported pathogenic variants was only 428 in Alliance genome database. About 45 percent of the pathogenic drosophila variants are point mutation or SNP, 35 percent are insertion, and the remaining variants are deletion delins, etc. (Figure 3.1.2.1) Unluckily we couldn't find any matching pathogenic variants with human variations.

3.3 Rat

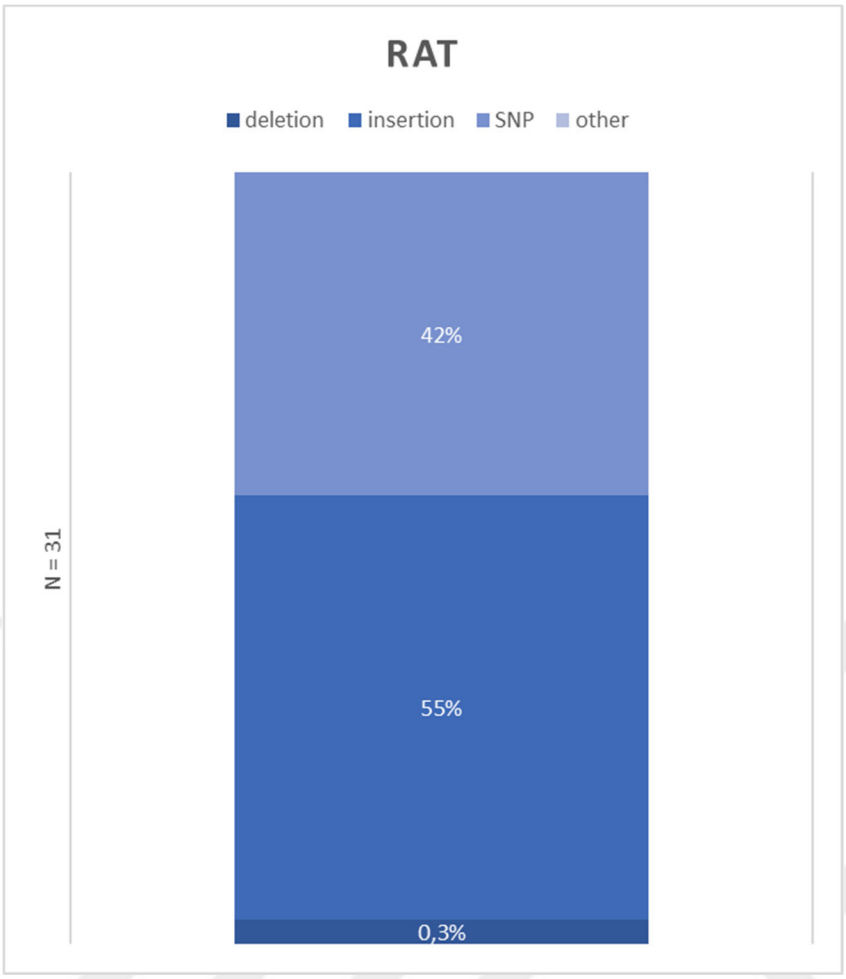

Figure 3.3.1 Types of Pathogenic Rat Variants

Among 81707 Rat variants in our data, the number of reported pathogenic variants was only 31 in Alliance genome database. 13 of the phenotypic drosophila variants is point mutation or SNP, 17 of them is insertion and there was only 1 reported deletion (Figure 3.1.3.1).

There were 2285 matching variants between rat and human. 5 of the matching rat variants have pathogenic features. 3 belongs to different regions of the lgi1 gene, and the other is the atm and plp1 genes, respectively (Figure 3.1.3.2).

3.4 Dog

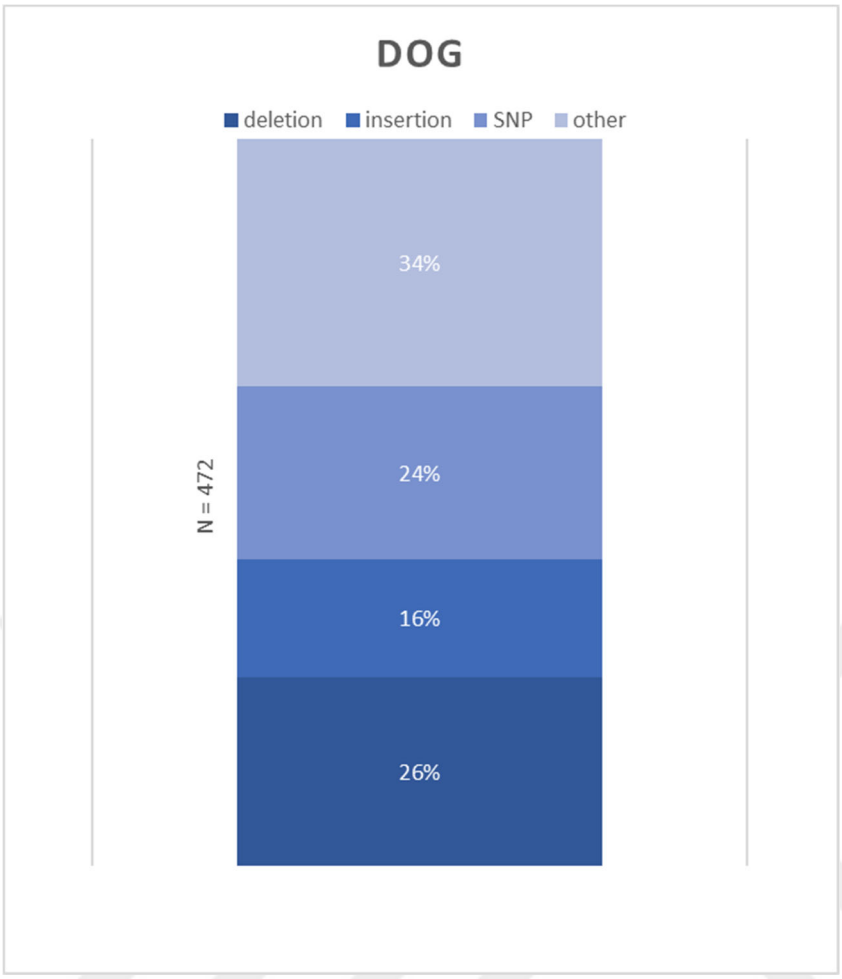

Figure 3.4.1 Types of Pathogenic Dog Variants

207217 is the total number of dog variations. Among them, 472 variants were reported as phenotypic in the OMIA database. 112 of the phenotypic dog variants are point mutation or SNP, 76 of them are insertions, and there were 123 deletions. The remaining 161 variants is another type of variants (Figure 3.1.4.1),

There were 7684 matching variants between dog and human. 66 of the matching dog variants have pathogenic features.

3.5 Cow

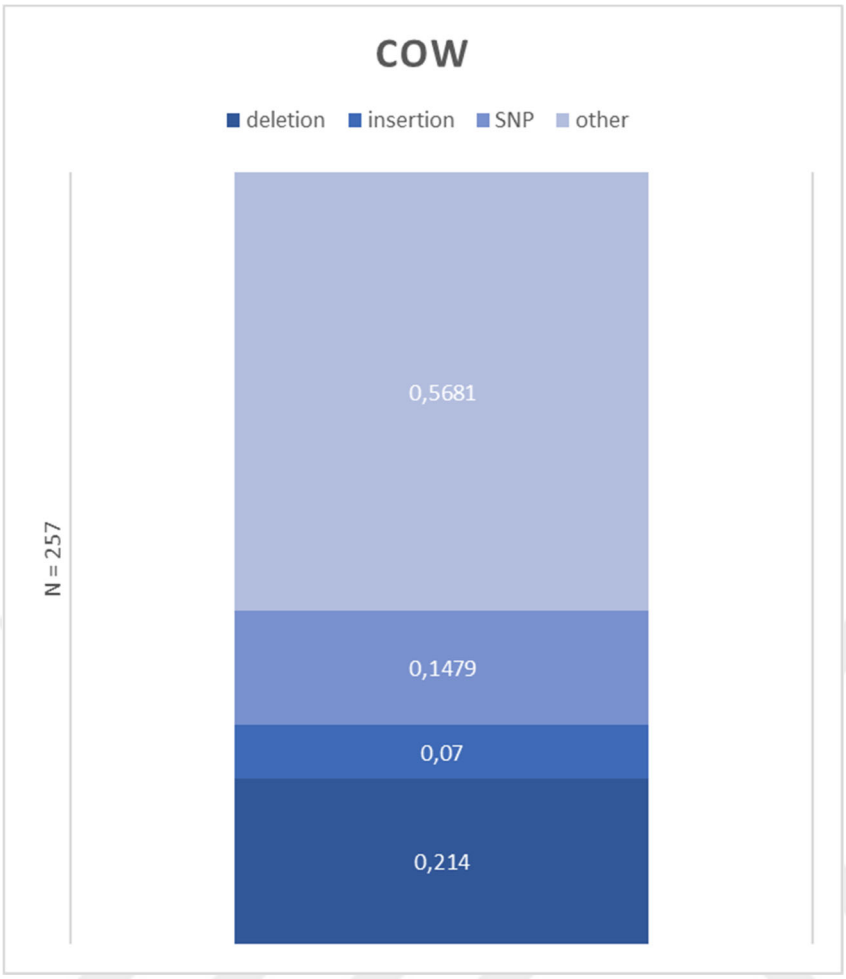

Figure 3.5.1 Types of Pathogenic Cow Variants

Cow has a vast number of variations. Since they have a dramatic commercial value, they have been studied and bred to have higher meat or milk production. As a result, high number of variations has been reported over the years. After amino acid conversion from Ensembl VEP software, we had 6304776 amino acid variations for cow. Most of them have no pathogenic indication in the literature. Only 257 variants were reported as pathogenic in the OMIA database. 38 of the phenotypic cow variants are point mutation or SNP, 18 of them are insertions, and there were 55 deletions. The remaining 146 variants is another type of variant (Figure 3.1.5.1)

There were 197226 matching variants between cow and human. 16 of the matching cow variants have pathogenic features.

3.6 Chimp

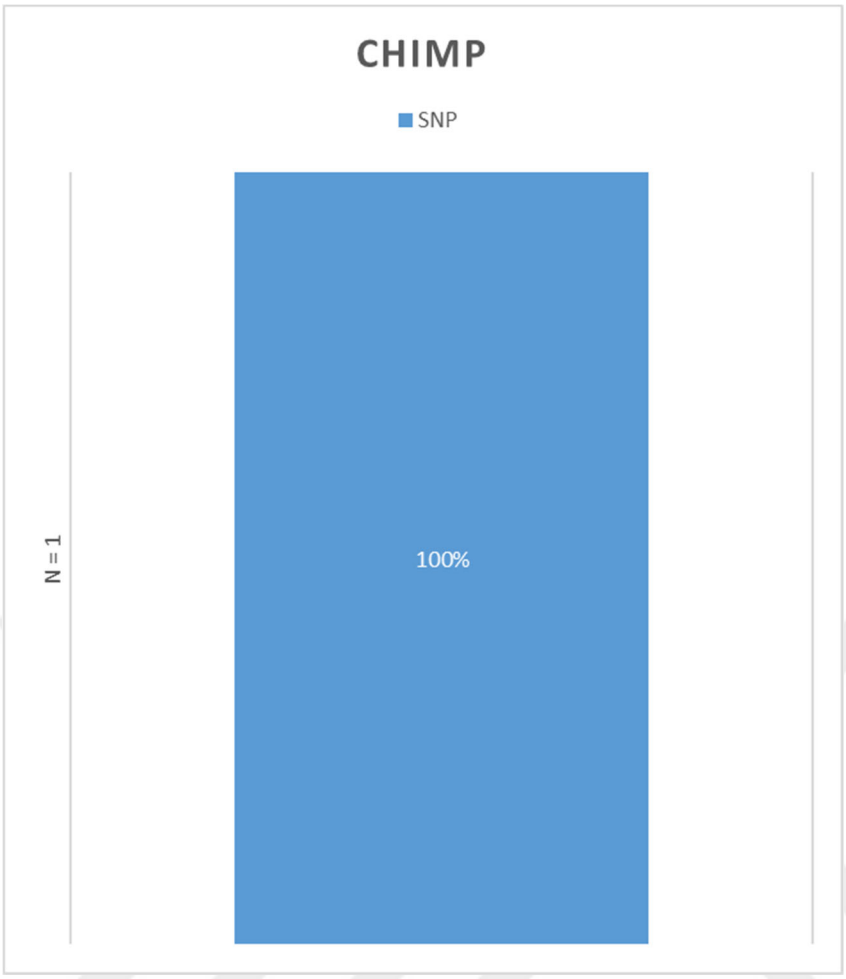

Figure 3.6.1 Types of Pathogenic Chimp Variants

Interestingly there were only one reported pathogenic variation for chimp on OMIA database, and it was SNP.

3.7 C elegans

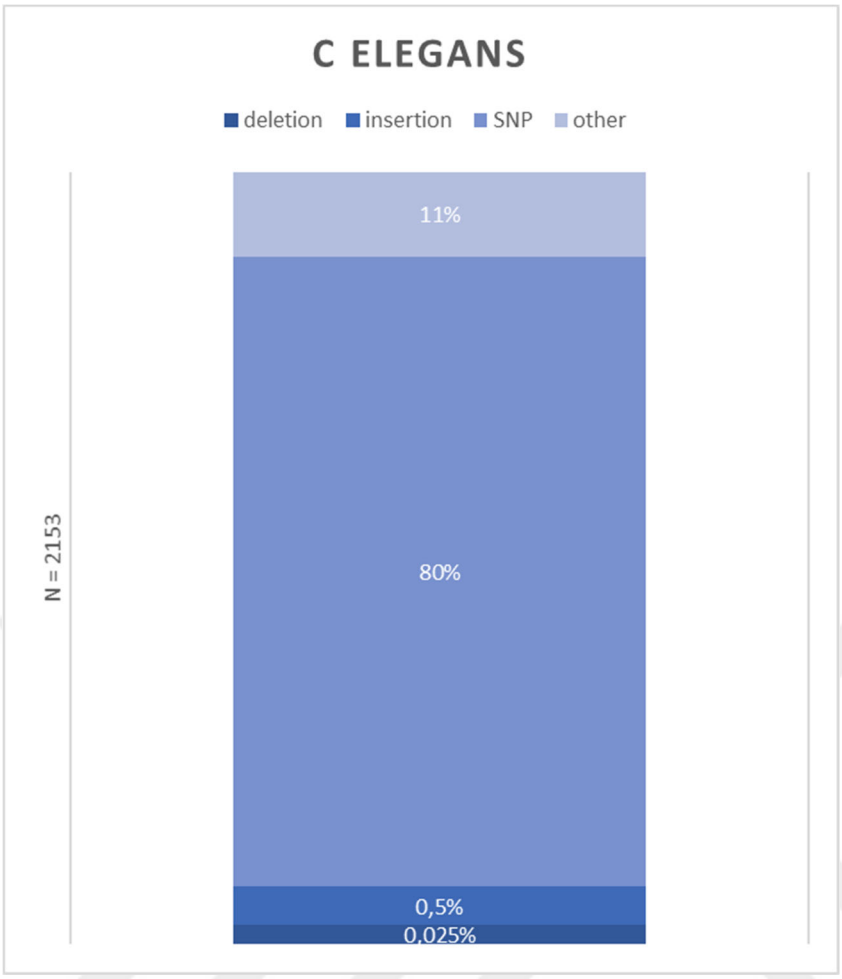

Figure 3.7.1 Types of Pathogenic *C elegans* Variants

382037 is the total number of the C elegans variation. Among them, 2153 variants were reported as phenotypic in Alliance genome database. 80 percent of the phenotypic C elegans variants are point mutation or SNP, about 5 percent of them are insertions, and the remaining variants belong to another type of variants (Figure 3.1.7.1)

Human and C elegans have over 86000 matching variants, but among them, only 987 variants have a pathogenic effect on C elegans.

3.8 Mouse

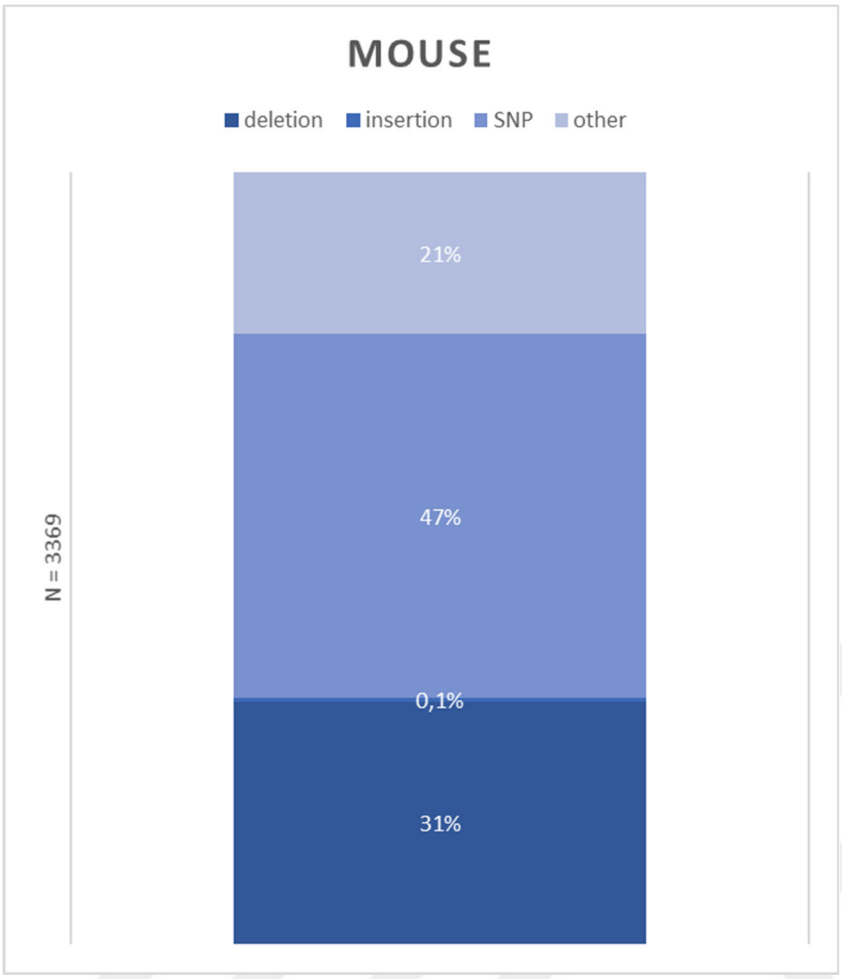

Figure 3.8.1 Types of Pathogenic Mouse Variants

382037 is the total number of mouse variations. Among them, 3369 variants were reported as phenotypic in Alliance genome database. 47 percent of the phenotypic mouse variants are point mutation or SNP, and about 31 percent of them are deletion (Figure 3.1.8.1)

Humans and mouse have over 40979 matching variants, and 4957 of them have a pathogenic effect.

3.9 Selected Gene

The ADAM-TS2 (a disintegrin and metalloproteinase with thrombospondin motifs 2) gene is involved in processing various procollagen proteins. Procollagen is a precursor to collagen, a protein that provides strength and support to many body tissues [50]. Many mutations in the ADAMTS2 gene have been reported to be pathogenic and result in significantly reduced production of enzymes made by the ADAMTS2 gene [50]. Without this enzyme, procollagen cannot be adequately processed. As a result, collagen fibrils are not assembled correctly and diseases occur. Since cows have the same gene and matching

variants with humans, the ADAMTS2 gene was selected to investigate. Luckily the amino acid sequences are very conserved for both organisms and have a high similarity rate with human and cow ADAMTS2 genes.

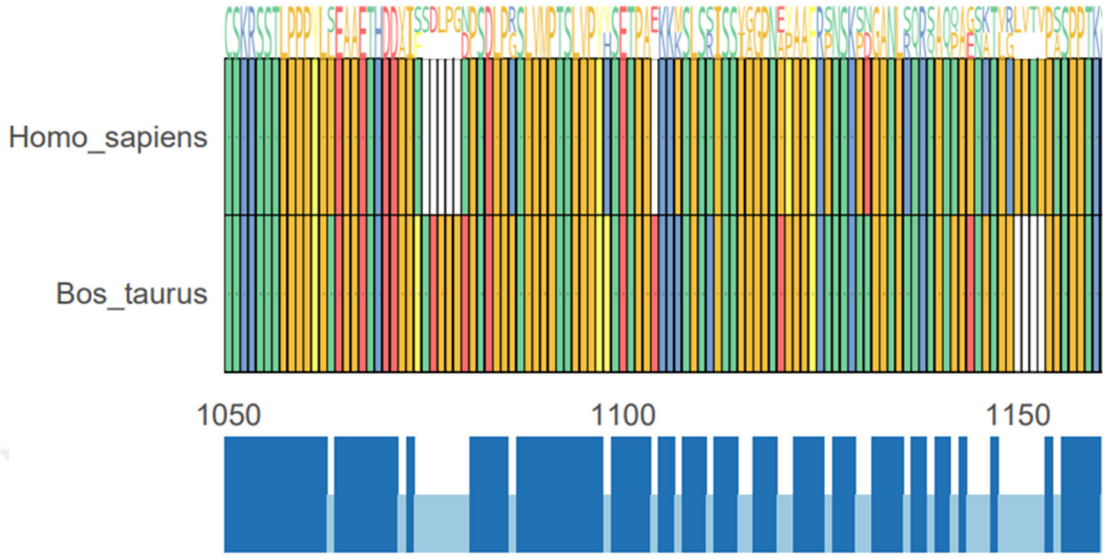

Figure 3.9.1 MSA of Human ADAMTS2 and Cow ADAMTS2 proteins

Seven variants in Bos taurus ADAMTS2 gene (Also ADAMTS2 in Human) which encode a disintegrin and metalloproteinase with thrombospondin motifs protein, were selected to investigate (ENSBTAP00000053777: p.T283A, p.A353T, p.T357I, p.H408T, p.Q444L, p.D746E, and p.H802Q). Functional outcomes of the variations represented in Figure 3.9.2.

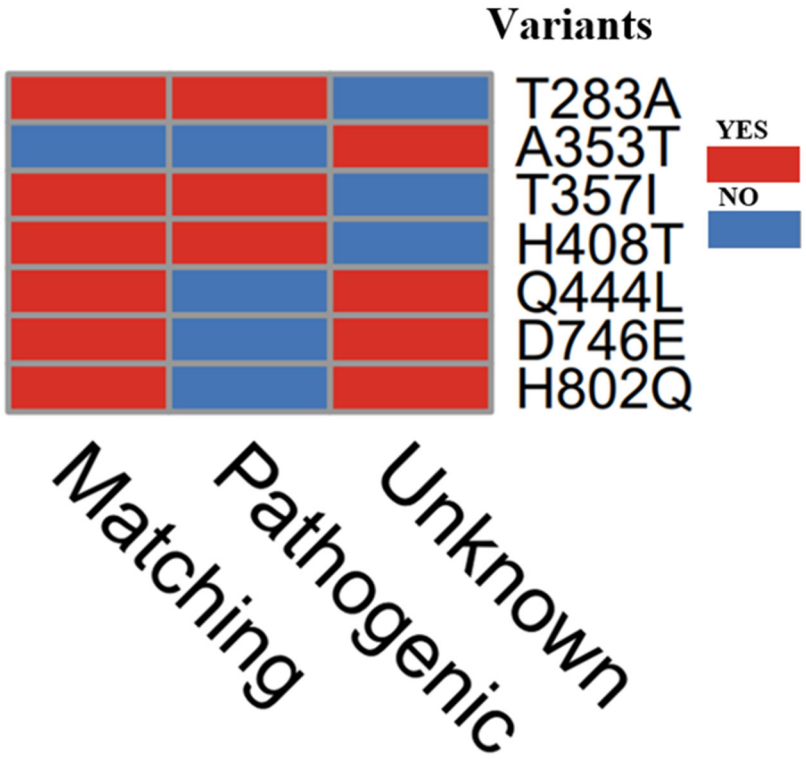

Figure 3.9.2 Functional outcomes of matching Cow ADAMTS2 gene variants with

human

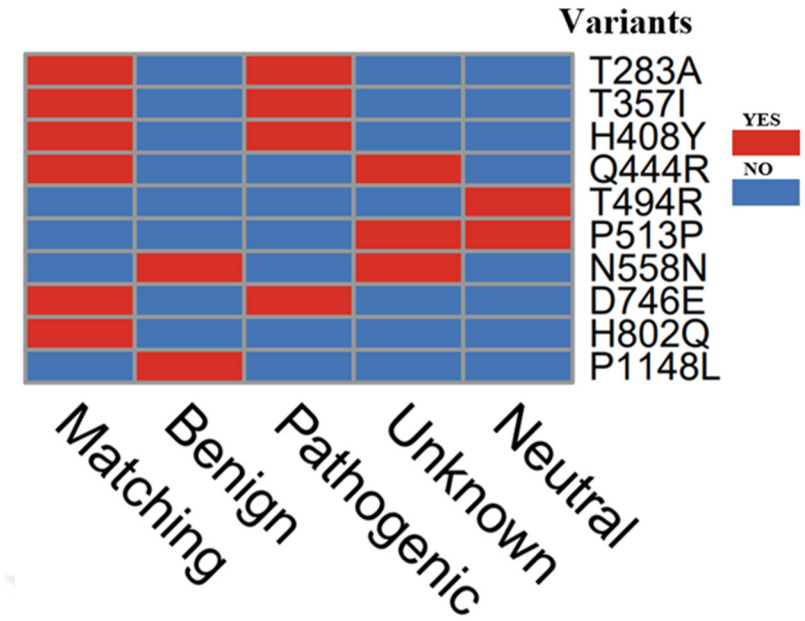

Figure 3.9.3 Functional outcomes of matching human ADAMTS2 gene variants with cow

Ten variants in *Homo sapiens* ADAMTS2 gene were selected to investigate (ENSP00000480055: p.T283A, p.T353I, p.H408T, p.Q444L, p.T494R, p.P513P, p.N558N, p.D746E, p.H802Q and p.P1148L). Functional outcomes of the variations represented in Figure 3.9.3.

Chapter 4

4.Discussion

The number of known genetic diseases and pathogenic variants are increasing with new studies and developments in sequencing and analyzing techniques. Collecting variant data from different databases, matching them with humans, and analyzing them as they have phenotypic annotations or not would be a beneficial approach to understanding unknown gene and variant properties. In this study, we aimed to analyze matching variants and predict possible pathogenic variants by in silico analysis methods.

For the zebrafish example, it is seen that the phenotypic ratio of all zebrafish variants is below %1. Rest of the higher percentage of variants are not phenotypic because zebrafish is a model organism, and non-phenotypic variants are also frequently reported. Similar rates were seen on all other organisms except Drosophila melanogaster.

Among the phenotypic variants, pathogenic variants are so low because many of them are unknown. For example, zebrafish have a %99.5 unknown phenotypic variant ratio. That's why it is important to predict their properties. After matching the similar variants were done, specific genes are selected from cow as a proof of concept. As an example, ADAMTS2 gene of cow were selected to investigate. Seven variants in *Bos taurus* ADAMTS2 gene were selected (ENSBTAP00000053777: p.T283A, p.A353T, p.T357I, p.H408T, p.Q444L, p.D746E, and p.H802Q). six of the cow ADMATS2 gene variants have a matching human variant version and all of them were conserved. Three of the matching cow variants (p.T283A, p.T357I and p.H408T) reported as pathogenic in the literature [51,52]. Consistently human equivalent variants (p.T283A, p.T357I, pT357W and p.H408Y) have predicted as deleterious by SIFT, probably damaging by Poly-Phen, likely disease causing by REVEL prediction tools. This data indicates that the experimental data of the cow, which is already available [51], may provide additional evidence to elucidate the functional outcomes of human variants.

Chapter 5

5.Conclusions and Future Prospects

5.1 Conclusions

In this study, we aimed to discover new pathogenic variants. After investigating and comparing matching variants of eight different organisms, we found many matching phenotypic variants and related diseases. However, in many cases, human variation does not have any phenotypic annotation in the literature, but it's matching non-human ortholog or vice versa. Thanks to our method, if one of the matching variants has a phenotypic annotation in the literature, the functional outcome of the other matching variant can be predicted and found easily.

5.2 Social Impact and Contribution

Inherited diseases are health problems caused by one or more abnormalities in the genome. It can be caused by changes in a single gene (monogenic) or multiple genes (polygenic), or by a damage on chromosomes [53]. There are many types of a genetic disorders. Their numbers are so high that almost 5% of people are affected by a genetic disease [54]. Those diseases reduce human life quality, and all the patients have an important place in the overall population. But unfortunately, there is still no known cure for many genetic diseases, and they are classified as rare diseases. Understanding genetic variants would illuminate how to diagnose and solve these diseases better.

Genetic variation is the differences in the DNA sequences that can be observed within a species or in alleles. Evaluation of genetic variants, together with reported phenotypic or pathogenic annotations from non-human organisms, facilitates the comparison of these variants with their human counterparts. In this work, we combined pathogenic and phenotypic annotations with variants, and these phenotypic orthologous variants from seven organisms can provide clues to the functional consequences of human genetic variants. Scientists may compare them to understand and predict diseases in human. Also,

veterinaries can predict the type of pathogeny of the animals by comparing matching human variants.

5.3 Future Prospect

Genetic diseases are pretty crucial for all organisms. Thanks to matching variants analysis, possible pathogenic variants of different organisms can be found before their wet lab verification; thus, the diagnosis or treatment process can be started. Also, scientists can use our method in different wet lab or dry lab studies.

Thanks to prepared pipelines and codes, the same systems and methods can easily be applied to new organisms and new variations from different databases. Thanks to our study and methods, which contains eight organisms, can be expanded and updated for new studies and new organisms.

BIBLIOGRAPHY

[1] Avery O.T., MacLeod C.M., McCarty M., "Studies on the chemical nature of the substance inducing transformation of pneumococcal types." Journal of Experimental Medicine, 79(2), 137–158 (1944).

[2] Watson J.D., Crick F.H., "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid." Nature, 171, 737–738 (1953).

[3] Sanger F., Nicklen S., Coulson A.R., "DNA sequencing with chain-terminating inhibitors." Proceedings of the National Academy of Sciences of the United States of America, 74(12), 5463–5467. (1977).

[4] Nurk S., Koren S., Rhie A., Rautiainen M., Bzikadze A.V., Mikheenko A., Vollger M.R., Altemose N., Uralsky L., Gershman A., Aganezov S., Hoyt S.J., Diekhans M., Logsdon G.A., Alonge M., Antonarakis S.E., Borchers M., Bouffard G.G., Brooks S.Y., Caldas G.V., Chen N.C., Cheng H., Chin C.S., Chow W., de Lima L.G., Dishuck P.C., Durbin R., Dvorkina T., Fiddes I.T., Formenti G., Fulton R.S., Fungtammasan A., Garrison E., Grady P.G.S., Graves-Lindsay T.A., Hall I.M., Hansen N.F., Hartley G.A., Haukness M., Howe K., Hunkapiller M.W., Jain C., Jain M., Jarvis E.D., Kerpedjiev P., Kirsche M., Kolmogorov M., Korlach J., Kremitzki M., Li H., Maduro V.V., Marschall T., McCartney A.M., McDaniel J., Miller D.E., Mullikin J.C., Myers E.W., Olson N.D., Paten B., Peluso P., Pevzner P.A., Porubsky D, Potapova T, Rogaev EI, Rosenfeld J.A., Salzberg SL, Schneider V.A., Sedlazeck F.J., Shafin K., Shew C.J.., Shumate A, Sims Y., Smit A.F.A., Soto DC, Sović I, Storer J.M., Streets A., Sullivan B.A., Thibaud-Nissen F., Torrance J., Wagner J, Walenz B.P., Wenger A, Wood JMD, Xiao C, Yan S.M., Young A.C., Zarate S., Surti U., McCoy R.C, Dennis M.Y., Alexandrov IA, Gerton J.L., O'Neill RJ, Timp W, Zook J.M., Schatz M.C., Eichler E.E.., Miga K.H., Phillippy A.M., "The complete sequence of a human genome." Science,376 (6588), 44-53 (2022).

[5] Zhong Y., Xu F., Wu J., Schubert J., Li M.M., "Application of Next Generation Sequencing in Laboratory Medicine." Ann Lab Med. , 41(1), 25-43 (2021).

[6] Bush,W.S., Moore J.H., "Genome-Wide Association Studies." PLOS Computational Biology, 8(12), 1-11. (2012.).

[7] Fidanoğlu P., Belder N., Erdoğan B., İlk Ö., Rajabli F., Özdağ H. "Genom Projeleri 5N1H: Ne, Nerede, Ne Zaman, Nasıl, Neden ve Hangi Popülasyonda?. Türk Hijyen Deneysel Biyoloji Dergisi," 71(1), 45-60. (2014).

[8] Weng L., Macciardi F., Subramanian A., Guffanti G., Potkin S.G., Yu Z., Xie X. "SNP-Based Pathway Enrichment Analysis for Genome-Wide Association Studies." BMC Bioinformatics, 12(99), 1-9.(2011).

[9] Stranger B.E., Forrest M.S., Dunning M., Ingle C.E., Beazley C., Thorne N., Redon R., Bird C.P., Grassi A., Lee C., Tyler-Smith C., Carter N., Scherer S.W., Tavaré S., Deloukas P., Hurles M.E., Dermitzakis E.T. "Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes." Science, 315, 848– 853.(2007).

[10] Wang K., Li M., Hadley D., Liu R., Glessner J., Grant S.F.A., Hakonarson H., Bucan M. "CNV: An Integrated Hidden Markov Model Designed for High-Resolution Copy Number Variation Detection in Whole-Genome SNP Genotyping Data." Genome Research, 17(11), 1665-1674. (2007).

[11] Sehn J.K., "Insertions and Deletions (Indels) in Clinical Genomics" .Elsevier Inc. ,1 ,129- 150 (2015).

[12] Thorisson G.A., Smith A.V., Krishnan L., Stein L.D., "The International HapMap Project" Genome Research, 15, 1592-1593.(2005).

[13] Rotimi C., Leppert M., Matsuda I., Zeng C., Zhang H., Adebamowo C., Ajayi I., Aniagwu T., Dixon M., Fukushima Y., Macer D., Marshall P., Nkwodimmah, C., Peiffer, A., Royal C., Suda E., Zhao H., Wang V.O., McEwen J., "The International HapMap Consortium. Community Engagement and Informed Consent in The International HapMap Project." Community Genetics, 10, 186–198.(2007).
[14] Ribas G., Gonza´lez-Neira A., Salas A., Milne R.L., Vega A., Carracedo B., Gonza´lez E., Barroso E., Ferna´ndez L.P., Yankilevich P., Robledo M., Carracedo A., Benı´tez J. Evaluating HapMap SNP Data Transferability in A Large-Scale Genotyping Project Involving 175 Cancer-Associated Genes.Human Genetics, 118, 669– 679. (2006).
[15] Harrow J., Frankish A., Gonzalez J.M., Tapanari E., Diekhans M., Kokocinski F., Aken B.L., Barrell D.,Zadissa A., Searle S., Barnes I., Bignell A., Boychenko V., Hunt T., Kay M., Mukherjee G., Rajan J.,Despacio-Reyes G., Saunders G., Steward C., Harte R. ,Lin M., Howald C., Tanzer A., Derrien T., Chrast J., Walters N., Balasubramanian S., Pei B., Tress M., Rodriguez J.M., Ezkurdia I., Baren J.V., Brent M. ,Haussler D., Kellis M., Valencia A., Reymond A., Gerstein M., Guigo´ R., Hubbard T.J. "GENCODE:The Reference Human Genome Annotation for The ENCODE Project." Genome Research, 22, 1760-1774.(2012).
[16] Qu H., Fang X. "A Brief Review on The Human Encyclopedia of DNA Elements (ENCODE) Project." Genomics Proteomics Bioinformatics, 11,135–141.(2013).
[17] Washington N.L., Stinson E.O., Perry M.D., Ruzanov P., Contrino S., Smith R., Zha, Z., Lyne R., Carr A., Lloyd P., Kephart E., McKay S.J., Micklem G., Stein L.D., Lewis, S.E. "The modENCODE Data Coordination Center: Lessons in Harvesting Comprehensive Experimental Details." Database (Oxford), 14, 1-17.(2011).
[18] Mouse ENCODE Consortium, Stamatoyannopoulos J.A., Snyder M., Hardison R., Ren B., Gingeras T., Gilbert D.M., Groudine M., Bender, M., Kaul R., Canfield, T., Giste E., Johnson A., Zhang M., Balasundaram G., Byron R., Roach V., Sabo P.J., Sandstrom R., Stehling A.S., Thurman, R.E., Weissman S.E., Cayting P., Hariharan M., Lian J., Cheng Y., Landt S.G., Ma Z., Wold B.J., Dekker J., Crawford G.E., Keller C.A., Wu, W., Morrissey, C., Kumar, S.A., Mishra, T., Jain, D., Byrska- Bishop, M., Blankenberg D., Lajoie B.R., Jain G., Sanyal A., Chen K.B., Denas O., Taylor J., Blobel G.A., Weiss M.J., Pimkin M., Deng W., Marinov G.K., Williams B.A., Fisher-Aylor K.I., Desalvo G., Kiralusha A., Trout D., Amrhein H., Mortazavi A., Edsall L., McCleary D., Kuan S., Shen Y., Yue F., Ye Z., Davis C.A., Zaleski C., Jha S., Xue C., Dobin A., Lin W., Fastuca M., Wang H., Guigo R., Djebali S., Lagarde J., Ryba T., Sasaki T., Malladi V.S., Cline M.S., Kirkup V.M., Learned K., Rosenbloom K.R., Kent W.J., Feingold E.A., Good P.J., Pazin M., Lowdon R.F., Adams L.B., "An Encyclopedia of Mouse DNA Elements (Mouse ENCODE)." Genome Biology, 13(418), 1-5. (2012).
[19] The ENCODE Project Consortium, "An Integrated Encyclopedia of DNA Elements in The Human Genome." Nature, 489(7414), 57–74.(2012).
[20] The 1000 Genomes Project Consortium, "A Map of Human Genome Variation From Population Scale Sequencing." Nature, 467(7319), 1061–1073.(2010).
[21] Dayhoff M.O., Robert S., "Comprotein: a computer program to aid primary protein structure determination." AFIPS ,62,345-352 (1962).
[22] Gauthier J., Vincent A.T., Charette S.J., Derome N., "A brief history of bioinformatics." Brief Bioinform. 20(6):1981-1996 (2019).
[23] Gregory J.M., "Emile Zuckerkandl, Linus Pauling, and the Molecular Evolutionary Clock, 1959-1965." Journal of the History of Biology, 31, 155–178, (1998).
[24] Needleman S.B., Wunsch C.D. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". J Mol Biol. 48(3):443-53 (1970).
[25] Mitsuo M., "An efficient algorithm for comparing two protein sequences:

Implementation for microcomputers," Computers & Chemistry,12, 21-25 (1988).
[26] Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman DJ. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res. , 25(17), 3389-402. (1997).
[27] Landrum M. J., Chitipiralla S., Brown G. R., Chen C., Gu B., Hart J., Hoffman D., Jang W., Kaur K., Liu C., Lyoshin V., Maddipatla Z., Maiti R., Mitchell J., O'Leary N., Riley G. R., Shi W., Zhou G., Schneider V., Maglott D., Holmes J.B., Kattman B. L. "ClinVar: improvements to accessing data". Nucleic Acids Res. 48,835-844 (2020).
[28] Landrum M.J., Lee J.M., Benson M., Brown GR, Chao C., Chitipiralla S., Gu B., Hart J., Hoffman D., Jang W., Karapetyan K., Katz K, Liu C, Maddipatla Z., Malheiro A., McDaniel K., Ovetsky M., Riley G., Zhou G., Holmes J.B., Kattman B.L., Maglott D.R., "ClinVar: improving access to variant interpretations and supporting evidence" Nucleic Acids Res. 13(4):452-455. (2018).
[29] Taliun D., Harris, D.N., Kessler, M.D., "Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program". Nature, 590, 290–299 (2021).
[30] Lek M., Karczewski KJ., Minikel EV., Samocha KE., Banks E., Fennell T., O'Donnell-Luria A.H., Ware J S., Hill AJ., Cummings BB., Tukiainen T., Birnbaum D. P., Kosmicki J.A., Duncan L.E., Estrada K., Zhao F., Zou J., MacArthur D.G. "Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans." Nature 536, 285–291 (2016).
[31] Tate J.G., Bamford S., Jubb H.C., Sondka Z, Beare D.M., Bindal N, Boutselakis H, Cole C.G., Creatore C., Dawson E., Fish P, Harsha B, Hathaway C, Jupe S.C., Kok C.Y., Noble K., Ponting L., Ramshaw C.C., Rye C.E., Speedy H.E., Stefancsik R, Thompson S.L., Wang S., Ward S., Campbell P.J., Forbes S.A.,"COSMIC: the Catalogue Of Somatic Mutations In Cancer" Nucleic Acids Research, 47, 941-947(2019).
[32] Bradford Y., Conlin T., Dunn N., Fashena D., Frazer K., Howe D.G., Knight J., Mani P, Martin R., Moxon S.A., Paddock H., Pich C., Ramachandran S., Ruef B.J., Ruzicka L., Bauer Schaper H., Schaper K., Shao X., Singer A., Sprague J., Sprunger B., Van Slyke C., Westerfield M.., "ZFIN: enhancements and updates to the Zebrafish Model Organism Database". Nucleic Acids Res., 39,822-829 (2011).
[33] Thurmond J., Goodman J.L., Strelets V.B., Attrill H., Gramates L.S., Marygold S.J., Matthews B.B., Millburn G., Antonazzo G., Trovisco V., Kaufman T.C., Calvi B.R., FlyBase Consortium, "FlyBase 2.0: the next generation". Nucleic Acids Research., 47,:759-765. (2019).
[34] Harris T.W., "WormBase: a comprehensive resource for nematode research". Nucleic Acids Res., 38,463-467 (2009).
[35] Nicholas F.W., "Online Mendelian Inheritance in Animals (OMIA): a comparative knowledgebase of genetic disorders and other familial traits in non-laboratory animals." Nucleic Acids Res., 34, 599-601 (2003).
[36] Kumar P., Henikoff S., Ng P.C., "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm." Nat Protoc, 4, 1073–1081 (2004)
[37] Ng P.C., Henikoff S., "Predicting deleterious amino acid substitutions." Genome Res., vol. 11, 863-74 (2001).
[38] Adzhubei I.A., Schmidt S., Peshkin L., Ramensky V.E., Gerasimova A, Bork P, et al. "A method and server for predicting damaging missense mutations." Nat Methods Nature Publishing Group., 7(4),248-249 (2010).
[39] Adzhubei I., Jordan D.M., Sunyaev S.R.," Predicting functional effect of human missense mutations using PolyPhen-2." Curr. Protoc. Hum. Genet. vol.5 Chapter 7 (2013).
[40] Davydov E.V., Goode D.L., Sirota M., Cooper G.M., Sidow A., Batzoglou S.

,“Identifying a high fraction of the human genome to be under selective constraint using GERP++.” PLoS Comput Biol. vol. 6,12 (2010).
[41] Cooper G.M., Stone E.A., Asimenos G., NISC Comparative Sequencing Program, Green E.D., Batzoglou S., et al. “Distribution and intensity of constraint in mammalian genomic sequence.” Genome Res., vol. 15,7 (2005).
[42] González-Pérez A., López-Bigas N., “Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score.” Condel Am J Hum Genet., 88 , 440-9 (2011).
[43] Kircher M., Witten D.M., Jain P., O’Roak B.J., Cooper G.M., Shendure J. ,“A general framework for estimating the relative pathogenicity of human genetic variants”. Nat Genet., 46(3), 310-5 (2014).
[44] Grimm D.G., Azencott C.A., Aicheler F., Gieraths U., MacArthur D.G., Samocha K.E., “The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity.” Hum Mutat Wiley Online Library, 36, 5 (2015).
[45] Capriotti E., Calabrese R., Casadio R., “Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information.” Bioinformatics., 22, 22 (2006).
[46] Wu C.H., Apweiler R., Bairoch A., Natale D.A., Barker W.C., Boeckmann B., “The Universal Protein Resource (UniProt): an expanding universe of protein information.” Nucleic Acids Res., 34,187-91. (2006).
[47] Sherry S.T., Ward M.H., Kholodov M., Baker J., Phan L., Smigielski E.M., “dbSNP: the NCBI database of genetic variation.” Nucleic Acids Res., 29, 308-311 (2001).
[48] Lek M., Karczewski K.J., Minikel E.V., Samocha K.E., Banks E., Fennell T. “Analysis of protein-coding genetic variation in 60,706 humans.” Nature. 536, 285–291 (2016).
[49] Pir M.S., Bilgin H.I., Sayici A., Coşkun F., Torun F.M., Zhao P., Kang Y., Cevik S., Kaplan O.I. , “ConVarT: a search engine for matching human genetic variants with variants from non-human species,” Nucleic Acids Research, Volume 50, 1172-1178 (2022).
[50] Colige A., Vandenberghe I., Thiry M., "Cloning and characterization of ADAMTS-14, a novel ADAMTS displaying high homology with ADAMTS-2 and ADAMTS-3". J. Biol. Chem., 277, 5756–66. (2002).
[51] Roberts J.H., Halper J. “Connective tissue disorders in domestic animals” Adv Exp Med Biol, 1348, 325-335, (2021).
[52] Malfait F., Castori M., Francomano C.A., Giunta C., Kosho T., Byers P.H.
“The Ehlers-Danlos syndromes” Nat Rev Dis Primers ,6, 64, (2020)
[53] Lvovs D., Favorova O.O., Favorov A.V., "A Polygenic Approach to the Study of Polygenic Diseases" Acta Naturae, 4 , 59–71. (2012).
[54] Jackson M., Marks L., May G.H.W., Wilson J.B., "The genetic basis of disease". Essays in Biochemistry, 62(5),643-723. (2018).

CURRICULUM VITAE

2013 – 2018	Biology, Middle East Technical Universty, Ankara, TURKEY
2020 – 2022	Present M.Sc., Bioengineering, Abdullah GÜL Universty, Kayseri,TURKEY

SELECTED PUBLICATIONS AND PRESENTATIONS

There is no publication and presentation.