Federal State Autonomous Educational Institution of Higher Education
«Ural Federal University named after the first President of Russia B.N.Yeltsin»

Federal State Budgetary Institution of Science
Institute of Immunology and Physiology of the Ural Branch of the Russian Academy of Sciences

On the rights of manuscript

SHINWARI KHYBER

NOVEL GENE VARIANTS IN THE EVALUATION OF INBORN ERRORS OF IMMUNITY: RBCK1 DEFICIENCY, CONGENITAL NEUTROPENIA, HENNEKAM SYNDROME

3.2.7 – Immunology

Dissertation for the degree of Candidate of Science in Biological science

Scientifics supervisers

- **Irina Aleksandrovna Tuzankina**, Doctor of Medical Sciences, Professor, Russian Academy of Sciences;
- **Valeriy Aleksandrovich Chereshnev**, Academician, Doctor of Medical Sciences, Professor, Russian Academy of Sciences.

Ekaterinburg - 2023

TABLE OF CONTENTS

INTRODUCTION

Relevance of the research topic. The immune system is a complex biological system designed to combat foreign antigens, recognize foreign external and internal antigens, destroy infected and abnormally developing cells, as well as to control tolerance to autoantigens and commensal microbiota, thereby fulfilling the important biological task of species preservation.

As a result, innate immune deficiencies (IID) or primary immunodeficiencies (PID) can lead to increased susceptibility to infections, autoimmune processes, autoinflammatory diseases, malignancies, or allergies. The cause of this may primarily be genetic changes, both at the level of the genome and of individual genes that encode protein molecules involved in immune mechanisms.

Although until recently PIDs were considered rare diseases and individual genetic disorders may not occur frequently, in aggregate they can affect a significant number of people. Moreover, due to improved diagnosis and the development of next-generation sequencing (NGS) technologies, the reported prevalence of primary immunodeficiencies (PIDs) has increased to approximately 40 per 100,000 population in recent years [89, 164].

In order to develop new methods for the diagnosis and therapy of immunopathology, a deep understanding of the functioning of the immune system at all levels of the organism is necessary. The emergence of high-throughput biological methods has allowed for an unprecedented understanding of the molecular mechanisms underlying the dynamics of the immune system and its interplay with other systems in the body. However, the tremendous complexity of all the parameters, spanning several orders of spatial and temporal scales, can only be grasped through the use of systems computational immunology - in particular, through the use of computational approaches for processing and modeling large immunological data.

Our work focuses on three diseases: congenital neutropenia (one of the most common forms of PID), Hennekam syndrome (one of the rarest), and RBCK1

deficiency, which is classified as an autoinflammatory PID but also has an increased susceptibility to pyogenic infections. RBCK1 deficiency was first described in 2012 [200], Hennekam syndrome in 1989 [25], and the first genetic discoveries of congenital neutropenia date back to 1999 [137,140]. However, diagnosing each of these syndromes remains a challenge, as new gene variants continue to be identified that lead to the phenotypes of these diseases, and the precise mechanisms of Hennekam syndrome and RBCK1 deficiency are still the subject of debate.

Furthermore, it is important not only to identify gene variants but also to demonstrate their influence on the final product - the protein, whose destabilization can be assessed by in silico tools. This will accelerate the assessment of the pathogenicity of the gene variant and enable the inclusion of identified variants in the list of causative factors to speed up diagnosis, and to approach methods of pathogenetic or gene therapy, which are the ultimate goals of studying congenital human pathology.

Thus, the importance of identifying causative gene variants in immunopathology, as well as searching for the mechanisms of pathology that lead to the phenotype of selected syndromes, have motivated the research goal.

Purpose of the study: To determine the role of potential pathogenic variants of causative genes in the pathogenesis of congenital immune disorders - RBCK1 deficiency, congenital neutropenia, and Hennekam syndrome - using bioinformatics analysis methods.

Research objectives:

1. We will conduct a comparative analysis of gene expression in RBCK1 deficiency relative to healthy children and patients with CINCA/NOMID syndromes, Macleod-Wells syndrome, and mevalonate kinase deficiency.
2. We will assess the pathogenicity of nonsynonymous single nucleotide variants in the ELANE and TCIRG1 genes in congenital neutropenia.
3. We will identify potential new candidate genes involved in the development of diseases belonging to the group of congenital neutropenias.

4. We will identify new variants of the CCBE1, ADAMTS3, and FAT4 genes that lead to the development of Hennekam syndrome.

Methodology and research methods: The work was carried out at the Department of Immunochemistry of the Chemical Technology Institute of UrFU, as well as at the Institute of Immunology and Physiology of the Ural Branch of the Russian Academy of Sciences (Ekaterinburg, Russia), in accordance with the program of fundamental scientific research "Immunological mechanisms of human ontogenesis and their role in the formation of pathological conditions" (state registration number - 01201352044).

Various data sources and research methods were used to solve the tasks set. To investigate the pathogenesis of RBCK1 deficiency, a comparative gene expression analysis was conducted, for which 2 datasets were downloaded from the NCBI Gene Expression Omnibus (GEO): 1) GSE31064, which included data obtained from skin fibroblasts of patients - 2 with RBCK1 deficiency, 1 with MYD88 deficiency, 1 with NEMO syndrome, and 3 healthy individuals (control); 2) GSE40561, which included data obtained from whole blood collected from 2 patients with CINCA/NOMID disease, 5 patients with Muckle-Wells syndrome, 2 patients with hyper Ig-D syndrome, 1 patient with RBCK1 deficiency, and 41 healthy children (control).

Differentially expressed genes in the disease may play a key role in the studied disease or condition and may be potential candidate genes for further research. To this end, a gene expression analysis of two datasets was conducted to search for candidate genes for congenital neutropenia, downloaded from NCBI GEO (https://www.ncbi.nlm.nih.gov). Dataset GSE142347 included 93 female patients, 193 control patients, and 95 affected males, while dataset GSE6322 included 2 parents and 2 children with neutropenia.

Data on various genes and single nucleotide polymorphisms (SNPs) in congenital neutropenia and Hennkam syndrome were downloaded from dbSNP-NCBI (https://www.ncbi.nlm.nih.gov/snp/) and Ensembl (https://www.ensembl.org/index.html). The following SNPs were downloaded for

the study of SNPs in congenital neutropenia genes: for the ELANE gene, 3646 SNPs, of which 301 were nonsynonymous SNPs (nsSNPs); for the TCGIR1 gene, a total of 5627 SNPs, of which 811 were nsSNPs. For the study of SNPs in Hennkam syndrome genes: CCBE1 - 73845 SNPs and 407 nsSNPs; FAT4 - 68257 SNPs and 3434 nsSNPs; ADAMTS3: 70876 SNPs and 911 nsSNPs.

The investigation of gene variants in patients with congenital neutropenia and Hennekam syndrome from the Sverdlovsk region was made possible thanks to sequencing results (performed at the Genome Center Genomed) voluntarily provided by patients for research purposes at the Institute of Immunology and Physiology of the Ural Branch of the Russian Academy of Sciences, and further anonymized.

To assess the harmfulness of nonsynonymous single nucleotide variants on protein structure and function, the following sequence of actions was used. Firstly, all identified nsSNPs in databases were evaluated using the SIFT tool. Then, the sorted probably deleterious mutations were processed through the PolyPhen-2 program, and subsequently sent for evaluation by other bioinformatics tools, including both software and online services, totaling up to 18 - PROVEAN, FATHMM, LRT, M-CAP, META SVM, METALR, Mutation Assessor, Mutation Taster, FATHMM MKLCoding, CAAD, PHD-SNP, Panter, SNP&GO, PON-P2, DANN, SNAP2 - all of which were accessible through VarCard [212] and MutPred [99].

The final result of the filtration, in which the prediction of harmfulness coincided in all tools, was considered as potentially harmful substitutions, and only they were evaluated for their impact on the secondary and tertiary structure of the protein, also assessed through molecular dynamics simulations.

To evaluate the impact of single nucleotide substitutions on the structure and stability of proteins, bioinformatics analysis programs I-Mutant and MU-PRO were used.

For the evaluation of protein-protein interactions, the software packages STRING and Cytoscape were used.

The KEGG database was used for functional enrichment analysis.

The program CemiTool was used for gene co-expression analysis.

In order to build 3D models of the wild-type and mutant protein structures and evaluate the impact of mutations on protein function, the following programs were used: HHPred, Alpha fold 2, Phyre2, I-Taser, Chimera UCSF Chimera, and PyMOL.

Molecular dynamics simulations were performed using the Maestro and Gromacs 4.5.3 packages from Schrödinger LLC. Analysis of whole-genome sequencing data and identification of single-nucleotide polymorphisms (SNPs) was performed on a supercomputer provided by the Shared-Access Equipment Center of the Institute of Mathematics and Mechanics, Ural Branch of the Russian Academy of Sciences (IMM UB RAS - SC Center) in Ekaterinburg. Informed consent was obtained from the parents of the patients for the use of de-identified research results. Statistical analysis, bioinformatics tools, and mathematical models were performed using Python version 3.7.1 (https://www.python.org/) and R version 3.4.3 (https://www.r-project.org/) on the Linux operating system.

Degree of reliability. The reliability of the research results was ensured by the careful application of established research methods and procedures, the use of appropriate data collection tools, and thorough analysis of the collected data. The study design was based on an extensive review of relevant literature, and hypotheses were tested using various statistical and bioinformatics analyses. The applicant's personal contribution consisted of direct participation in all stages of the dissertation research, including the creation of the main idea, planning of scientific work, formulation of working hypotheses, objectives, tasks, determination of the methodology of the dissertation research, interpretation, and analysis of the results obtained, which were conducted by the applicant with scientific supervisors - I.A. Tuzankina, Doctor of Medical Sciences, Professor, Honored Scientist of Russia, and V.A. Cheresnev, Academician of the Russian Academy of Sciences, Doctor of Medical Sciences, Professor. A large and diverse population sample was used for the study, collected from online databases, and the data were analyzed using bioinformatics and computational biology methods. The author, together with a

biotechnologist (an associated professor and a candidate of biological sciences, Dr. Hafiz Musamil Rahman) and a candidate of sciences and senior researcher at the Institute of Immunology and Physiology of the Ural Branch of the Russian Academy of Sciences, Dr. Mikhail Bolkov, conducted bioinformatics analysis, which included analysis of differential gene expression, pathway analysis, gene ontology analysis, protein-protein interactions, in silico nsSNP analysis, protein modeling, and molecular dynamic simulation.

The provisions for defense:

1. A deficit of RBCK1 is associated with reduced expression of genes involved in immune response signaling pathways, inflammatory response, and protein phosphorylation.

2. Congenital neutropenia is associated with a list of genes that expands the spectrum of known genes associated with primary immunodeficiencies.

3. Newly identified non-synonymous single nucleotide variants in the TCIRG1 and ELANE genes have a destabilizing effect on the TCIRG1 and ELANE proteins, respectively.

4. Newly identified non-synonymous single nucleotide variants in genes associated with Hennekam syndrome result in destabilization of the structure and function of CCBE1, ADAMTS3, and FAT4 proteins.

The scientific novelty of research

1. The scientific novelty of the research lies in several key aspects. Firstly, the study identified differences in gene expression in peripheral blood mononuclear cells in individuals with RBCK1 deficiency compared to healthy individuals. This finding sheds new light on the underlying mechanisms of RBCK1 deficiency and may contribute to the development of new diagnostic and treatment approaches.

2. Secondly, the study identified new pathogenic variants in the TCIRG1 and ELANE genes, which were analyzed for their impact on the corresponding proteins for the first time. This information is important for understanding the molecular basis of diseases associated with these genes and could lead to new therapeutic strategies.

3. Thirdly, the study identified new candidate genes for congenital neutropenia, which may be useful in the diagnosis and treatment of this disease in the future.

4. Finally, the study identified new non-synonymous single nucleotide polymorphisms (nsSNPs) in the causative genes of Hennekam syndrome (CCBE1, FAT4, and ADAMTS3) that have a significant impact on the structure and function of these proteins. This information adds to our understanding of the molecular basis of this syndrome and could contribute to the development of new therapeutic approaches.

Theoretical and practical significance of the research

The theoretical significance of the study lies in obtaining new data on genetic-phenotypic relationships, which form the pathogenetic basis of diseases associated with inborn errors of immunity, namely RBCK1 deficiency, congenital neutropenia, and Henneman syndrome, through the development of a program for the sequential use of bioinformatic analysis methods, including molecular dynamics simulations. This will allow the use of the obtained information in further research aimed at identifying therapeutic targets for these diseases.

The practical significance of the study lies in the emerging possibility of using predicted gene variants in the differential diagnostic process when identifying primary immunodeficiency syndromes, such as RBCK1 deficiency, congenital neutropenia, and Henneman syndrome. The developed program for sequential use of bioinformatic analysis methods can be used in the search for new candidate genes associated with inborn errors of immunity.

Implementation of research results in practice

The results of this work have been implemented in the educational process of the graduate program at the Institute of Immunology and Physiology of the Ural Branch of the Russian Academy of Sciences, as well as in the Department of Immunohemistry at the Chemical-Technological Institute of the Ural Federal University named after the first President of Russia, Boris Yeltsin. Additionally, the findings have been incorporated into the scientific research practices of the

Inflammation Immunology Laboratory at the Institute of Immunology and Physiology of the Ural Branch of the Russian Academy of Sciences. Furthermore, the obtained results have been applied in the work of the State Budgetary Healthcare Institution "Maternal and Child Healthcare Clinical Diagnostic Center".

Publications. Regarding publications, the applicant has published 13 works based on the dissertation results, including 5 publications in journals recommended by the Higher Attestation Commission (VAK) with a K1 category rating and indexed in international electronic databases Web of Science Q1-2, Web of Science (Q2-Q4), and Scopus with a total of 6 publications.

Volume and structure of the thesis. The dissertation is composed of 209 pages of typewritten text, and includes an introduction, literature review, materials and methods, three chapters with the results of original research, conclusions, practical recommendations, a list of abbreviations, and a list of references (228 sources, including 12 domestic and 216 foreign). The work is illustrated with 20 tables, 79 figures, and 2 formulas. The language used in the dissertation is both grammatically and scientifically sound.

Acknowledgments. I am sincerely grateful to Mikhail Artemovich Bolkov, senior research fellow at the Laboratory of Inflammation Immunology at the Institute of Immunology and Physiology of the Ural Branch of the Russian Academy of Sciences, for his insightful comments and illuminating discussions.

I am grateful to Professor Yevgeny Yuryevich Gusev, head of the Laboratory of Inflammation Immunology at the Institute of Immunology and Physiology of the Ural Branch of the Russian Academy of Sciences, for his systematic approach to the issues raised in our work, as well as for his important observations on the general molecular mechanisms and signaling pathways investigated in our research.

I am grateful to Boris Hermanovich Yushkov, Doctor of Medical Sciences, professor at the Institute of Immunology and Physiology of the Ural Branch of the Russian Academy of Sciences, and corresponding member of the Russian Academy of Sciences, as well as to Alexei Petrovich Sarapulcev, Doctor of Medical Sciences,

for their precise comments, which allowed us to adjust our work in the necessary direction.

CHAPTER 1 - REVIEW OF THE LITERATURE

Primary immunodeficiencies (PIDs), which arise from congenital errors of the immune system, are classified into more than 485 monogenic syndromes and can affect up to 1% of the population [4, 89]. Without proper detection and treatment, individuals with PIDs are susceptible to serious, long-lasting, and often life-threatening infections, autoimmune and autoinflammatory processes, reparative disorders, and tumors. Despite the achieved successes, awareness of PIDs remains a critical issue for both the medical community and the general population, as it should lead to improved diagnosis and timely application of modern and effective therapeutic methods, which will not only improve the quality of life but also save it for those suffering from these diseases. Meanwhile, patients with various manifestations of the disease do not always have a chance of recovery due to the delayed identification of the underlying cause - the presence of primary immunodeficiencies, which can be observed, for example, in patients with oncological pathology, despite the well-known fact of an increased risk of malignant tumors in PIDs compared to the population without PIDs. The same can be observed in autoimmune or autoinflammatory pathology, which may be the only manifestation of PIDs [149].

1.1 - PID classification, Prevalence, Diagnosis, treatment and prevention of PID

The symptoms of primary immunodeficiencies can include infectious diseases, which may be caused by inherited immune system errors [115], often characterized by many other deviations, including increased susceptibility to malignant tumors, lymphoproliferative, autoimmune diseases, and autoinflammatory disorders [4, 89]. Since 2017, all primary immunodeficiencies have been classified as inherited immune system errors, but the term "primary immunodeficiency" still appears in the International Classification of Diseases, 11th revision [21]. Access to research related to primary immunodeficiencies is difficult

due to the rarity of disorders and the lack of known causal genetic defects [103]. Phenotypic manifestations of diseases, like other consequences of genetic pathology, can appear at any age, but more severe forms of the disease are more commonly encountered in infancy or early childhood. Since PID often manifests as infections with various symptoms and clinical manifestations, in practical healthcare, infections are often treated while overlooking the underlying cause [9]. This often leads to recurrent infections, early chronicity, development of severe complications and disease progression, disability, irreversible organ damage, or even death. For example, in the United States in the 2010s, the average time from the onset of symptoms to the diagnosis of PID was 12.4 years [228]. This means that many people with primary immunodeficiency (PID) face recurrent infections with negative consequences that can affect their personal, social, and professional life for over a decade. However, after recognition of inherited immune system defects and treatment of PID, patients can lead a normal, productive life provided that they receive pathogenetically justified therapy or radical curative measures [2, 10, 51]. To address this critical problem, the Jeffrey Modell Foundation (JMF) has established a global network of specialized centers and developed 10 warning signs of PID [74]. The full original document can also be found on the official JMF website. However, at present, an improved version of 12 warning signs is used in Russia, which places special attention on non-infectious manifestations. They are more diverse, and in many syndromes, infectious manifestations are absent or secondary.

In November 2021, the National Association of Experts on Primary Immunodeficiencies (NAEPID), together with the charitable foundation "Podsolnukh," which provides assistance to children and adults with immune disorders, updated the list of warning signs for primary immunodeficiencies for use by healthcare specialists of various profiles. This list focuses on the signs of primary immunodeficiencies that manifest primarily in the first year of life and often lead to fatal outcomes in early childhood (Figure 1).

Warning signs of primary immunodeficiencies:

1. Family history - the presence of PID cases in relatives.
2. Frequent bacterial infections.
3. Severe course of bacterial infections requiring the use of intravenous antibiotics.
4. Infections caused by opportunistic pathogens.
5. Severe atypical skin manifestations, edema.
6. Inflammatory bowel disease with early onset and/or severe course.
7. Decreased values in complete blood count.
8. Prolonged enlargement of lymph nodes, liver, spleen.
9. Significant reduction in the size of the thymus, lymph nodes, and tonsils.
10. Recurrent fevers without foci of infection.
11. Combination of multiple autoimmune disorders, including endocrinopathies.
12. Facial features (congenital malformations and minor developmental anomalies).

Figure 1 - Warning signs of primary immunodeficiencies

C. Picard et al. (2015) determined the prevalence of primary immunodeficiency disorders (PID) as 1 case per 1200, with a range of 1:600 for IgA deficiency and 1:20,000 for all immunodeficiencies, 1:50,000 for T-cell immunodeficiency, and 1:100,000 for X-linked agammaglobulinemia [227]. Recent data suggest that inborn errors of immunity occur in 1% of the population [79].

Immunodeficiency disorders are considered more significant for healthcare planning in countries where deaths from common infections have been almost completely eliminated, and children with PID survive long enough to be identified. Therefore, Pilania et al. (2019) claim that PIDs are most often detected in countries where infant mortality does not exceed 15/1000 births [57]. Epidemiological observations of PIDs in Asian countries such as Japan and Korea date back to the 1950s. The first survey and registration program for PID patients in Japan was created in 1974 with the establishment of the Immunodeficiency Registration Center in the Pediatrics Department of the University of Tokyo [84]. 497 patients were registered in Japan from 1966 to 1975. Among them, the most commonly diagnosed PIDs were IgA deficiency, X-linked agammaglobulinemia (XLA), and ataxia-telangiectasia. The Ministry of Health, Labor and Welfare of Japan established a research program that created a clinical research group to conduct epidemiological, pathological, diagnostic, and therapeutic research on PIDs, and by 2008, the Primary Immunodeficiency Diseases Network (PIDJ) database network was created to expand research opportunities and patient service. In 2011, 1240 PID patients were registered, and the prevalence of the disease was 2.3 per 100,000 population [142]. Although this prevalence was higher than in earlier reports in Japan, it was much lower than in Western countries and the Middle East. Several reasons for this discrepancy have been postulated.

In Japan, several factors may contribute to the lower prevalence of primary immunodeficiency disorders (PIDs) compared to Western countries, including low levels of consanguinity in the region, sampling bias (asymptomatic selective IgA deficiency, transient hypogammaglobulinemia of infancy, and some other PIDs were not included in this study), and lower detection rates of PIDs in adults [142].

Currently, there are facilities for the diagnosis and treatment of PIDs in 66 hospitals throughout Japan [209].

The earliest reports of PIDs from China were published in the 1960s. Interest in PIDs in China became more apparent in the 1980s [116]. In 1981, a section of pediatric immunology was established at the Chinese Pediatric Society of the Chinese Medical Association, and in 1998, a joint network and patient registry for PIDs was established. The largest cohort of PID patients was registered at the Children's Hospital of Chongqing Medical University, where a diagnosis was made for 352 patients between 2005 and 2011, with genetic analysis performed in 203 patients [219]. Large cohorts of PID patients have also been identified at other medical centers in China, including the Children's Hospital of Fudan University in Jiaotong, the Beijing Children's Hospital, and the Guangzhou Children's Hospital.

In the region of Taiwan, the Primary Immunodeficiency Care and Research Institute (PICAR) at Chang Gung Memorial Hospital in Taoyuan serves a population of approximately 23 million people and has diagnostic and treatment facilities for various primary immunodeficiency disorders (PIDs) [57]. Another similar center is located in Taipei. The incidence of PIDs in Taiwan was 2.17 per 100,000 live births, and Taiwan was the first region in Southeast Asia where a nationwide newborn screening for PIDs was conducted in 2012.

The University of Hong Kong established a specialized service for children with PIDs in 1988, and in 1995, conditions for molecular diagnosis of PIDs were first established. Currently, the University of Hong Kong conducts genetic diagnosis for several PIDs using whole genome sequencing.

Due to the high degree of consanguinity in the Middle East, a large number of PID cases have been reported in Turkey and Iran [15, 28]. Autosomal recessive disorders are more common in these countries. The first department of pediatric immunology was established in the children's hospital of Hacettepe University in 1972. In 1974, the Turkish Society of Immunology was founded. There are also opportunities for hematopoietic stem cell transplantation (HSCT), and to date, about 80 patients with SCID have received HSCT in Turkey. Recently, two Jeffrey Modell

Foundation (JMF) Centers for Immunodeficiencies have been established in Turkey: the Department of Pediatric Allergy and Immunology at Marmara University in Istanbul and the JMF Center at Hacettepe University [208].

The first center of clinical immunology and allergy in Iran was established by Professor Abolhassan Farhoudi at the Children's Medical Center of Tehran University of Medical Sciences in 1977 [6]. In 1999, a database for registering Iranian patients with primary immunodeficiency disorders (PIDD) was created - the Iranian Registry of Primary Immunodeficiencies (IPIDR), which is located at the Children's Medical Center and covers major hospitals throughout Iran. By 2018, it had registered 3,056 patients (with 1,395 new cases) [2]. The Iranian Primary Immunodeficiency Association (IPIA) was founded in 1998 with the goal of improving the diagnosis, management, and treatment systems, as well as promoting research and education in the field of PIDD. Several centers also have the ability to perform hematopoietic stem cell transplantation for PIDD patients.

Significant progress has been made in understanding the pathogenesis, diagnosis, and treatment of these diseases over the past three decades. However, in many developing countries, these diseases still remain insufficiently recognized. This is mainly due to the lack of awareness among doctors, as well as the absence of diagnostic equipment in resource-limited countries.

The earliest reports of primary immunodeficiency (PID) cases in India date back to the late 1960s. Initially, cases of patients with Wiskott-Aldrich syndrome (WAS), agammaglobulinemia, and ataxia-telangiectasia were reported [128, 131, 214]. In 2012, Gupta et al. published a study comparing the clinical profile of PID patients in two large pediatric centers in India, the Advanced Pediatrics Centre at the Postgraduate Institute of Medical Education and Research (PGIMER) in Chandigarh and the National Institute of Immuno-hematology (NIIH) and B.J. Wadia Children's Hospital in Mumbai [165]. The profile of PID patients differed between these two centers. Antibody deficiency was the most common PID in Chandigarh, while familial hemophagocytic lymphohistiocytosis (HLH) was the most common PID diagnosed in Mumbai. Other common PIDs diagnosed in Chandigarh were WAS,

hyper-IgE syndrome, ataxia-telangiectasia, and hereditary angioedema. More cases of neutropenia, leukocyte adhesion deficiency (LAD), IFNγ-IL12 pathway disorders, and autoimmune lymphoproliferative syndrome were registered in Mumbai [165]. With increasing awareness, more cases of PID are being diagnosed throughout the country.

In other Southeast Asian countries, including Singapore, centers for diagnosis and management of PID patients are also developing. In Singapore, 39 patients were registered in 2003, and data were collected from three major centers, including the Children's Medical Institute, National University Hospital, Tan Tock Seng Hospital, and Women's and Children's Hospital [168]. Antibody deficiency was the most common PID, followed by phagocytic defects. Since then, there has been a significant increase in the number of diagnosed PIDs in Singapore, and many centers in Singapore have the capabilities to perform genetic testing.

Malaysia and Thailand are also catching up in terms of awareness and diagnostic base for PID. The national PID initiative was initiated in 2007 with the aim of improving the diagnostic base in various centers in Malaysia, which led to an improvement in the treatment and outcomes of PID patients [43]. The Malaysian Primary Immunodeficiency Network (MyPIN) was established in 2009 with the aim of improving the diagnostic and therapeutic base for PID. More than 300 PID patients are registered here [127]. A study published in Thailand reports on 72 patients with various PID from the Ramathibodi Pediatric Allergy/Immunology/Rheumatology Clinic from 1991-2011 [167]. Intravenous immunoglobulin (IVIG) therapy is also available to most PID patients at a subsidized rate in Malaysia and Thailand [43]. The structure of PID diseases in Asia varies in different Asian countries. Due to the high level of consanguinity in the Middle Eastern countries, autosomal recessive (AR) diseases are relatively more common [15]. In some other Asian countries, X-linked forms of the disease are more common. Studies conducted in Japan and China claim that X-linked forms of SCID and chronic granulomatous disease (CGD) are more common than autosomal

recessive forms [197, 219]. As a result of consanguineous marriages, autosomal recessive diseases are also very common in some Southeast Asian countries, such as India, Pakistan, and Bangladesh [169]. Genotype can determine the clinical profile of inherited diseases, it can be altered by many environmental factors and determine the final phenotype [115]. Environmental factors affect the gut microbiota, significant factors may include socio-economic standards and the spectrum of available medical institutions.

In addition to the differences in the PID spectrum observed in Asia compared to the rest of the world, PID patients in Asia also have a unique and distinct pattern of infections that can contribute to morbidity and mortality in these patients. Among these infections, Mycobacterium tuberculosis, Mycobacterium bovis, Burkholderia pseudomallei, and Talaromyces marneffei are prevalent [57]. It has been established that patients with chronic granulomatous disease (CGD) in Asia have a remarkably high prevalence of tuberculosis infection compared to CGD patients from other countries [94]. Due to the higher endemicity of tuberculosis in many Asian countries, Bacillus Calmette Guerin vaccine is administered at birth. Therefore, disseminated BCG infection is a major clinical manifestation of many PIDs in many Asian countries, such as severe combined immunodeficiency (SCID), CGD, hyper-IgM syndrome, and IL12-IFN-γ-mediated defects [57].

It has also been reported a high frequency of arthritis in XLA patients from Asian countries [216]. This is likely due to delayed diagnosis and subsequent delay in the initiation of immunoglobulin replacement therapy in these patients. Chromobacterium violaceum has been registered as an opportunistic infection in phagocytic defects (e.g., CGD) in many Asian countries. Initially, it was reported in patients from Malaysia, then it was reported in Vietnam, Thailand, Sri Lanka, India, as well as in Hong Kong and Taiwan in China [144]. Mortality rates of up to 50% have been reported in infections with this microorganism [224]. Similarly, melioidosis caused by Burkholderia pseudomallei is also endemic in many countries and is a major problem among patients with PID in Asia [98].

In recent times, several primary immunodeficiencies (PIDs) have been identified in association with a predisposition to endemic mycoses (such as Talaromyces marneffei, disseminated coccidioidomycosis, histoplasmosis, and paracoccidioidomycosis) in this region. These fungal infections are usually linked to defects in the IL-12/IFN-γ-mediated pathway, enhanced STAT1 function, and other diseases mediated by Th17 lymphocytes [113, 155].

The oral live polio vaccine is still in use in several Asian countries and poses a significant problem for many patients with PIDs from these countries. Patients with hypogammaglobulinemia often receive it even before the diagnosis of immunodeficiency is established. These patients can also become infected with the vaccine strain of the virus through close contact in family and community, and it is very difficult to eliminate it from the body. Thus, immunodeficiency-associated vaccine-derived polioviruses (iVDPVs) remain a significant problem for these patients [75, 165]. They are also a potential reservoir for poliovirus transmission. In an international multicenter study, poliovirus shedding was studied in 653 patients with PIDs (570 had primary antibody deficiency and 65 had combined immunodeficiency). Thirteen patients (2%) shed polioviruses, and non-polio enteroviruses were detected in 30 patients. Five of them (0.8%) were classified as patients with immunodeficiency-related vaccine-derived poliovirus (iVDPV) [6, 153].

In Russia, according to Mukhin et al., 2020, the minimum overall prevalence of PID is estimated at 1.3 per 100,000 people, with significant variations across federal districts (from 0.9 to 2.8 per 100,000; 3) (Figure 2) [166].

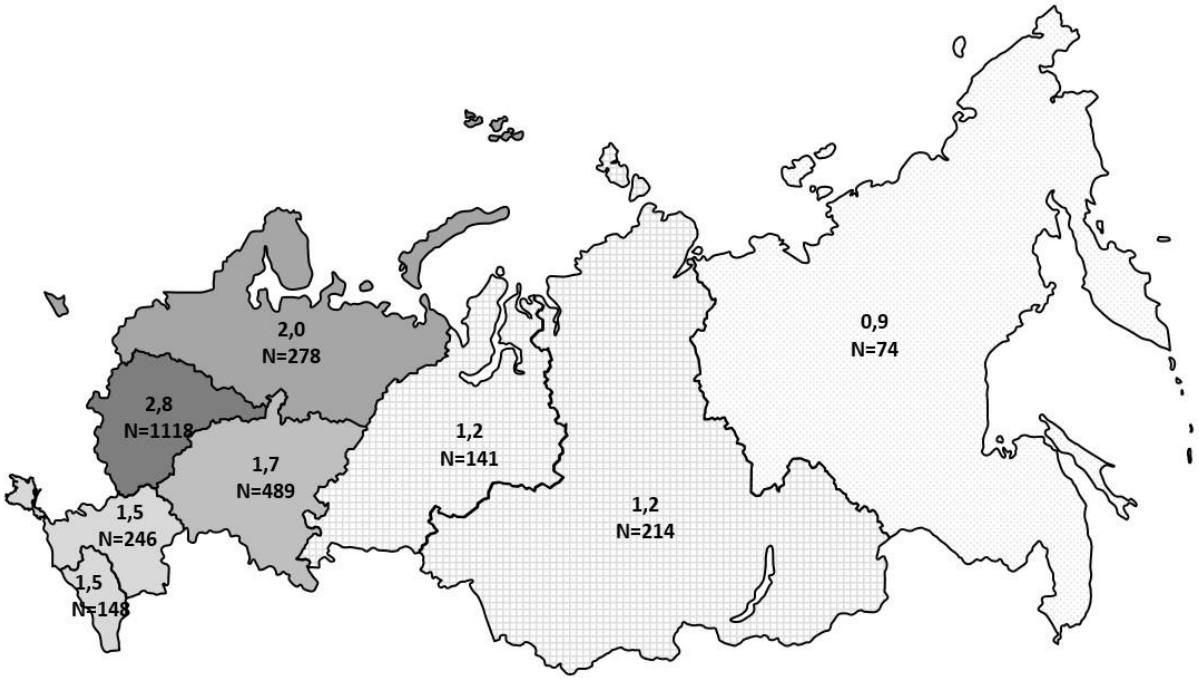


Figure 2 - Distribution of PID (STIs) in Russia by Federal Districts [166]

The Russian National Registry contains information on nearly 3000 patients (60% male, 40% female) from all federal districts of the Russian Federation, with 68% being alive in 2020, of which 77% were children and 23% were adults. PID was diagnosed before the age of 18 in 88% of cases. The most common PID groups were antibody deficiencies (26%) and PID with syndromic features (22%). The overall prevalence of PID in the Russian population was minimal at 1.3 per 100,000 individuals; the calculated birth rate of PID was 5.7 per 100,000 live births. The median delay in diagnosis was 2 years, with this indicator ranging from 4 months to 11 years depending on the PID category [12, 166].

Since 1999, the International Union of Immunological Societies (IUIS) has classified inborn errors of immunity (IEIs) into ten groups, depending on which part of the immune system is affected. One of the identified groups, the tenth, includes autoimmune conditions and somatic variants that mimic genetically determined IEIs. Each IEI group is associated with unique phenotypic manifestations of

infections, autoimmunity, or inflammation. For example, patients with antibody deficiencies usually suffer from bacterial respiratory infections, while patients with deficiencies in the terminal complement fractions are prone to recurrent meningitis caused by Neisseria bacteria. The IUIS report identifies ten PID (IEI) groups, each of which is described in terms of its genetic cause [89].

Classification of Primary Immunodeficiencies 2022:

1. Immunodeficiencies with a combination of cellular and humoral immune defects.
2. Combined immunodeficiencies associated with syndromic manifestations.
3. Predominantly antibody deficiencies.
4. Immunodysregulation disorders.
5. Inherited defects in the number and function of phagocytes.
6. Defects in innate and adaptive immunity.
7. Autoinflammatory syndromes.
8. Complement deficiencies.
9. Bone marrow failure.
10. Phenocopies of primary immunodeficiencies.

In brief, the main differences between primary immunodeficiencies (PIDs) depend on the level of genetic defects and corresponding defects in receptors or proteins. Functionally, the immune system is divided into two main components - innate and adaptive immune responses, and depending on which component of the immune response is primarily impaired, two major groups of immunodeficiencies can be conditionally distinguished. However, this classification most fully reflects the structure and diversity of innate immune errors [112]

Modern methods for treating PIDs include symptomatic support, targeted therapy, replacement therapy, and two types of radical surgery: hematopoietic stem cell transplantation (HSCT) and gene therapy [110] It should be noted that gene therapy is still in the experimental stage of research, although it is already actively implemented in clinical practice for some forms of genetic pathology. For some

children with PIDs, HSCT is the most important and even the only way to treat the disease and restore immune system functions. Moreover, the absence of genetic confirmation is not a contraindication for HSCT [110].

One of the most important achievements of modern medicine is the speed of diagnosis, including screening technologies that allow for the identification of patients at preclinical stages of disease development. This enables timely treatment, preventing the establishment of pathological phenotypes. The analysis of TREC and KREC molecules in blood drops deposited on a Guthrie card is conducted in infants during the first few days of life and allows for the detection of severe combined immunodeficiencies and antibody formation defects that cause life-threatening diseases [1]. As of 2023, such tests for the quantitative determination of TREC and KREC are included in the expanded neonatal screening for all newborns in the Russian Federation (Ministry of Health of the Russian Federation Order No. 274n dated 21.04.2022) [7].

However, the diseases under consideration are not amenable to screening by this method and, like many other congenital conditions, are not immediately detected at birth but rather after a prolonged period of time. In primary immunodeficiencies, the speed of diagnosis is a critically important factor. The identification of SNPs that lead to diseases allows for their inclusion in automatic methods for assessing pathology, such as various diagnostic test panels (including NGS technology) and bioinformatics databases.

1.2 - Primary immunodeficiencies and innate immunity mechanisms

It is known that innate immunity includes epithelial and mucosal barriers, natural antimicrobial products, pattern recognition receptors, and cytokines. It phylogenetically precedes adaptive immunity and is present in all multicellular organisms, including plants, insects, and animals. Although innate immune cells are somewhat primitive, they organize a discrete immune response by recognizing different pathogens through pattern recognition receptors [108, 130, 157]. Neutrophils, macrophages, dendritic cells, natural killer (NK) cells, and NKT cells

in combination with natural barriers (primarily the skin, mucous membranes of the gastrointestinal and respiratory tracts), antimicrobial agents, opsonins (such as complement), and cytokines are the key components of innate immunity.

Inherited immunodeficiencies that lead to increased susceptibility to tuberculosis and nontuberculous mycobacteria are collectively called Mendelian susceptibility to mycobacterial diseases (MSMD) [27, 36].

Macrophages phagocytize mycobacteria, leading to the production of interleukin (IL)-12 p70, the heterodimer of IL-12 p40 and IL-12 p35, as well as IL-23, the heterodimer of IL-12 p40 and IL-12 p19. IL-12 and IL-23 stimulate T and NK cells to phosphorylate signal transducer and activator of transcription (STAT)4 through their cognitive receptors, resulting in the production of interferon (IFN)-γ. The latter acts through its heterodimeric receptor, mainly phosphorylating STAT1 and activating interferon-responsive genes that contribute to mycobacterial clearance. Inherited immunodeficiencies leading to increased susceptibility to tuberculosis and nontuberculous mycobacteria are collectively referred to as Mendelian susceptibility to mycobacterial diseases (MSMD) [27, 36]. In recent years, it has been established that patients with MSMD have mutations in seven different genes: IFNGR1, IFNGR2, STAT1, IL12B (IL-12p40), IL12RB1, TYK2, and IKBKG (NEMO), all of which are involved in IL-12/23-dependent, IFN-γ-mediated immunity. Recently, mutations in the IRF8 gene have also been found to be associated with the development of mycobacterial diseases (MSMD). Specific mutations in these loci account for different forms of inheritance patterns (autosomal recessive, autosomal dominant, or X-linked), presence or absence of protein expression (missense or nonsense mutations), severity of the phenotype (complete or partial deficiency), and specific affected function. These syndromes are clinically heterogeneous and range from locally limited to life-threatening, widespread mycobacterial diseases. In addition to mycobacteria, other intracellular bacteria (such as Salmonella), viruses (such as the varicella-zoster virus), and fungi (such as histoplasmosis, coccidioidomycosis, and paracoccidioidomycosis) have been reported in patients with MSMD [181, 220].

Mutations in Interferon-gamma receptor 1 (IFNGR1) were the first to be identified as causing innate susceptibility to mycobacteria. This gene can be mutated in a way that leads to recessive or dominant transmission. The dominant form is most commonly characterized by non-tuberculous mycobacterial osteomyelitis [45]. While recessive complete mutations usually do not allow for protein expression, dominant mutations are characterized by excessive accumulation on the surface of the mutated receptor, which still binds to IFN-gamma but significantly suppresses intracellular signaling. Deficiency of IL12RB1, detected in more than 140 patients worldwide, is the most common form of innate susceptibility to mycobacteria, but appears to be highly sensitive to the environment. Individuals carrying biallelic mutations may demonstrate very weak susceptibility to mycobacteria, Salmonella, or fungi, which correspond to partial penetrance and variable expressivity of this deficiency [21].

Male patients with mutations in NF-κB essential modulator (NEMO) exhibit a wide clinical heterogeneity. NEMO encodes the main modulator of the nuclear factor-κB, also known as IκB kinase (IKK)γ, a critical component of the IKK complex. Mutations in this gene cause various diseases: amorphic alleles, leading to null mutations, result in the development of pigmentation incontinence in females but are lethal for male fetuses. On the other hand, hypomorphic alleles can also lead to the development of pigmentation incontinence in females but manifest in males as different combinations of X-linked anhidrotic ectodermal dysplasia and immunodeficiency syndrome. X-linked anhidrotic ectodermal dysplasia with immunodeficiency is likely the most common phenotype, but there is a tremendous heterogeneity in this syndrome. Genotype-phenotype associations in this disease are surprisingly elusive, but mutations at the very C-terminus, including stop codons, have been linked to osteopetrosis and/or lymphedema [14, 158].

To date, only two patients with complete Tyrosine Kinase 2 (TYK2) deficiency have been described [62]. Tyk2 is a member of the Jak/STAT signaling family and is constitutively bound to receptors for type I interferons (IFN-alpha and IFN-beta), interleukin-6, interleukin-10, interleukin-12, and interleukin-23. One

patient had a homozygous deletion of 4 base pairs, resulting in an early stop codon. He had a complex clinical phenotype characterized by viral (contagious molluscum, herpes simplex), fungal (oral candidiasis), bacterial (Staphylococcus aureus, atypical Salmonella), and mycobacterial (localized Calmette-Guérin bacillus) susceptibilities, as well as atopic dermatitis, moderate eosinophilia, and increased IgE levels in the blood serum. Recently, a second case was reported in which the patient did not have any allergic manifestations. Thus, Tyk2 deficiency disrupts signaling of type I IFN (susceptibility to viruses), IL-12/IL-23 signaling (susceptibility to mycobacteria and superficial fungi) [177], and IL-6 signaling (susceptibility to S. aureus) [175].

Mutations in STAT1 can be recessive or dominant, leading to deep susceptibility to broad infection in infancy (recessive complete deficiency) or milder susceptibility to mycobacteria, which manifests later in childhood (partial dominant deficiency) [18].

Epidermodysplasia verruciformis (EV) is a rare genodermatosis characterized by selective susceptibility to keratinocytic-tropic infections of human papillomavirus (subgroup B1) and usually manifests in early childhood [66, 117]. The WHIM syndrome (MIM 193670) is a rare autosomal dominant disorder with a frequency of approximately 1 case per 4.3 million live births [129]. The term "WHIM" is an abbreviation of its main clinical features, including warts, hypogammaglobulinemia, infections, and myelokathexis. Myelokathexis is characterized by a delay in the release of neutrophils from the bone marrow, leading to a decrease in their numbers in the blood. This can result in recurrent bacterial infections, especially of the skin, lungs, and sinuses. Hypogammaglobulinemia associated with WHIM syndrome is characterized by a reduction in all classes of immunoglobulins, making patients vulnerable to bacterial and viral infections. Warts, which are also a common feature of WHIM syndrome, can be persistent and recurrent [129]. Patients with WHIM syndrome may also have delayed bone marrow development, which can lead to myelodysplastic syndrome or acute myeloid leukemia. WHIM syndrome is caused by dominant heterozygous gain-of-function

(GOF) pathogenic variants in the gene encoding chemokine receptor 4 (CXCR4). Since CXCR4 is involved in the retention of neutrophils in the bone marrow, embryonic GOF mutations exacerbate this process, thereby slowing down the release of neutrophils, leading to neutropenia [129]. The ubiquitin system plays an important role in the regulation of TLR signaling. The ubiquitin system is a post-translational modification system that regulates protein function [184]. In some situations, the ubiquitin molecule is attached to target proteins to form polyubiquitin chains. During the synthesis of these polyubiquitin chains, sequential conjugation of the C-terminal glycine residue involves conjugation of the glycine residue in one ubiquitin molecule with one of the seven lysine residues in another ubiquitin molecule [111].

1.3 - Autoinflammatory Syndromes and RBCK1 Deficiency

Autoinflammatory diseases are a broad class of human pathologies associated with innate immunity errors and defects in inflammation mechanisms. This class of diseases was discovered relatively recently, but more than 40 autoinflammatory diseases are now known. The main characteristic of these diseases is uncontrolled autoinflammation in the absence of autoantibodies. Therefore, these syndromes were previously called idiopathic fevers, considering that spontaneous inflammation accompanied by fever is typical for them. However, autoinflammation mechanisms have been identified in many long-known diseases that are not directly related to the fever syndrome, such as obesity, rheumatoid arthritis, and Bechterew's disease. They represent inflammation of serous membranes - pleura, peritoneum, synovial membranes of joints, and eyes [161]

Thus, RBCK1 deficiency (RanBP-Type And C3HC4-Type Zinc Finger-Containing Protein 1) is an autoinflammatory syndrome, characterized by increased susceptibility to infections. In addition, RBCK1 deficiency is characterized by glycogen metabolism disorder leading to its accumulation in muscles (amylopectinosis). Patients with RBCK1 deficiency have broad and variable clinical manifestations, including fever, infectious syndrome (various skin inflammations,

recurrent bacterial infections, up to sepsis), as well as myopathies, cardiomyopathies, and encephalopathies [161].

It is known that RBCK1, also known as HOIL-1, is involved in the assembly of the linear ubiquitin chain complex. Ubiquitins are proteins that play the role of "death kisses" for proteins inside cells, marking them with a black tag for cleavage into amino acids in the proteasome. The linear ubiquitin chain assembly complex (LUBAC) includes RBCK1, RNF31 (ring finger protein 31, also known as HOIL-1-interacting protein or HOIP), and SHARPIN (SHANK-associated protein with RH domain). The linear ubiquitin chain assembly complex (LUBAC) binds to linear (Met1) ubiquitin chains and directs several proteins into the classical NF-kB signaling pathway, preventing inflammation and participating in apoptosis [121, 189].

Studies have shown that LUBAC-catalyzed linear ubiquitination in response to TNF-alpha stimulation participates in the activation of the canonical NF-kB pathway and prevents cell death [184]. RBCK1 (58 kDa, also known as HOIL-1) with two RanBP-type zinc fingers and a C3HC4-type RING finger is involved in the recognition of substrates for LUBAC catalysis. Therefore, RBCK1 deficiency affects the regulation of the immune system, leading to the development of autoinflammatory syndromes. RBCK1 (also known as HOIL-1) is a protein that forms a complex of approximately 600 kDa with two other proteins, SHANK-associated RH domain-interacting protein (SHARPIN) and HOIL-1 Interacting Protein (HOIP-1) [32, 111].

Defects in each of the LUBAC proteins individually lead to autoimmune inflammatory syndromes. Known cases of HOIP deficiency in humans are associated with decreased expression not only of HOIP, but also of other LUBAC proteins. It is known that two patients with a HOIP defect have autoimmune inflammation (especially small joint polyarthritis from an early age), recurrent fevers, severe bacterial, viral, and fungal infections, and a pathological reaction to pneumococcal antigens during vaccination. Patients with a RBCK1 defect have a wide range of clinical outcomes, but the cause of this individual heterogeneity is

unknown. However, all cases are accompanied to some extent by defective glycogen accumulation [14, 95, 174].

Patients may simultaneously exhibit chronic autoimmune inflammation and immunodeficiency, including recurrent sepsis [161]. Patients identified to date with RBCK1 mutations (also known as RANBP2-type and C3HC4-type zinc finger-containing protein 1) significantly differ in clinical outcome (skeletal muscle, cardiac muscle, autoimmune inflammation, or immunodeficiency). The explanation for this individual heterogeneity remains unclear, although it has been suggested that the precise location of the variant in the gene may be a predictor of the predominant phenotype, with mutations primarily leading to immunological dysfunction in the N-terminal region of RBCK1, and mutations in the middle or C-terminal portions leading to a (cardio-)myopathy phenotype [161]. M1-linked linear polyubiquitination is mediated by LUBAC, a complex modification that makes nuclear factor-kB (NF-kB) and its pleiotropic immune system critical for nuclear translocation and transcriptional control. RBCK1 and HOIP contain a RING-between-RING (RBR) domain. The linear ubiquitin assembly complex (LUBAC), which includes HOIL-1-interacting protein (HOIP), Heme-oxidized IRP2 ubiquitin ligase-1 (HOIL-1), and SHANK-associated RH domain interactor (SHARPIN), often associates linear (Met1) ubiquitin chains in the canonical NF-kB pathway with many target proteins [73]. The linear ubiquitin-specific deubiquitinase OTULIN controls the function of LUBAC. Immune dysregulation is observed in mice and humans with defects in the processes of linear ubiquitination and K63 deubiquitination [13].

HOIP is the catalytic subunit of the linear ubiquitination assembly complex (LUBAC), which is essential for NF-kB signaling and therefore for proper innate and adaptive immunity. To date, HOIP deficiency has been identified in only one individual with symptoms such as immunodeficiency, systemic lymphangiectasia, and autoinflammation [186].

HOIP deficiency is also manifested by lymphangiectasia in systemic edema, gastrointestinal tract, and hypoalbuminemia, which can cause malabsorption.

Molecular studies have established that fibroblasts and B-cells from patients who are not responsive to immune stimuli and unable to maintain stable regulation of NF-kB activity have an immunodeficient phenotype observed in the patient. Compared to immune responses in fibroblasts, HOIP and HOIL1-deficient peripheral blood mononuclear cells (PBMCs) were highly reactive to IL-1 stimulation and expressed proinflammatory cytokines IL-6 and MIP-1a. [198]. The HOIL-1 deficiency in patient cells resulted in a decrease in IKK kinase phosphorylation, a slowing of alpha IIB degradation, and a decrease in NEMO ubiquitination in response to TNF or IL-1β stimulation, and a lower level of NF-aB activation in patient cells was associated with a decrease in NF-aB transcriptional activity. The catalytic center of the Linear Ubiquitin Chain Assembly Complex (LUBAC) in fibroblasts and B cells from patients with HOIL-1 deficiency was relatively undetectable, indicating a deficiency in LUBAC in these patients. LUBAC is active in the NF-κB pathway and binds linear polyubiquitin chains to unique Lys residues of the NEMO protein. Human fibroblasts with HOIL-1 deficiency exhibit weakened NF-κB activation [13, 193], resulting in weak transcription of genes controlled by NF-κB and cytokine development in response to TNF and IL-1β. These data are consistent with the results of mouse cell studies with RBCK1 knockout or knockdown gene [33].

1.4 - Monogenic, Multigenic, and Allogeneic Defects in Congenital Neutropenia

Neutropenia is a common disorder that pediatricians regularly encounter, and it is a serious health problem. In neutropenia, the absolute number of polymorphonuclear cells decreases, making the body more susceptible to infections. As a result, infections often become exceptionally severe or occur with an unusually high frequency. Neutrophils are an important component of innate immunity and a key product of hematopoiesis. The number of neutrophilic granulocytes in peripheral blood is used to determine the severity of neutropenia. In most cases, the etiology of neutropenia is iatrogenic and well-known to the treating physician. Allogeneic or

autoantibodies are the second most common cause of neutropenia. The ability of certain viral infections to induce neutropenia is well known [133, 144].

In addition to deviations from the normal range of leukocyte counts in childhood and variations in the average number of neutrophils in individuals of different races, an absolute neutrophil count below 1500/μl is considered neutropenia and is often the initial symptom of this condition. Neutropenia can take several forms, classified as mild when the absolute neutrophil count is between 1000-1500/μl, moderate when it is between 500-1000/μl, and severe when it is less than 500/μl. Neutropenia is a common hematological condition in multiple primary immunodeficiencies with various genetic defects, ranging from congenital phagocytic defects to complicated immunodeficiencies, and can be used to screen for acute infections [4, 90].

Inheritance of congenital neutropenia associated with PID is complex, ranging from isolated severe congenital neutropenia to mental retardation, organ anomalies, facial dysmorphisms, and depigmentation of the skin. Phagocytic innate anomalies are divided into two categories according to the IUIS classification: deficiency of phagocytes (neutropenia) and dysfunction of phagocytic cells [4, 89]. Chronic or intermittent neutropenia may occur in a variety of inherited immune system disorders, including various forms of antibody deficiency, reticular dysplasia, WHIM syndrome, and other diseases. The main pathophysiological causes of severe chronic neutropenia in patients with PID include abnormal differentiation of bone marrow cells, improper release of granulocytes from the bone marrow, increased apoptosis or increased death of peripheral blood granulocytes [4, 89, 144, 233]. Bone marrow studies have shown that in most patients, myelopoiesis maturation stops at the level of promyelocytes, leading to a decrease in the number of neutrophils but an increase in the number of atypical promyelocytes [223]. Such infectious conditions as otitis, gingivitis, skin infections, pneumonia, deep abscesses, and sepsis in these patients begin in the neonatal period and, without appropriate treatment, persist throughout life. In addition, patients with severe combined

immunodeficiency (SCID) are at increased risk of developing leukemia. The cause of SCID can be variants of various genes [90, 49, 187].

The causes of congenital neutropenia may include defects in neutrophil maturation and function, immune dysregulation syndromes (various hemophagocytic lymphohistiocytosis), some severe combined immunodeficiencies (such as reticular dysgenesis (AK2 defect) and PAC2 activation defect), as well as primary autoimmune neutropenia at different stages of neutrophil development. Typically, patients with congenital neutropenia require antimicrobial prophylaxis and treatment with granulocyte colony-stimulating factor, and radical cure is impossible without hematopoietic stem cell transplantation. Currently, there are over 30 inherited errors of immunity (or primary immunodeficiencies) that can cause neutropenia, and while each condition is rare, the overall prevalence of these conditions in the population is significant, and a good and timely diagnosis is necessary to prescribe adequate therapy. [104, 227].

It is known that mutations in the following PID genes lead to the development of neutropenia and congenital neutropenia: ELANE, HAX1, G6PC3, WASP, JAGN1, GFI1, SEC61A1, CSF3R, LYST, AP3P1, TCIRG1, VPS45, LAMTOR2, SBDS, DKC1, SLC37A4, BTK, CD40, CXCR4, AK2, GATA2, STK4, RMRP, and VPS13B.

Classical congenital neutropenia depends on the function of elastase. Defects in elastase lead to severe congenital neutropenia (SCN) types 1 (ELANE deficiency), 2 (GFI1 deficiency), 3 (HAX1 deficiency or Kostmann's disease), 4 (G6PC3 deficiency), 5 (VPS45 deficiency), glycogen storage disease type 1b (G6PT1 deficiency), X-linked neutropenia/myelodysplasia (WAS GOF mutation), P14/LAMTOR2 deficiency, Barth syndrome (3-methylglutaconic aciduria, type II) (TAZ deficiency, X-linked), Cohen syndrome (VPS13 B deficiency), Clericuzio syndrome (USB1 deficiency), JAGN1 deficiency, 3-methylglutaconic aciduria (CLPB deficiency), G-CSF receptor deficiency (CSF3R), SMARCD2 deficiency, specific granule deficiency (CEBPE), Shwachman-Diamond syndrome (caused by

defects in at least 3 genes, SBDS, DNAJC21, and EFL1), HYOU1 deficiency, and SRP54 deficiency [188].

ELANE (OMIM #130130) encodes neutrophil elastase, a serine protease expressed in myelomonocytic cells and their precursors. Neutrophil elastase is mainly produced at the promyelocytic stage of neutrophil maturation and is retained in the azurophilic neutrophil granules that participate in the destruction of microorganisms [10, 96]. However, even when only this protein is mutated, different clinical pictures of congenital neutropenia are observed, and the exact pathogenesis of each condition remains unclear [68,137,139, 190].

The main mechanisms of neutropenia in the case of a defect in neutrophil elastase are related to endoplasmic reticulum stress (unfolded protein response) caused by the accumulation of misfolded elastase in the endoplasmic reticulum, leading to the activation of death signals [81, 96]. It is known that ELANE becomes the most abundant protein at the promyelocyte stage of neutrophil development, reaching millimolar concentrations in neutrophils, supporting the theory that accumulation of misfolded protein may cause a deficiency of chaperone proteins, which activates death signals and apoptosis of immature neutrophils [31, 33]. On the other hand, mutated neutrophil elastase blocks further differentiation, leading to neutropenia [137]. In addition, the ELANE p.G185R polymorphism is associated with impaired neutrophil differentiation and decreased expression of genes encoding critical hematopoietic transcription factors, cell surface proteins, and neutrophil granule proteins [96, 137].

The T-cell immune regulator 1 gene (TCIRG1) encodes a subunit of the large protein complex known as vacuolar H+-ATPase (V-ATPase). This protein complex acts as a pump for moving protons across membranes. This proton movement helps regulate the pH of cells and their surrounding environment. V-ATPase-dependent acidification of organelles is necessary for intracellular processes such as protein sorting, zymogen activation, and receptor-mediated endocytosis. V-ATPase consists of a cytosolic V1 domain and a transmembrane V0 domain. Alternative splicing

results in many transcript variants. Mutations in this gene are associated with infantile malignant osteopetrosis and severe congenital neutropenia.

The TCIRG1 gene in humans is primarily associated with autosomal recessive osteopetrosis. Molecular analysis has identified six new genes (TNFSF11, TNFRSF11A, CLCN7, OSTM1, SNX10, and PLEKHM1) associated with autosomal recessive osteopetrosis in humans. More than half of all patients with autosomal recessive osteopetrosis have mutations in the TCIRG1 gene [20, 202]. Studies have shown that mice with disrupted Atp6i gene function develop severe osteopetrosis [23, 41]. Despite significant progress in understanding the mechanisms of osteoporotic diseases, the genetic basis of 30% of cases remains unclear [148]. According to research, TCIRG1 mutations include missense, nonsense, small deletions/insertions, splice-site mutations, significant genomic deletions, and intronic mutations [26, 34, 60, 138]. Autosomal recessive osteoporosis type 1 is caused by mutations in the TCIRG1 gene, leading to impaired bone resorption and abnormal accumulation of dense bone tissue. This can lead to fractures, bone marrow insufficiency, neurological problems, and immunodeficiency, which can ultimately result in premature death. This problem can be detected as early as 10 days of age. The most common symptoms of the disease are pathological fractures, bone marrow insufficiency, and compression of cranial nerves, which are caused by abnormalities in bone tissue structure, metabolism, and insufficient foramen expansion of cranial nerves [26]. High bone density can result from impaired bone resorption caused by osteoclast dysfunction, which can lead to serious abnormalities. Some defects may arise at early stages of fetal development, such as microcephaly, progressive deafness, blindness, hepatosplenomegaly, and severe anemia. Secondary intracranial hypertension can often lead to deafness and blindness [198].

There are numerous examples of multigenic (or polygenic) causes of congenital neutropenia (CN), where mutation variants in multiple genes may contribute to the formation of similar or different phenotypes of this disease [48]. Why is CN more multigenic compared to other PID? One possible reason is that the multigenic nature of CN is a result of complex interactions between genes.

Understanding the mechanism of their interaction should help doctors and researchers gain insight into the pathophysiology of PID, enabling improved diagnosis and treatment approaches. The key mechanisms underlying the protein-protein interaction network of genes in congenital neutropenia (CN) remain unclear, lacking a systematic level of interpretation. With the recent accumulation of new gene expression data in CN [16, 50] and modern computational methods [97], there is an urgent need to identify candidate genes for CN. The use of systems biology and bioinformatics methods will accelerate and improve the accuracy of identifying new CN genes, allowing for a deeper understanding of the pathogenetic mechanisms of this disease. Furthermore, this is a cost-effective and fast method that will assist clinicians in diagnosing patients with CN phenotype and unknown genetic causes.

1.5 - Hennekam syndrome, phenotype and genotype

Hennekam syndrome is an autosomal recessive disorder and is one of the rarest forms of primary immunodeficiency, characterized by developmental defects of the lymphatic system [87]. The underlying cause of Hennekam syndrome is primary lymphedema-lymphangiectasia, which is attributed to defects in the development and/or functioning of the lymphatic system. It can affect any part of the body, with a predominance in the lower extremities, intestines, abdominal and pleural cavities. Additionally, patients with this condition often have flattened facial features, a broad nasal bridge, hypertelorism, epicanthus, and other anomalies [25]. Currently, 27 different genes have been associated with primary lymphedema (either isolated or as part of a syndrome). It was previously believed that the common signaling pathway in the pathogenesis of lymphedema was the VEGFR3 receptor signaling pathway. However, this pathway is only responsible for a third of all cases of primary lymphedema, highlighting the existence of additional genetic factors. Hennekam lymphangiectasia-lymphedema syndrome may be caused by mutations in the CCBE1 gene (in 25% of cases), as well as in the FAT4 and ADAMTS3 genes, each of which influences the VEGF-C / VEGFR-3 signaling pathways [37, 67, 136, 226].

Hennekam syndrome type 1, also known as CCBE1-associated Hennekam syndrome, was first described by Dutch physician Raoul Hennekam in 1989 [25]. The main molecular mechanism of lymphedema in Hennekam syndrome type 1 is the reduced ability of mutated CCBE1 (collagen and calcium-binding protein 1, containing an epidermal growth factor domain) to accelerate and concentrate the activation of the primary lymphangiogenic growth factor VEGF-C [226].

For Hennekam syndrome type 2, the cause is a homozygous or complex heterozygous mutation in the FAT4 gene on chromosome 4q28. Interestingly, a mutation in the FAT4 gene can also cause Van Maldergem syndrome (VMLDS2), another disorder in which some symptoms overlap with those of Hennekam syndrome [83].

In a 2017 study, a group of authors led by P. Brouillard identified Hennekam syndrome type 3, in which a heterozygous mutation was found in the ADAMTS3 gene on chromosome 4q13. More importantly, the researchers highlighted the close functional relationship between ADAMTS3 and CCBE1 proteins in the activation of the VEGFR3 molecule, which is a cornerstone for the differentiation and functioning of lymphoid endothelial cells [124]. However, mutations in these genes are only found in some patients, and the genetic etiology of most Hennekam syndrome patients remains unclear, mainly because the syndrome is genetically heterogeneous.

Knowledge about the genetic cause of Hennekam syndrome has allowed for the identification of the involvement of the mTOR (mammalian target of rapamycin) signaling pathway and the discovery of a potential therapeutic target, specifically mTOR inhibitors such as rapamycin and its analogues. mTOR is a protein that plays a role in regulating cell growth and metabolism, and its dysregulation has been implicated in the pathogenesis of several genetic diseases, including Hennekam syndrome. mTOR inhibitors halt the progression of lymphedema and lymphatic malformations, and also have anti-inflammatory and anti-fibrotic effects, but they do not cure patients of existing lymphatic system abnormalities and their overall

efficacy is not high. Therefore, understanding the genetic nature and pathogenesis of Hennekam syndrome will help identify more effective targets for therapy.

1.6 - Problems in the study of primary immunodeficiencies

The following facts make primary immunodeficiency a complex group of diseases for both practicing physicians and scientific researchers [3, 5, 8].

In the coming years, it will be extremely important to ensure universal access to numerous scientific achievements and create a sustainable mechanism for timely consideration of these achievements in future developments [118]. Although next-generation sequencing is a revolutionary method for PID diagnosis, it is not available in many countries, especially in low-income countries. Therefore, there is a real task of achieving accessibility of this diagnostic method and reducing the cost of genetic testing. It is also necessary to make other express tests for screening of antibody deficiency syndromes easily accessible, which potentially could facilitate testing in remote areas of countries with limited resources. In addition, newborn screening for SCID and other lymphopenias represents hope for early diagnosis and treatment of PID, but it needs to be implemented more widely in public and private medical institutions, as it allows for the early detection of PID [9]. Following the United States, several European countries have started pilot studies to implement neonatal screening, or have already introduced it as a government project, as has been done in the Russian Federation [7].

Access inequality to treatment and care for patients with PID, including issues of reimbursement, availability, and creating an organizational structure for access to medical care, etc., needs to be addressed. Additionally, a quantitative analysis of the need for care in various regions of the world, especially in the Asia-Pacific region, is necessary to support advocacy efforts to increase government investment in the treatment and research of PID [56].

As there is still limited awareness among the general public about PID, they are often perceived as "exotic" diseases. Improving awareness, understanding, and timely recognition of new forms of PID can change the lives of many patients in the future. It is necessary to continue working together to maintain the supply of plasma-

derived medicines worldwide, including during times when healthcare systems experience difficulties in blood and plasma supply.

The constant need for the discovery of new advanced treatment methods is the second obstacle because we are uncovering new types of diseases and better understanding their nature.

The second stumbling block is that the phenotypic changes associated with PID are usually very diverse. For example, in the case of patients with Wiskott-Aldrich syndrome, the exact nature of the gene defect, such as missense or nonsense mutation, the exact location of splice site anomalies, can significantly alter the phenotype of the syndrome. The manifestations of mutations in this gene can range from very severe to mild, such as X-linked thrombocytopenia, B-cell lymphoma, frequent bacterial and fungal infections, eczema, low platelet count, or neutropenia. The diagnosis of PID usually requires an in-depth analysis of clinical manifestations in combination with an assessment of the patient's history and genealogy.

The major problem faced by researchers and clinicians is the difficulty in finding information. Publicly available databases contain limited samples, which are also less diverse compared to oncology data. There are very few information resources that connect clinical descriptions and functional genomic data, protein-protein interactions, and signaling pathways. Specialized databases include UniProt, IntAct, STRING, and KEGG.

Several databases, registries, knowledge bases, prediction tools, and expert systems are rapidly evolving in response to diagnostic requirements. According to Richardson A.M. et al.'s 2018 article, the disease spectrum is further refined due to the expansion of immunological, genetic, and epigenetic knowledge. The careful application of these diagnostic tools and bioinformatics will not only help understand these complex disorders, but also enable personalized therapeutic approaches for disease treatment [61]. Krina Samargiti et al. in 2009 explained that tools useful for PID diagnosis can be classified into the following seven categories (Figure 3) [183].

Firstly, the primary resources on PID provide a large amount of information ranked at different levels, ranging from genes to protein structures, disease models to specific diagnostic groups, and so on.

Secondly, there are classifications of PID that contain clinical features.

Thirdly, there are laboratory criteria and corresponding tools.

Fourthly, there are national and international patient registries for PID, supplemented with mutation databases (the fifth category), whose information can be used to compare the case under consideration with previously described cases.

Fifthly, there are bioinformatics tools available for predicting or prioritizing new PID candidate genes, which are also used in PID diagnosis – this is the seventh category.

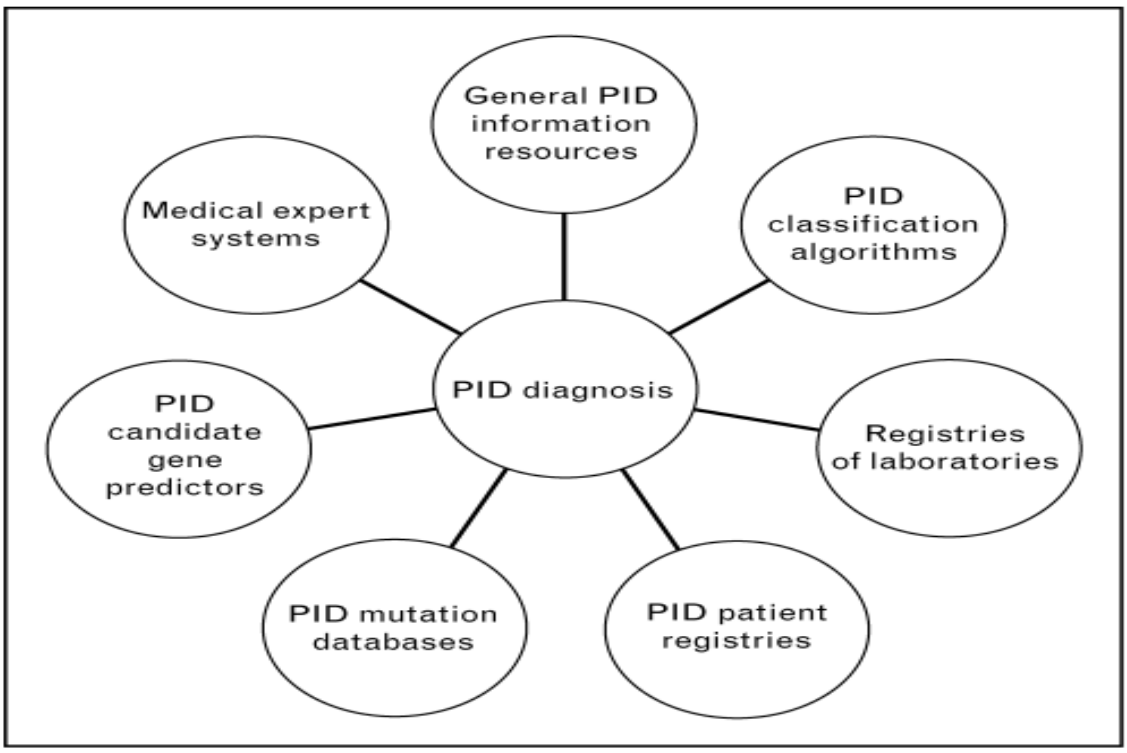


Figure 3 - Schematic grouping of bioinformatics resources and tools that provide information on primary immunodeficiencies [183]

A rational approach to selecting and interpreting genomic analysis in primary immunodeficiencies facilitates the integration of clinical data with immunological and genetic data for establishing a diagnosis [61].

Currently, genome-wide association studies (GWAS) are being conducted on common diseases to identify common low-penetrance causal variants. Some of these variants will alter protein sequences, the most common of which is a non-synonymous single nucleotide polymorphism (nsSNP). The advantage of nsSNPs is the ability to predict their functional impact on protein structure and function, both for the final identification of the causal variant in the disease-associated chromosome region and for further functional analysis of the nsSNP and the associated protein [72].

It is known that non-synonymous SNPs (nsSNPs) alter protein function and are more likely to cause disease in humans. Recent nsSNP studies using computational approaches show the potential impact of mutations on understanding the molecular mechanisms of various diseases [17, 41].

Studies have demonstrated that bioinformatics analysis of gene expression profiles has significant potential in uncovering potential key genes and pathways in the development of complex diseases [30, 180, 218].

The large number of different primary immunodeficiencies (PIDs) poses difficulties in diagnosis, including at the clinical level. Additionally, many diseases are so rare that it is impossible to find a sufficient number of families for analysis. Screening and early detection of PIDs is a serious challenge for physicians. In recent years, high-throughput sequencing has yielded a greater number of known genetic defects. The identification of new candidate genes for PIDs will help prioritize genes for confirmation in PID patients whose exact causal gene has not yet been identified.

In 2009, Keerthikumar et al. used a support vector machine method to classify all human genes into PID genes and non-PID genes. The classification principle was based on calculating a confidence score for each PID gene candidate based on 69 features observed for 148 known PID genes at that time [163]. Based on a literature search, we found that the attention of scientific researchers involved in identifying PID genes has also focused on integrating functional gene ontology (GO) annotations and building datasets of protein-protein interaction networks. In 2018, Guojun Liu and colleagues identified 172 candidate genes for common variable

immunodeficiency (CVID) with similar biological functions to known CVID genes, and eight genes were recently announced as CVID-associated genes [38, 94]. Ortutay et al. (2008) predicted 26 candidate PID genes by analyzing protein-protein interaction network properties (PPI) of all known human immune system genes and their ontologies (GO). In addition, 3,110 candidate disease genes related to PID were predicted based on the calculation of the so-called biological distance (indicating functional interdependence) [133, 159].

Researchers who wish to participate in the study of PID problems face the question of whether PID is a multigenic or monogenic disease. Initially, PID was considered to be a congenital and monogenic disease that follows the principles of Mendelian inheritance [174, 178]. Monogenic diseases result from changes in a single gene that occur in all cells of the body. However, progress in DNA sequencing has led to the discovery of multigenic and somatic causes of PID, and a wide phenotypic variability has been observed for these diseases [80, 211]. Understanding that most PIDs are multigenic in nature is the first step in understanding the pathogenesis of all diseases. According to the multigenic concept, PIDs are the result of complex interactions between genes. Based on this, scientists tried to find the "biological distance" between PID genes and other human protein-coding genes; it was found that PID genes, compared to other human genes, are usually located in the central node of the human genomic network and interact more closely with each other [107]. In addition, PID genes form several closely related subclusters, most of them having at least one functionally close neighbor among a wide range of biological mechanisms [38, 107. 201]. Uncovering these relationships may provide a better understanding of the diversity of genetic pathways underlying PID, which, in turn, will help unlock new opportunities for drug development and therapeutic approaches.

Genetic changes can lead both to a complete or partial loss of a protein (nonsense variants), a decrease in its function (LOF - loss of function), and a gain of function (GOF - gain of function). This is true for any proteins, including key molecules involved in the immune response. To date, pathological variants of more

than 485 genes included in the classification of congenital immunity errors (CIs) are known [89], but a larger number of genes, whose research is in the earlier stages of study, are awaiting detailed description. An important factor complicating the search for causative genetic changes in a large number of diseases is the huge phenotypic heterogeneity of congenital immunity errors, including defects in antibodies, lymphocyte populations and subpopulations, complement system deficiency, autoimmune and autoinflammatory pathologies, lymphoproliferative syndromes, bone marrow failure, and immune dysregulation diseases. In a large cohort of immunodeficiencies, combined immune-dependent processes such as autoimmune and/or immune dysregulation can be observed, especially in cases in which genetic errors lead to alterations in the molecules that regulate the immune response or are involved in providing immune tolerance processes [146].

Despite the fact that, until recently, PIDs were considered rare diseases and individual genetic disorders may be infrequent, collectively they can affect a significant number of people. Moreover, as a result of improved diagnosis, due to the development of next-generation sequencing (NGS) technologies, the reported prevalence of primary immunodeficiencies (PIDs) has increased in recent years to approximately 40 per 100,000 population [89, 164].

It is necessary to consider the complex interrelationships of all genes and proteins in the body, since a simple genotype-phenotypic correlation very often remains unclaimed - patients with a defect of the same gene can have a fundamentally different phenotypic presentation [11].

If there is an assumption that there is an association between a PID phenotype and a gene that has not been previously described from this point of view, thorough functional studies confirming or refuting this association are required to make a statement of a new disease or its new phenotypes. Investigating the values of genetic alterations for immune system function has the unique advantage that immune cells are readily available, usually requiring simple blood sampling to obtain the relevant cells, in contrast to mutations affecting other hard-to-reach tissues (54).

In order to prove causality, studies must demonstrate the significance of a specific pathological gene variant with abnormalities of a specific immune process leading to the corresponding disease phenotype. That is, a functional validation must be performed, which includes assessment of the number and function of proteins, analysis of signaling pathways, and the biological mechanism of pathology implementation [89, 164].

Determining the causal relationship of new mutations is easier when several unrelated families with similar genetic variants and phenotype are identified. However, new diseases may have single descriptions. Some limitations of single-patient studies are the lack of statistical power or the presence of confounding genetic modifiers, which reduces the ability to identify a particular variant as a disease-causing mutation. Experimental modeling of genetic changes in cell lines or animal models overcomes these limitations. To confirm a new gene whose pathological variants can lead to the development of PIDs in a single individual, the following criteria must be met: the genotype found cannot be in individuals without a clinical phenotype; experimental studies must demonstrate that the variant damages, destroys, or alters the function or expression of the gene product; the causal relationship between the genotype and the clinical phenotype must be confirmed in an appropriate cell or animal model [80].

It should be noted that bioinformatics is now becoming an increasingly prominent part of various fields of biology, including molecular biology, statistics and genetics, which play a crucial role in analyzing the expression and regulation of genes and proteins [195]. The study of the effect of single-nucleotide polymorphisms - SNPs - in the coding part of the genome, which directly affect the structure of proteins, is the focus of the vast majority of the scientific community. According to estimates by various researchers, about 90% of genetic variations in humans are due to single-nucleotide polymorphisms. They are determined with a frequency of 1% to 5%, depending on the pathology under study. The values of allele distribution frequencies are important for determining the relevance of SNPs in a particular population and for understanding the potential effect of this SNP on

susceptibility to diseases or other characteristics of interest [179]. The HapMap project, an international collaborative effort aimed at identifying common genetic variations among humans, has described and genotyped over 4 million DNA samples. This has made it possible not only to validate SNPs and estimate the frequency of their alleles in the general population, but also to assess the degree of linkage disequilibrium between them. Moreover, SNP genotyping technologies have recently advanced to the point where hundreds of thousands of SNPs can be typed in thousands of people, for example, using the case-control method. Consequently, the discovery of causal variants for common diseases will accelerate, and it would be helpful if the functional effects of SNPs could be predicted bioinformatically to guide functional studies and narrow down the best candidate SNPs in areas of the genome that exhibit a high degree of disequilibrium [46]. This is why the science of bioinformatics is becoming an integral part of modern research.

The most identifiable category of SNPs is a small fraction of mutations (less than 1%) that alter the protein sequence, and these are usually nonsynonymous substitutions (nsSNPs). The nsSNP prediction tools are used to predict the potential structural and functional impact caused by these variants. In order to more accurately assess the structural impact caused by changes in the amino acid sequence, bioinformatic analysis and protein structure modeling is required to account for changes in the amino acid sequence. Knowledge of the three-dimensional structure of a gene product is of great help in predicting and understanding its function, its role in intracellular processes and in pathology formation, molecular dynamics modeling can be performed to observe changes in many parameters such as protein stability and flexibility. Interdisciplinary modeling (bioinformatics, pathophysiology, genetics and immunology) is gradually becoming a major trend in the development of technologies for clinical research [46].

The identification of candidate genes for various types of pathology requires their verification, which requires not only the use of clinical data but also experimental data, as well as analysis of gene co-expression, activation of biological

signaling pathways, protein-protein interaction, and evaluation of the functioning of the altered protein in the simulation.

Methods for integrating expression profiles and protein-protein interaction (PPI) data are an important part of the ongoing research. Bioinformatics methods are used to study the differential mechanisms of protein interactions in all immune cell lines, transcriptional activators and modules, which are analyzed in the context of examples obtained by clustering the PPI network. The results of such studies demonstrate that integration of protein interaction networks with the most comprehensive database of immune cell gene expression profiles can be used to generate hypotheses about the mechanisms underlying differentiation and differential functional activity across immune cell lines. Comparative analysis of the detected differences between diseased and healthy states helps to obtain pathogenetic characterization of immune-dependent diseases and ultimately lead to the development of new curative methods of pathology correction.

Currently, research on differentially expressed genes defines one of the special scientific directions in which the identification of genes that are differentially expressed in diseases is assumed. In pharmaceutical and clinical research, the results of evaluating differentially expressed genes can be valuable targets for identifying candidate biomarkers, therapeutic targets, and gene signatures for diagnosis. Although specific changes in gene expression do not always lead to subsequent biological activity, such data can nevertheless be combined with other biological data and, with the ability to provide high throughput to create complex analyses, such as building a target disease landscape [123, 213], can be an indispensable research tool [63].

In our work, aimed at finding significant pathophysiological mechanisms for the formation of certain types of immune-dependent pathology, various sites with congenital immune disorders, including congenital neutropenia, RBCK1 deficiency autoflammatory syndrome and Hennecam syndrome, were chosen as models of immune-dependent pathology using research methods of bioinformatics analysis in disorders characteristic of primary immunodeficiencies.

List of work published by the 1st chapter

1. Ретроспективный анализ случаев первичных иммунодефицитов у детей с врожденными пороками сердца / С.С. Дерябина, Д.А. Черемохин, И.А. Тузанкина, Х. Шинвари // Российский иммунологический журнал. – 2020. – Т. 23, № 4. – С. 505-514. (RSCI, ВАК К1).

2. Классификация врожденных ошибок иммунитета человека, обновленная экспертами комитета международного союза иммунологических обществ в 2019 году / М.А. Болков, И.А. Тузанкина, Х. Шинвари, Д.А. Черемохин // Российский иммунологический журнал. – 2021. Т. – 24 (1). – С. 7-68. (RSCI, ВАК К1).

3. Роль врожденных ошибок иммунитета в группе детей с летальными исходами на первом году жизни / Д.А. Черемохин, И.А. Тузанкина, В.А. Черешнев, М.А. Болков, Х. Шинвари // Российский иммунологический журнал. – 2022. – Т. 25 (4). – С. 555-560. (RSCI, ВАК К1).

4. Analysis of the TREC and KREC Levels in the Dried Blood Spots of Healthy Newborns with Different Gestational Ages and Weights / D.A. Cheremokhin, K. Shinwari, S.S. Deryabina, M.A. Bolkov, I.A. Tuzankina, D.A. Kudlay // Acta naturae. –2022. –V. 14 (1). – P. 101–108. (RSCI, ВАК К1).

CHAPTER 2 - MATERIALS AND METHODS USED IN THE WORK

2.1 - Study materials

The data were collected from open data sources, various gene variant databases, as well as two blinded sequencing results of Sverdlovsk patients, which were provided for study to the Institute of Immunology and Physiology, UrB RAS, previously approved by the ethical committee and published.

Two datasets (datasets) from the NCBI GEO database were used to perform the task of investigating the pathogenesis of RBCK1 deficiency, viz: GSE40561, which includes data from whole blood collected from 2 patients with CINCA/NOMID disease, 5 patients with Muckle-Wells syndrome, 2 patients with mevalonatkanase deficiency, 1 patient with RBCK1 deficiency and 41 healthy children (for comparative analysis); GSE31064, which included data obtained from skin fibroblast cells of 2 patients with RBCK1 deficiency, 1 patient with MYD88 deficiency, 1 patient with NEMO syndrome, and 3 healthy patients (from a control group).

Data sets from NCBI GEO were used for the task of investigating candidate genes associated with congenital neutropenia: GSE142347 (patients with congenital neutropenia - 93 women and 95 men, and 193 control patients); GSE6322 (family case - 2 healthy parents and 2 children with congenital neutropenia). A list of 442 known PID genes (and microdeletions) at the time of the study, including 31 genes associated with congenital neutropenia, were obtained from the European Immunodeficiency Society website. CD3D, CD3E, CD3Z, CORO1A, IL2RG, IL7R, JAK3, LAT, PTPRC, ADA, AK2, DCLRE1C, LIG4, NHEJ1, PRKDC, RAC2, RAG1, RAG2, B2M, BCL10, CARD11, CD3G, CD40 (TNFRSF5), CD40LG (TNFSF5), CD8A, CIITA, DOCK2, DOCK8, FCHO1, ICOS, ICOSLG, IKBKB, IKZF1, IL21, IL21R, ITK, LCK, MALT1, MAP3K14, MSN, POLD1, POLD2, REL, RELA, RELB, RFX5, RFXANK, RFXAP, RHOH, STK4, TAP1, TAP2, TAPBP, TFRC, TNFRSF4, TRAC, ZAP70, ZAP70, ARPC1B, WAS, WIPF1, ATM, BLM (RECQL3), CDCA7, DNMT3B, GINS1, HELLS, LIG1,

MCM4, NBS1, NSMCE3, PMS2, POLE1, POLE2, RNF168, ZBTB24, 11q23del, 22q11.2, CHD7, Del10p13-p14, FOXN1, FOXN1, SEMA3E, TBX1, EXTL3, MYSM1, RMRP, RNU4ATAC, SMARCAL1, CARD11, ERBB21P, IL6R, IL6ST, PGM3, SPINK5, STAT3, TGFBR1, TGFBR2, ZNF341, MTHFD1, SLC46A1, TBH2, IKBKB, IKBKG, NFKBIA, ORAI1, STIM1, BCL11B, CCBE1, EPG5, FAT4, KDM6A, KMT2A, KMT2D (MLL2), NFE2L2, PNP, RBCK1, RNF31, SKIV2L, SP110, STAT5B, STAT5B, TTC37, TTC7A, BLNK, BTK, CD79A, CD79B, IGHM, IGLL1, PIK3CD, PIK3R1, SLC39A7, TCF3, TCF3, TOP2B, ARHGEF1, ATP6AP1, CD19, CD20, CD21, CD81, IKZF1, IRF2BP2, MOGS (GCS1), NFKB1, NFKB2, PIK3CDGOF, PIK3R1, PTEN, RAC2, SEC61A1, SH3KBP1, TNFRSF13B, TNFRSF13C, TNFSF12, TRNT1, AICDA, AICDA, INO80, MSH6, UNG, CARD11, IGKC, Mutation or chromosomal deletion at 14q32, FAAP24, PRF1, SLC7A7, STX11, STXBP2, UNC13D, AP3B1, AP3D1, LYST, RAB27A, BACH2, CTLA4, DEF6, FERMT1, FOXP3, IL2RA, IL2RB, LRBA, STAT3, AIRE, AIRE, ITCH, JAK1, PEPD, TPP2, IL10, IL10RA, IL10RB, NFAT5, RIPK1, TGFB1, CASP10, CASP8, FADD, TNFRSF6, TNFSF6, CARMIL2, CD27, CD70, CTPS1, MAGT1, PRKCD, RASGRP1, SH2D1A, TNFRSF9, XIAP, CEBPE, CLPB, CSF3R, DNAJC21, EFL1, ELANE, G6PC3, G6PT1, GFI1, HAX1, HYOU1, JAGN1, LAMTOR2, SBDS, SMARCD2, SRP54, TAZ, USB1, VPS13B, VPS45, WAS, ACTB, CFTR, CTSC, FERMT3, FPR1, ITGB2, MKL1, RAC2, SLC35C1, WDR1, CYBA, CYBB, NCF1, NCF2, NCF4, CYBC1, G6PD, GATA2, CSF2RA, CSF2RB, CYBB, IFNGR1, IFNGR1, IFNGR2, IL12B, IL12RB1, IL12RB2, IL23R, IRF8, IRF8, SG15, JAK1, RORC, SPPL2A, STAT1, TYK2, TYK2, CIB1, CXCR4, TMC6, TMC8, FCGR3A, IFIH1, IFNAR1, IFNAR2, IRF7, IRF9, POLR3A, POLR3C, POLR3F, STAT1, STAT2, DBR1, IRF3, TBK1, TICAM1, TLR3, TLR3, TRAF3, UNC93B1, CARD9, IL17F, IL17RA, IL17RC, STAT1, TRAF3IP2, IRAK1, IRAK4, MYD88, TIRAP, APOL1, CLBH7, HMOX, NBAS, NCSTN, OSTM1, PLEKHM1, PSEN, PSENEN, RANBP2, RPSA, SNX10, TCIRG1, TNFRSF11A, TNFSF11, IL18BP, IRF4, ACP5, ADA2, ADAR1, DNASE1L3, DNASE2, IFIH1 (GOF), OAS1,

RNASEH2A, RNASEH2B, RNASEH2C, SAMHD1, TMEM173, TREX1, USP18, POLA1, MEFV, MEFV, MVKNLRC4, NLRP1, NLRP1, NLRP12, NLRP3, NLRP3, NLRP3, PLCG2, ADAM17, ALPI, AP1S3, CARD14, COPA, HAVCR2, IL1RNIL36RN, LPIN2, NOD2, OTULIN, PSMB8*, PSMB8*, PSMG2, PSTPIP1, SH3BP2, SLC29A3, TNFAIP3, TNFRSF1A, TRIM22, C1QA, C1QB, C1QC, C1R, C1R, C1S, C1S, C2, C3C3, C4A, C4B, C5, C6, C7, C8A, C8B, C8G, C9, CD46, CD55, CD59, CFB, CFB, CFD, CFH, CFH, CFHR1, CFHR2, CFHR3, CFHR4CFHR5, CFHR1CFHR2, CFHR3CFHR4, CFHR5, CFI, CFP, FBH3, MASP2, SERPING1, THBD, ACD, ACD, BRCA1, BRCA2, BRIP1, CTC1, DKC1, DNAJC21, ERCC4, ERCC6L2, FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCI, FANCL, FANCM, MAD2L2, NOLA2, NOLA3, PALB2, PARN, RAD51, RAD51C, RFWD3, RTEL1, RTEL1, SAMD9, SAMD9L, SLX4, SRP72, STN1, TERC, TERT, TERT, TINF2, TINF2, TP53, UBE2T, WRAP53, XRCC2, XRCC9.

The list of 31 genes associated with congenital neutropenia used in this study includes: MTHFD1, LYST, CSF3R, ELANE, JAGN1, LAMTOR2, SMARCD2, VPS13B, WAS, WDR1, CXCR4, TCIRG1, HAX1, G6PC3, GFI1, GATA2, SLC37A4, SBDS, STK4, CLPB, AP3B1, USB1, VPS45, CXCR2, EIF2AK3, RAB27A, AK2, RMRP, TBN2, TAZ, and CD40LG.

A study on gene variants in patients with congenital neutropenia and Henneman syndrome from the Sverdlovsk region was conducted using de-identified data voluntarily provided by the patients' parents for bioinformatic analysis with the approval of an ethics committee. Only VCF files with missense mutations in the FAT4 (Henneman syndrome) and TCIRG1 (congenital neutropenia) genes, as well as de-identified clinical data, were used. Medical observation and clinical research on the patients were carried out prior to our study in medical organizations in the Sverdlovsk region.

Data on various missense mutations for genes associated with the investigated diseases were obtained from public databases. Specifically, FAT4, ADAMTS3, CBEE1, ELANE, and TCIRG1 were obtained from the publicly available dbSNP database on the National Center for Biotechnology Information (NCBI) portal

(https://www.ncbi.nlm.nih.gov/snp/), as well as the Ensmble database (https://www.ensembl.org/index.html), Swiss-Prot database (http://expasy.org./), OMIM (https://www.omim.org/) and HGMD (https://www.hgmd.cf.ac.uk/).

The dbSNP database is an online resource designed to aid researchers in the field of biology. Its aim is to create a unified database containing all identified genetic variations (single nucleotide polymorphisms) that can be used to investigate a wide range of genetically determined natural phenomena. In particular, access to molecular variations catalogued in dbSNP helps to carry out fundamental research, such as physical mapping, population genetics, evolutionary relationships, and enables rapid and quantitative assessment of variations in a particular genomic region (Figure 4). Most of these nucleotide sequence variations were identified through DNA sequencing and genotyping of samples from the general population, in addition to the group of patients (Figure 5).

The Ensembl database (USA) allows for the analysis of transcription for a specific gene, as well as corresponding protein sequences and their various variants. Specifically, for our analysis, we uploaded a CSV file of variants for the genes we investigated into the database (Figure 6).

Figure 4 - Data representations of missense mutations using the CCBE1 gene as an example in the dbSNP database on the NCBI portal

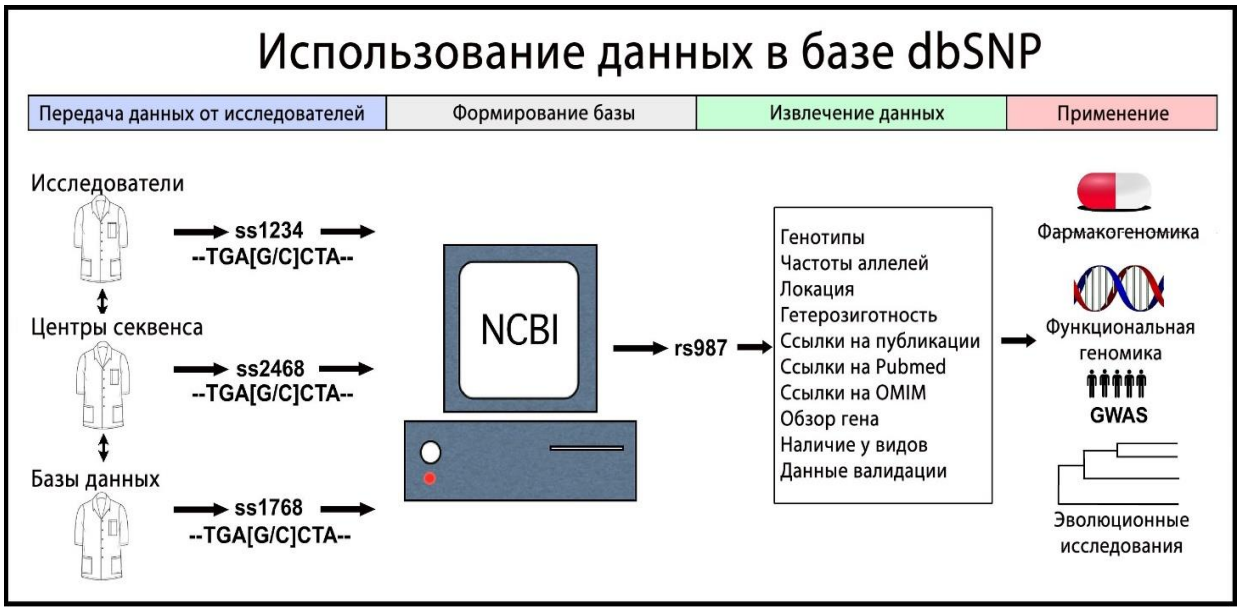


Figure 5 - Data sources and paths of data usage in the dbSNP (non-synonymous single nucleotide polymorphisms) database of the National Center for Biotechnology Information (NCBI, USA)

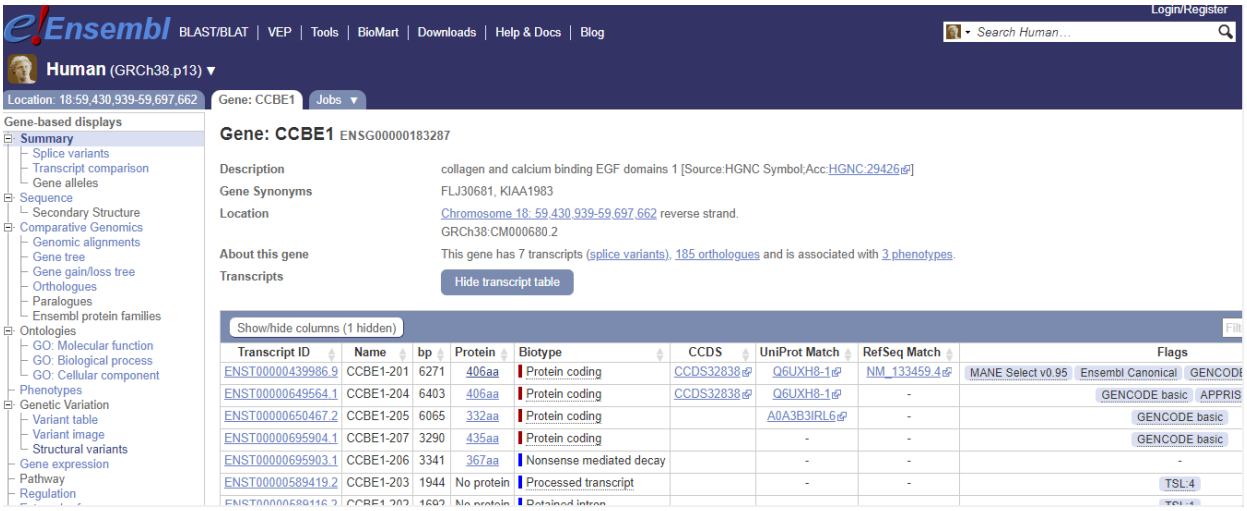


Figure 6 - Ensemble database example for searching information on the CCBE1 gene

Other databases were similarly utilized. For cross-checking of non-synonymous single nucleotide polymorphism (nsSNP) data, information was searched in the Exome Aggregation Consortium (ExAC), Genome Variation Server, Functional Single Nucleotide Polymorphism (F-SNP), Human Gene Mutation Database (HGMD), which compiles known gene mutations responsible for inherited human diseases. Similarly, the Genetic Association Database (GAD), which contains an archive of more than 3600 dbSNP records, and the Human/Genome

Variation Database (HGVBase), which reports associations with diseases based on published clinical studies, were used, although very few of these statistical associations have been confirmed.

The Online Mendelian Inheritance in Man (OMIM) database used in this study is a catalog of genetic disorders of inherited diseases, associated with human genes not only highly penetrant but also rare (MAF - minor allele frequency of less than 0.01 in the population).

We collected data on non-synonymous SNPs from these portals, associated with the studied genes FAT4, ADMATS3, CCBEI, ELANE, and TCIRG1; data related to other factors were excluded. The number of SNPs for the listed genes is displayed in Table 1.

Table 1 - SNPs loaded from the dbSNP and Ensemble databases

Genes	SNP	nsSNP
ELANE	3646	301
TCIRG1	5627	811
CCBE1	73845	407
FAT4	68257	3434
ADAMTS3	70876	911

2.2 - Methods used in the work

2.2.1 - Differential gene expression analysis from data from patients with HOIL-1/RBCK1 deficiency and patients with congenital neutropenia

Differential gene expression analysis (DEG) is a process used to identify genes differentially expressed between two or more conditions, such as normal and disease or conditions under different treatments. This analysis can be performed using bioinformatics tools and pipelines [222].

The DEG analysis procedure involves several steps, including quality control of raw sequencing data, mapping reads to a reference genome or transcriptome,

quantification of gene expression levels, data normalization, statistical analysis to identify differentially expressed genes, and functional analysis of identified genes.

A standard differential gene expression method was used to determine the differences in gene expression in the data set for the RBCK1 deficiency study. The analysis was performed using the Bolstad R package. Differences in gene expression between patients with RBCK1 deficiency and normal samples were evaluated as significant with a P-value < 0.05, |logFC| > 1, and a false discovery rate (FDR) p-value of 0.57 was used as a threshold value [28].

False discovery rate (FDR) determination is a method for conceptualizing the first-order error rate when testing null hypotheses in multiple comparisons. The Log2-value is a cutoff value important for calculating the difference between expression levels.

The false discovery rate method is one of the main statistical tools when annotating genes using GO.

GO is a standardized vocabulary of terms that are used to describe the functions of genes, cellular components, and biological processes in various organisms. Each gene can be annotated to one or more GO terms, which can be used to infer gene function and to compare the functions of different genes [71].

Gene expression differences were calculated using the R Limma package. Functional enrichment analysis of genes characteristic of various congenital primary immunodeficiencies and autoinflammatory diseases was performed using the R Bioconductor package.

Gene Set Enrichment Analysis (GSEA) is a set of methods to link a set of genes to a change in phenotype [225]. Such methods often use databases of previously annotated gene sets to formalize existing phenotype data (e.g., Gene Ontology Project (GO) terms: molecular functions, biological processes, or cellular components [134]. The result of the method (program release) in this case is a set of preannotated sets that help determine whether the ordered list of genes depends on the phenotype or whether they are simply random [225]. Such preannotated sets are

called overrepresented (if the frequency is higher than the background) or underrepresented (if the frequency is lower than the background).

Enrichment coefficient (ES) is a statistical coefficient determined by the Kolmogorov-Smirnov method, reflecting the degree of overrepresentation of genes at the top or bottom of the ranked list of genes.

Over-enrichment analysis (ORA) and Gene Ontology (Gene Ontology) and the signaling pathways involved were performed using the analysis of the borrowed signaling pathways in the KEGG, WikiPathways, reactome, and DAVID databases. Subsequently, DAVID was used to perform analysis in the KEGG database and gene annotation (GO) [71, 109, 200]. The major genes were selected according to their level of connectivity and depicted using Metaphase software [221].

2.2.2 - Prediction of candidate genes for congenital neutropenia

To predict candidate genes for congenital neutropenia, we took the following steps.

First, we used the STRING database to obtain protein-protein interaction (PPI) data for PID and congenital neutropenia genes. The data include genomic context, co-expression, and known and predicted interactions from previous data. The minimum required interaction value was set at 0.4 [192].

Cytoscape (version 3.5.1) was used to estimate gene network density (Dnetwork) and biological distance for congenital neutropenia genes and other primary immunodeficiency genes [58]. Density (Dnetwork) is the most widely used concept in gene regulation and the study of networks of protein-protein interactions (PPIs) and can be used to determine whether a network is dense or not. Network density (Dnetwork) is determined by the formula [88].

$$D_{network} = \frac{\sum_{i=1}^{n} \sum_{j \neq i} a_{ij}}{n(n-1)} \quad (1)$$

where aij is pairwise adjacency, ∑ represents connectivity (network connectivity equals unweighted network connectivity equal to the number of genes that directly the i-th gene), and n is the number of genes in the network.

Note that aij = 1 if gene i and gene j interact in the STRING database, whereas aij = 0 otherwise.

Congenital neutropenia PPI group data (based on the published 32 congenital neutropenia genes) and ten random groups (each group consists of 41 PID genes) were respectively converted into a symmetric adjacency matrix (aij, i, j = 1, □□n) using the "igraph" R package [55].

The network density was used to compare their functional cohesion and proximity. The higher the network density in a group, the closer the interaction of genes in the group. The concept of biological distance (Bi,j) was first introduced by Ethan J. et al. in 2013. With biological distance, researchers studying the functional relationships of genes in a network of genomic interactions do not mean the actual distance between genes in a DNA molecule or on a chromosome, but rather the functional proximity between pairs of genes or within a group of genes [201].

Using the value of biological distance, Itan Y. et al. showed that primary immunodeficiency genes are usually located in the center of the human genomic network and form several closely related subgroups according to different biological mechanisms [107,201]. Biological distance (Bi,j) is determined by the formula:

$$B_{i,j} = \begin{cases} \dfrac{C}{S_{i,j}} & if\ C = 1 \\ \dfrac{C}{S_{i,1} + S_{1,2} + S_{2,3} + \cdots + S_{C-2,C-1} + S_{C-1,j}} & if\ C > 1 \end{cases} \quad (2)$$

where Si,j is the combined index between gene i and gene j obtained from the STRING database, and C is the number of direct connections between gene i and the desired gene. The smaller the biological distance between the groups, the closer the biological relationship between the genes in the group.

The biological distance of a group of known congenital neutropenia genes (32 genes) and two random groups of PID genes (each group consisted of 41 PID genes)

was calculated using the Python package for Human Gene Conectome (HGC), provided by Y. Itan et al., 2015 [107].

Using the "igraph" R package, congenital neutropenia group PPI data and 10 random groups (each with 41 PID genes) were transformed into a symmetric adjacency matrix (aij, I j = 1, n) [55]. Network cohesion or density was determined using network density analysis (a higher network density represents a closer interaction of genes in the group).

Further, the biological distance between genes (Bij) was estimated, which can be used to calculate the shortest functional distances between all possible pairs of human genes [201].

The calculation of the biological distance between the congenital neutropenia gene group (31 genes) and two random PID groups (41 PID genes in each) was performed using the Human Gene Connectome (HGC) tool in Python [201].

The direct search for candidate genes after the preparatory steps was performed in three ways.

1) A Pearson correlation analysis (PCC) was performed to assess the expression of 31 congenital neutropenia genes and each protein-coding gene (or candidate gene) based on data sets GSE142347 and GSE6322 (Downloaded from NCBI using GEO transcriptomic profiles of congenital neutropenia patients). $|r|>0.9$ and $p<0.05$ were used.

2) PPI data for all human protein-coding genes were obtained from J. Cheng et al, 2006 [40], including 217160 interactions provided by eleven databases such as BioGRID [29], HI-II-14_Net [19], HPRD [91], Instruct [101], InnateDB [100], IntAct [102], MINT [205], PINA [160], SignaLink2.0 [191], KinomeNetworkX [172] and PhosphositePlus [118]. The candidate gene was then conserved if the interaction between the congenital neutropenia gene and the candidate gene from the previous step was found in the PPI data.

3) Kyoto Gene and Genome Encyclopedia (KEGG) analysis was performed using the R package "clusterProfiler" to evaluate the biological function enrichment of congenital neutropenia genes [217]. KEGG analysis was then performed for the

remaining candidate genes for congenital neutropenia. A gene was defined as a true congenital neutropenia candidate gene if it was enriched in the same pathway as the congenital neutropenia gene.

To determine whether our method is suitable for predicting congenital neutropenia candidate genes, we calculated the biological distances (Bi, j) of the predicted candidate genes and compared them with 32 known congenital neutropenia genes. A "functional genomic alignment" (FGA) and phylogenetic cluster analysis were then performed. These steps were performed using the APE package available in R to assess the biological correlation between candidate genes and known genes [152, 201]. Specifically, we first created a biological distance matrix between congenital neutropenia genes and congenital neutropenia candidate genes, and then applied a neighbor-joining algorithm (function nj) to create a phylogenetic fan tree showing a hierarchical clustering of known and congenital neutropenia candidate genes. If the candidate genes were evenly distributed throughout the range of known congenital neutropenia genes, this meant that these candidate genes were closely related to the known genes. If the candidate genes and known genes were separated into two or more groups, it meant the opposite.

Using the R package "limma," we searched for genes with differences in expression between patients with congenital neutropenia HC and healthy controls and showed a |log 2|-fold change. Values were taken as cutoff (Threshold) > 0.4 and P-value <0.05 [120]. Data overlap between information on differentially expressed genes obtained from analysis of the GSE142347 and GSE6322 datasets was determined using a Venn diagram in the R package [39].

2.2.3 - Sequence evaluation of nonsynonymous single nucleotide substitutions (missense-SNP) of CCBE1, FAT4, ADAMTS3, TCIRG1, ELANE genes and prediction of pathogenicity of substitutions

We used various in silico tools to test the functional evaluation of the listed immune sietm genes with nsSNPs of pathological or benign nature. We used the following tools: SIFT [145], POLYPHEN-2 [70] PROVEAN [42], FATHMM [69],

LRT [215], M-CAP [125], VEST3 [59], CAAD [52], MetaLR [176], Mutation Assessor [105], MutationTaster [135, 141], and FATHMM-MKL [22], SNP&GO, PhD-SNP [77], PANTHER [150], SNAP2 [82]. All of these tools were available through VarCard [212] and MutPred [99].

SIFT (Sorting Intolerant From Tolerant) is a bioinformatics algorithm used to predict the possible effect of amino acid substitutions on protein function. The algorithm works by comparing an amino acid at a given position in a protein sequence with a set of related protein sequences and estimating how much of the amino acid is conserved in different species.

The SIFT algorithm calculates a score for each amino acid substitution ranging from 0 to 1. A score of 0 means that the substitution is highly likely to be harmful, while a score of 1 means that the substitution is likely benign. To sort gene variants into pathological and benign, the threshold value in SIFT was set at >0.5 (Figure 7).

SIFT results (dbSNP)

Processing... If your browser times out before results are shown, html results can be seen at https://sift.bii.a-star.edu.sg/www/sift/tmp/58b8f27d5f_dbSNP.html and tsv results at https://sift.bii.a-star.edu.sg/www/sift/tmp/58b8f27d5f_dbSNP.tsv Both files are stored for 24 hours before being deleted.

Done.

SNP	ORGANISM/BUILD	CHR	COORDINATE	REF ALLELE	ALT ALLELE	AMINO ACID CHANGE	GENE NAME	GENE ID	TRANSCRIPT ID	PROTEIN ID	REGION	SIFT SCORE	SIFT MEDIAN	N SE POS
rs2288598	Homo_sapiens/GRCh37.74	18	57363917	G	A	I52I	CCBE1	ENSG00000183287	ENST00000439986	ENSP00000404464	CDS	1	3.71	13
rs61745250	Homo_sapiens/GRCh37.74	18	57106987	G	A	P279P	CCBE1	ENSG00000183287	ENST00000439986	ENSP00000404464	CDS	1	3.32	385
rs80008675	Homo_sapiens/GRCh37.74	18	57364452	G	T	D41E	CCBE1	ENSG00000183287	ENST00000439986	ENSP00000404464	CDS	0.016	4.25	11
rs116596858	Homo_sapiens/GRCh37.74	18	57133983	G	A	P181S	CCBE1	ENSG00000183287	ENST00000439986	ENSP00000404464	CDS	0.007	3.58	22
rs116596858	Homo_sapiens/GRCh37.74	18	57133983	G	A		CCBE1	ENSG00000183287	ENST00000398179	ENSP00000381241	UTR_5			
rs116596858	Homo_sapiens/GRCh37.74	18	57133983	G	A		CCBE1	ENSG00000183287	ENST00000589419	ENSP00000467710	UTR_5			
rs116675104	Homo_sapiens/GRCh37.74	18	57134025	G	A	R167W	CCBE1	ENSG00000183287	ENST00000439986	ENSP00000404464	CDS	0.017	3.52	25
rs116675104	Homo_sapiens/GRCh37.74	18	57134025	G	A		CCBE1	ENSG00000183287	ENST00000398179	ENSP00000381241	UTR_5			
rs116675104	Homo_sapiens/GRCh37.74	18	57134025	G	A		CCBE1	ENSG00000183287	ENST00000589419	ENSP00000467710	UTR_5			

Figure 7 - An example of presenting the results of SNP pathogenicity analysis in SIFT

PolyPhen-2 (Polymorphism Phenotyping v2) is a bioinformatics tool used to predict the possible functional impact of an amino acid substitution in a protein. The algorithm analyzes the amino acid sequence of the protein, the position of the variant, and the properties of the amino acids involved in the substitution to predict whether the substitution will be damaging or benign.

PolyPhen-2 uses a combination of evolutionary conservation and structural information to make predictions. First, the algorithm aligns the amino acid sequence of the protein with those of other related species to determine which amino acids are highly conserved and therefore may be functionally important. The algorithm then uses a number of structural characteristics, including solvent availability and the presence of hydrogen bonds, to predict the effect of amino acid substitution on protein structure and function.

The output of PolyPhen-2 is a prediction of the functional impact of the amino acid substitution, which is expressed as a score from 0 to 1. Variants with a score of more than 0.5 are considered harmful, and variants with a score of less than 0.5 are considered benign. PolyPhen-2 has shown high accuracy in predicting the effect of amino acid substitutions, which makes it a useful tool for researchers studying the effects of genetic variations on protein function [70, 92] (Figure 8).

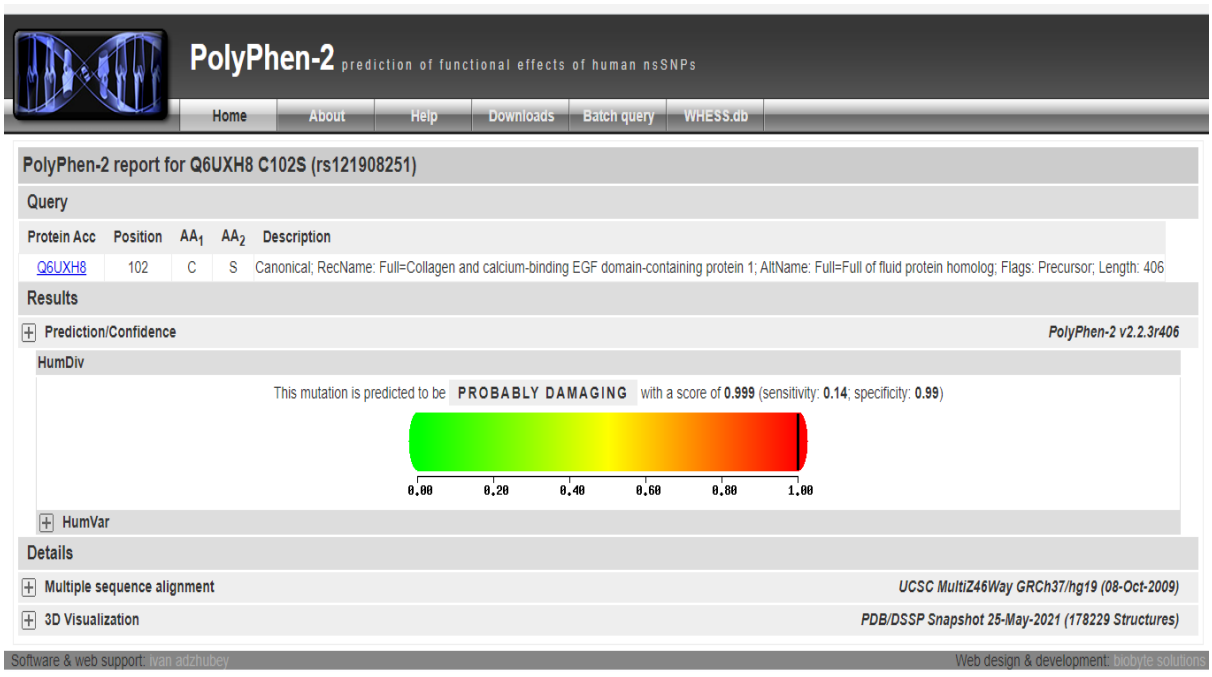


Figure 8 - Example of a presentation of the results of SNP pathogenicity analysis in PolyPhen-2

VarCards is a bioinformatics tool used to analyze genetic variants and predict their potential impact on human health. It combines data from a variety of sources, including public databases, the literature, and experimental data, to provide

comprehensive information about genetic variants and their potential clinical significance.

The VarCards algorithm works in several steps:

1) Collecting variant data: First, the algorithm collects information about genetic variants from various sources, including public databases such as dbSNP and ClinVar, as well as literature and experimental data.

2) Variant annotation: The algorithm annotates variants with information about their genomic location, functional impact and frequency in the population. This information is taken from various sources, including Ensembl, dbNSFP, and ExAC.

3) Prediction of pathogenicity: The algorithm uses various computational tools, such as SIFT and PolyPhen-2, to predict the potential effect of each variant on protein function and estimate its effect on changes in protein function.

4) Association with diseases: The algorithm also integrates information about the association of each variant with human diseases from various sources, including the Human Gene Mutation Database (HGMD), ClinVar, and PubMed.

5) Clinical interpretation: Finally, the algorithm provides clinical interpretation of each variant, including its potential pathogenicity, association with diseases and relevance to specific clinical conditions.

VarCards provides a user-friendly interface for querying and analyzing genetic variants, as well as a customizable pipeline for integrating additional data sources and analysis tools. It is widely used in clinical and research settings to analyze genetic variants and identify potential disease-causing mutations.

We have used VarCARD for the results of tools such as: LRT, Mutation Taster, Mutation Accessor, PROVEAN, FATHMM, VEST3, MTA SVM, METALR, M-CAP, CADD, DANN, FATHMM-MKK, PhD-SNP, PANTHER, SNP-GO, P-MUT [212] (Figure 9).

The threshold values for the aforementioned tools were as follows: Mutation Taster: <0.5; CADD: >15; MetaLR: >0.5; M-Cap: >0.025; PANTHER: probably damaging at time > 450my, possibly damaging (less likely) at 450my > time >

200my, likely benign at time < 200my; VEST3: >0.5; LRT: >0.001; PROVEAN: >-2.667; FATHMM-MKK: <0.5; PhDSNP: >0.5; SNP-GO: >0.5; SNAP2: scale from -100 (completely neutral) to +100 (strong effect); DANN: >0.5; Mutation Assessor: >0.65 (from -5.545 to 5.975, with higher values indicating greater damaging effects); FATHMM: >0.453; PON-P2: >0.5.

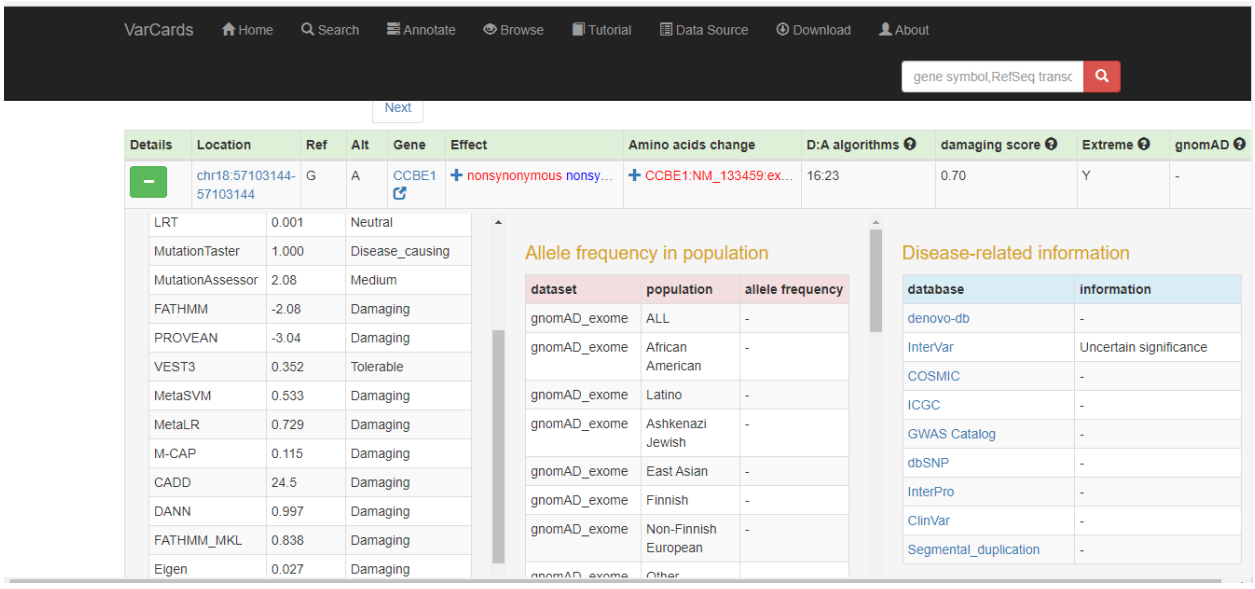


Figure 9 - An example of presenting the results of SNP pathogenicity analysis in VarCards

The online tool MutPred (http://mutpred.mutdb.org/) is used as a search tool for predicting the molecular basis of disease associated with amino acid substitution in a mutant protein. It employs several attributes related to the structure, function, and evolution of the protein. MutPred uses three other services - PSI-BLAST, SIFT, and Pfam - as well as algorithms TMHMM, MARCOIL, and DisProt. This allows for the prediction of most structural damage and achieves even greater prediction accuracy by combining the ratings of all three services [99].

2.2.4 - Assessment of nsSNP effects of CCBE1, FAT4, ADAMTS3, TCIRG1, ELANE genes using in silico tools on protein structure and function

The Mupro method uses support vector machine learning to predict protein stability changes in single-amino acid mutations using both sequence and structural information, as does the IMutant 3.0 method.

iMutant 3.0 is a web server that predicts the effect of single point mutations on protein stability and produces an estimate indicating the probability of destabilizing or stabilizing the mutation. The algorithm is based on a support vector method (SVM) trained on a large data set of experimentally characterized mutants to predict the effect of a mutation on protein stability. The SVM model is trained to distinguish between stabilizing and destabilizing mutations based on the extracted features.

iMutant 3.0 has shown high accuracy in predicting the effect of single-point mutations on protein stability, making it a useful tool for researchers studying the effects of genetic variations on protein function. The algorithm can be used for a wide range of applications, including the construction of stable and functional proteins and the detection of disease-causing mutations.

Some methods use sequence conservation of certain amino acids in a sequence family or look for certain features of the protein structure to predict whether the substitution affects the function of the protein. Amino acid substitutions caused by nsSNPs can alter the stability of the native protein, which can lead to effects on the protein and ultimately to disease [40].

Using the met classifier, iStable 2.0, we predicted changes caused by nsSNP missense substitutions on protein stability. The met classifier uses machine learning and investigates whether protein stability increases or decreases. This is due to amino acid substitution, which is based on the prediction of 8 structural (I-Mutant 3.0, CUPSAT, PoPMuSiC, AUTO-MUTE2.0, SDM, DUET, mCSM, MAESTRO and SDM2) and 3 sequential (I-Mu-tant2.0, MUpro and iPTREESTAB) protein stability prediction tools. A 4-letter PDB code or FASTA-formatted protein sequence is used as input, but the structural predictor achieves better performance than the sequential one. The iStable 2.0 can be found on the Web server at http://ncblab.nchu.edu.tw/iStable2. [106]. I-Mutant 3.0 https://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi).

The ΔΔG Mut dataset from Pro Therm was used to pre-train the algorithm. The value of ΔΔG (kcal/mol) can be used to identify a single-site mutation that

depends on the structure or sequence of the protein. A ΔΔG value less than zero indicates that the variant changes the structure or sequence of the protein [35].

Project HOPE was used in the ELANE study to assess the structural consequences of the substitution sought. Project HOPE is a web server (https://www.projecthope.org/) that proposes to estimate the altered protein in the observed 3D structure in interaction with UniProt and the 3DAS prediction algorithm. The protein sequence is used as an input source in Project HOPE and then a structural comparison is made with the wild type.

In addition, the secondary structure of the ELANE protein was evaluated using the SOPMA program [74]. This is a more sophisticated version of the self-optimized prediction method (SOPM), which can predict the secondary structure (helix, turn and twist) of 69.5% of amino acids in a database of 126 non-homologous (less than 25% homologous) protein chains. SOPMA and the neural network approach (PhD) correctly predict 82.2% of residues and 74% of amino acids predicted when used together.

2.2.5 - Assessment of the effect of nsSNPs on posttranslational modification of immune system proteins

The effect of amino acid substitutions at sites affecting posttranslational modification of a protein was assessed to predict changes in its structure and function [4, 143]. Software available online, GPSMSP v3.0 (https://msp. biocuckoo.org/online.php) was used to predict methylation sites.

We used NetPhos 3.177 (https://www.cbs.dtu.dk/services/NetPhos/) [156] and GPS 5.078 (https://gps.biocuckoo.cn/) [76] to predict potential phosphorylation sites. The NetPhos 3.1 service predicts serine, threonine, and tyrosine phosphorylation sites in proteins using ensembles of neural networks. Phosphorylation sites with a score greater than 0.5 are more likely to be phosphorylated [31].

We used GPSMSP 1.0 (https://msp.biocuckoo.org/), BDMPUB (https://www.bdmpub.biocuckoo.org), and UbPred [93] (https://www.ubpred.org)

to evaluate potential methylation and ubiquitination sites. NetOglyc4.0 additionally predicts glycosylation sites using glycosylation [162] (https://www.cbs.dtu.dk/services/ NetOGlyc/). Glycosylation sites with a score greater than 0.5 are more likely to be glycosylated.

2.2.6 - Assessment of the effect of nsSNPs on functionally different regions of immune system proteins

Conservation analysis is a bioinformatics method used to identify functionally important regions in protein structures by analyzing evolutionary conservation in related protein sequences. The method is based on the principle that evolutionarily conserved regions in protein structures are likely to be functionally important, while variant regions are likely to be less important for protein function. A neural network algorithm and corresponding web service Consurf [53] were used for conservation analysis.

The algorithm converts the estimated rate of evolution into a conservation score relative to other related protein sequences, which typically ranges from 1 (high variability) to 9 (high conservation). Conservation scores are then plotted on the protein structure to identify conserved and variant regions. This can be visualized using various tools such as PyMOL or Chimera.

Conservation analysis can be used to identify functionally important regions in protein structures such as active sites, binding sites, and structural domains. It can also be used to study the evolution of protein function and to design experiments to verify the functional importance of certain regions in the protein structure.

Based on the location and functional importance of different regions of the protein, the amino acid sites in a protein can be divided into several categories, including functional, open, buried, and structural residues.

Functional residues are amino acids that contribute directly to the function of the protein, such as active sites, binding sites, or catalytic residues. Functional residues tend to be highly conserved in related proteins and are often located on the surface of the protein where they can interact with other molecules.

Open residues are amino acids that are located on the surface of the protein and are accessible to the environment. Open residues can play a role in protein-protein interactions, ligand binding, and other functions requiring interaction with the external environment.

Buried residues are amino acids that are located in the interior of the protein and are inaccessible to the environment. Buried residues play an important role in maintaining the overall structure and stability of the protein because they participate in the formation of the protein core.

Structural residues are amino acids that are not directly involved in the functioning of the protein, but are important for maintaining its structure and stability. Structural residues include those that form the secondary structure of the protein, such as alpha-helices and beta-sheets, and those that contribute to the overall stability of the protein, such as disulfide bonds.

Classification of amino acids according to these categories can provide insight into the structure and function of the protein, as well as its evolutionary history and potential for engineering or modification.

2.2.7 - Construction of a 3D model of the structure of immune system proteins to identify the influence of amino acid substitutions

The data source to obtain the wild-type (original) protein sequence was the UniProt database (Universal Protein Resource, https://www.uniprot.org/), an online database of protein sequences and functional information about proteins that is freely available. UniProt is a centralized repository of protein sequences, annotations, and other related information that comes from various databases [171].

Prediction of three-dimensional protein models in order to further compare three-dimensional models of wild (original) types and mutant (altered) types of proteins was performed by their 3D modeling (in Phyre2, I-Tasser, HHpred and AlphaFold2 programs), structure overlay, comparison and further by molecular dynamics simulation (MDS). These programs resulted in .pdb files containing the coordinates of atoms in 3D space [154].

At the same time, the programs HHpred and AlphaFold2 allow MDS without the use of third-party applications from Schrodinger, which will be discussed below, and allow to estimate the standard deviation (RMSD) of the distances between the carbon bases of natural and mutant models over time.

The HHPred (Homology Detection and Structure Prediction by Hidden Markov Model Comparison) application is a bioinformatics tool that uses a Hidden Markov Model (HMM) profile comparison to identify homologous sequences and predict protein structure. The HHPred algorithm compares the target sequence with a database of HMMs derived from protein families in the Pfam database to identify homologous sequences and predict protein structure.

HHPred is a widely used tool for protein structure prediction and is highly accurate and successful in identifying homologous sequences and in predicting protein structure. It is particularly informative for the study of proteins that do not have significant sequence similarity with proteins with known structures. HHPred is freely available as a web server and can be used to predict the structure and function of a wide range of proteins.

AlphaFold 2 is a deep neural network-based protein structure prediction software developed by DeepMind's artificial intelligence research group. AlphaFold 2 uses deep learning techniques to predict the 3D structure of proteins with high accuracy, reaching, in some cases, accuracy close to the atomic level. The software has been used by Jumper J. et al., 2021, to predict the structure of many proteins, including those involved in diseases such as COVID-19, and has the potential to accelerate drug discovery and protein development. AlphaFold 2 was released as an open-source tool, making it freely available to researchers worldwide [85].

Phyre2 is a set of tools available online to predict and analyze protein structure, function, and mutations. The main goal of Phyre2 is to provide biologists with a simple and intuitive interface to state-of-the-art protein bioinformatics tools [207].

I-Tasser, the Iterative Thread Assembly Refinement Server, is an integrated platform for automated prediction of protein structure and function based on the

sequence-structure-function paradigm. Starting from the amino acid sequence, I-TASSER first generates three-dimensional (3D) atomic models based on multi-threaded alignment and iterative structural assembly modeling. Protein function is determined by structurally comparing the 3D models to other known proteins. The result of a typical server contains predictions of the full-length secondary and tertiary structure as well as functional annotations on ligand binding sites, enzyme commission numbers and Gene Ontology terms. An estimate of the accuracy of the predictions is provided based on the confidence score of the simulation. This protocol provides new insights and guidelines for the design of server systems for state-of-the-art predictions of protein structure and function. The server is available at http://zhanglab.ccmb.med.umich.edu/I-TASSER [182].

The resulting .pdb files were visualized in PyMOL, Chimera, and the online service Discovert Studio.

Chimera UCSF is a program for interactive visualization and analysis of molecular structures and related data, including density maps, trajectories, and sequence alignments [210]. PyMOL is a cross-platform molecular graphics tool and is widely used for 3D visualization of macromolecules.

The capabilities of PyMOL have been greatly extended by various plug-ins, including macromolecular analysis, homology modeling, protein-ligand docking, pharmacophore modeling, VS and MD modeling. We used the programming languages R and Python to access these programs.

Discovert Studio (https://discover.3ds.com/) is a program for molecular modeling and various ways of 3D visualization of the resulting models.

In the study of ELANE proteins, calculation of differences between models of wild-type and mutant versions of the proteins after creating models in Phyre2 and I-Tasser was performed using Zhanggroup online service (https://zhanggroup.org/TM-score/, University of Michigan Medical School, USA).

Validation of 3D models was performed using PROCHECK and the Ramachandran plot service.

PROCHECK is a program used to validate the three-dimensional structures of proteins. It was developed by Roman Laskowski at the European Bioinformatics Institute (EBI) and is now widely used in structural biology.

PROCHECK analyzes protein structures in terms of their geometry, including bond lengths, bond angles and torsion angles, and compares them with ideal values for well-functioning structures. The program generates a series of graphical results that summarize the quality of the structure and highlight any areas that may be problematic [170].

Ramachandran plots serve as an indirect tool to check the stereochemistry and geometry of the complex by establishing that none of the geometries are in the forbidden electrostatically unfavorable regions of the plot [170, 173]. This online service applying this method was used in the work: https://swift.cmbi.umcn.nl/servers/html/ramaplot.html (Netherlands).

A similar method that complements the simulation results is MolProbity. It is a web-based all-atom structure validation application for macromolecular crystallography that integrates validation programs from the Richardson lab at Duke University designed to assess the quality of three-dimensional protein structures.

One of the main features of MolProbity is the Ramachandran graph analysis, which examines the torsion angles of the main chain of the protein structure and compares them to the expected values for a properly coiled protein. The program also assesses the quality of the protein geometry, including bond lengths, angles and non-bonding interactions, and identifies potential collisions or steric overlaps. In addition, MolProbity includes tools to assess the consistency of a protein's structure with experimental data, such as electron density maps or nuclear magnetic resonance data. The program also provides recommendations for optimizing the hydrogen bond network in the protein structure and identifying potential errors in the placement of ligands or other non-protein molecules.

In a simulated protein molecule, MolProbity identifies a favorable region, a resolved region, and an outlier region, which correspond to different regions on the

Ramachandran graph, which is a graphical representation of the torsion angles of the main part of the protein structure.

The favorable region corresponds to the area of the diagram where most high-quality protein structures are located. In this region, the torsion angles of the main part are close to the ideal values for a well coiled protein, indicating a well-functioning and stable structure.

The tolerable region is adjacent to the favorable region and represents an area in which the base torsion angles are slightly less than ideal, but still acceptable. Protein structures with torsional angles within this region are considered to be of sufficient quality, although they may have some minor structural problems.

The outlier region is the area of the graph where the torsion angles of the main part differ significantly from the ideal values, indicating a potentially unstable or poorly folded protein structure. Protein structures with torsions in this region are considered low quality and may require significant structural refinement or correction.

Comparison of 3D models of wild-type and mutant variants of proteins was performed taking into account the model comparison metric (TM-score).

TM-score (Template Modeling score) is a widely used metric for comparing structural similarity between two protein structures. It is a measure of structural similarity between two protein structures, taking into account both the standard deviation (RMSD) of aligned residues and the length of the aligned region.

TM-score ranges from 0 to 1, with higher values indicating greater structural similarity between the two proteins. A TM-score score of 1 indicates complete structural similarity between the two proteins, while a TM-score score of 0 indicates no structural similarity.

2.2.8 - Docking methods to study the effect of substitutions on the function of immune system proteins analyzed

Protein docking analysis is the simulation of molecular interactions between two proteins to determine which specific atoms of one protein bind to atoms of the

other protein in three dimensions. This analysis can help understand how two proteins can bind and which specific atoms are involved in this process.

The Discovery Studio and PyMol programs described earlier were used for this purpose. Interactions between atoms at specific amino acid residues were calculated to identify binding forces that were crucial in stabilizing the formation of receptor-ligand complexes.

In addition, AutoDock (Scripps Research Institute) was used in the study of the ELANE protein. AutoDock is designed to perform both rigid and flexible docking simulations. In rigid docking, the protein remains stationary and only the ligand can move during the simulation. In flexible docking, both the protein and ligand can move during the simulation. This flexibility allows AutoDock to simulate conformational changes in the protein that may occur during ligand binding [24].

2.2.9 - Molecular dynamics simulation to assess the pathogenicity of newly identified nsSNPs

Molecular dynamics simulation (MDS) is a computer simulation technique for analyzing the physical motion of atoms and molecules. Atoms and molecules are allowed to interact for a fixed period of time, giving insight into the dynamic "evolution" of the system. In the most common version, the trajectories of atoms and molecules are determined by numerically solving Newton's equations of motion for a system of interacting particles, with the forces between particles and their potential energies often calculated using the interatomic potentials or force fields of molecular mechanics.

A particularly important application of molecular dynamics simulation is to determine how a biomolecular system will respond to some perturbation. In each of these cases, it is usually necessary to run several simulations of both the perturbed and unperturbed system in order to identify consistent differences in the results.

Molecular dynamics simulations were performed using the packages Maestro and Gromacs 4.5.3 from Schrödingern LLC [78].

Maestro creates the preparation for the simulation, in particular, it adds hydrogen atoms to the virtual environment, assigns hydrogen bonds, and minimizes the energies of the molecule. In addition, a "dissolution" of the molecule is performed.

Wild-type and mutant proteins were pretreated using Protein Preparation Wizard in Maestro, which included optimization and complex minimization. A tool available in the Maestro software package that is designed to preprocess protein structures before performing molecular dynamics simulations. It automatically optimizes the geometry of the protein structure, adds hydrogenic atoms, corrects missing or incorrect atoms, removes water and ligands, creates an extended vacuum layer around the protein, and more. All of these steps help eliminate possible problems with the protein structure and prepare it for molecular dynamic simulations. All systems were prepared using the System Builder tool. TIP3P, a solvent model with an orthorhombic cell, was chosen. (Transferable Intermolecular Interaction Potential 3 Points). The OPLS 2005 force field [196] was used in the simulations. To make the models neutral, counter ions were introduced. To simulate physiological conditions, 0.15 M sodium chloride (NaCl) was added. An NPT ensemble with a temperature of 300 K and a pressure of 1 atm was chosen for the entire simulation. The models were "relaxed" prior to simulation. The trajectories were stored for study every 100 ps, and the stability of the simulations was checked by comparing the standard deviation mean square (RMSD) of the protein and ligand over time.

Gromacs produces the following simulation results:

1. Root Mean Square Deviation (RMSD) - a measure of structural deviation over time compared to the structure at T=0 ns. RMSD is calculated by measuring the average distance between atoms of two protein structures after aligning them with each other. Alignment is usually performed by comparing the positions of atoms in the backbone of the two structures. The value of RMSD reflects the degree of deviation between the two structures, with smaller RMSD values indicating greater similarity or coincidence.

2. Root Mean Square Fluctuation (RMSF) - the mean square fluctuation is a measure of the degree of mobility or flexibility of each atom or residue in the protein structure. RMSF is calculated by taking the root mean square deviation of each atom or residue in the protein structure from its mean position during a given simulation or trajectory. The obtained RMSF values are a measure of the variability or fluctuation of the position of each atom or residue, which can indicate the degree of its mobility or flexibility.

3. Differences in the secondary structure of the protein.

4. Radius of gyration (Rg) - a measure of the "expansion" of the protein. The radius of gyration is calculated as the root mean square distance of all atoms in the protein from the center of mass of the protein. Thus, an assessment is made of the overall shape and compactness of the protein. Rg is influenced by various factors, such as the size, shape, and flexibility of the protein. For example, a more compact protein will have a smaller Rg value, while a more elongated or flexible protein will have a larger Rg value. The Rg value can be used to monitor the stability and folding of the protein over time during molecular dynamics simulations.

5. The number of hydrogen bonds formed between different groups of atoms during molecular dynamics simulation. The most commonly used tool for calculating hydrogen bonds in GROMACS is the "g_hbond" command, which identifies hydrogen bonds between donor and acceptor groups of atoms based on geometric criteria. In particular, the tool calculates distance and angle criteria for each potential hydrogen bond and reports the number of hydrogen bonds that satisfy these criteria.

6. The Solvent Accessible Surface Area (SASA) is a measure of the surface area of a protein or other biomolecule that is accessible to the surrounding solvent. It is commonly used in molecular dynamics simulations to analyze the conformational properties of proteins and their interactions with solvents. In GROMACS, SASA is calculated as the surface area of the protein or other biomolecule that is accessible to a probe sphere with a specified radius, typically a water molecule. The calculation involves dividing the surface of the biomolecule

into a grid of small triangles or squares and computing the area of each grid element that is accessible to the solvent.

7. Principal Component Analysis (PCA) is a statistical method used in molecular dynamics simulations to analyze the motion and conformational changes of proteins and other biomolecules. PCA analysis is performed on trajectory files obtained during molecular dynamics simulations. The first step in PCA analysis involves constructing a covariance matrix from the atomic coordinates of the protein or other biomolecule at each time step of the simulation. The covariance matrix is then diagonalized to obtain a set of eigenvectors and eigenvalues that describe the collective motions of the system. The results of PCA analysis can be used to determine the most important collective motions of the protein or other biomolecule, such as domain movements, loop bending or loop fluctuations. These motions can provide insight into the functional properties of the protein, such as enzyme catalysis, ligand binding or protein-protein interactions.

8. The Free Energy Landscape is a graphical representation of the free energy of a system as a function of one or more collective variables, which are typically chosen to describe important degrees of freedom of the system. It is a powerful tool used in molecular dynamics simulations to study the thermodynamics and kinetics of complex systems, such as protein folding, ligand binding, or conformational changes. In GROMACS, free energy landscapes are often constructed using the umbrella sampling method, which involves applying an external biasing potential to constrain the system along the chosen collective variable. Several simulations are then performed, each with a different value of the biasing potential, to sample the entire range of the chosen collective variable.

The free energy of a system as a function of a chosen collective variable can be obtained from the probability distribution of the collective variable, which is estimated from data obtained from umbrella sampling simulations. This probability distribution can be further analyzed using methods such as weighted histogram analysis (WHAM) to obtain a landscape of free energy. The free energy landscape can provide valuable information about the thermodynamics and kinetics of the

system under investigation. For example, it can show stable and metastable states of the system, barriers for conformational changes or ligand binding, as well as protein folding or aggregation mechanisms.

The structure of the native and mutant Aurora-A kinase was used as a starting point for the simulation of molecular dynamics in Maestro. Simulation parameters were set according to our previous work performed for the Aurora-A protein and other proteins. The systems were solvated (dissolved) in a rectangular box with TIP3P water molecules with an edge radius of 10 Å. The systems were neutralized by adding 3 sodium ions (Na+) to the simulation field using the "genion" tool that accompanies the Gromacs package. Energy minimization was performed over 5000 iterations using the conjugate gradient method using the GROMOS96 43a1 force field. The Emtol convergence criterion, which serves as a measure of the stability of the molecular dynamics, was set to 1000 kJ/mol/nm. The Berendsen temperature coupling method was applied to regulate the temperature inside the simulation box. This method ensures that the system temperature remains constant during the simulation by adjusting the temperature of the box based on the instantaneous temperature of the system.

Electrostatic interactions were calculated using the Ewald method with a particle grid. The systems were simulated with position constraint for 5 ns and then simulated without constraint for 200 ns. A comparative analysis of structural deviations in the native and mutant structure was then performed. RMSD, RMSF, SAS, and Rg were analyzed using the tools g_rms, g_rmsf, g_sas, and g_gyrate, respectively. The number of individual hydrogen bonds (NHbonds) was calculated using g_hbond.

In addition, we used g_densmap to obtain the atomic density distribution of the native and mutant protein. All graphs were plotted using the Grace GUI toolbox version 5.1.22. Next, we performed principal component analysis using the Essential Dynamics (ED) method according to the protocol in the Gromacs software package. This section is an abbreviated version of our previously published work.

We used 100 nanoseconds, Desmond, software from Schrödingern LLC, was used to simulate molecular dynamics. Integrating Newton's classical equation of motion, MD simulations typically calculate the motion of atoms over time. Simulations have been used to predict protein stability in a physiological environment [86, 132, 185].

2.2.10 - Identification of possible genetic causes of disease in patients with clinical diagnoses of "congenital neutropenia" and "Hennekam syndrome"

Sequencing results were aligned to the standard human genome sequence hg38 using the Burrows-Wheeler Aligner (BWA) program [119]. The SAM files were then sorted, indexed and converted to BAM format using the SAMtools program [203]. Single nucleotide variants (SNVs) and insertion/deletion (indel) variants were identified using the Genome Analysis Toolkit version 4.1.2.0 (GATK4, http://www.broadinstitute.org/gatk/) [199]. Only exonic variants with a read depth (or coverage) >10× and a minimum mapping quality score of 30 were retained using the VCFtools program to reduce the number of false calls due to mapping errors [64, 204].

All synonymous SNVs, indels without a shift in frame coordinates, and variants with an exonic function annotated as "not applicable" or "unknown" were discarded. Candidate SNVs and indels obtained from the previous steps were further filtered for the presence of SNVs and indels in genes associated with primary immunodeficiency (PID). Candidate SNVs (or indels) were then classified as less common, rare or uncommon if the minor allele frequency (MAF) of the SNV (or indel) was less than 0.01 in all data from the Exome Aggregation Consortium (ExAC), 1000 Genomes (1000g), and the Genome Aggregation Database (gnomAD). All SNVs (or indels) were considered pathogenic if they were nominated as deleterious in at least one model from the following.

Functional analysis was performed using hidden Markov models (FATHMM), protein variation effects analyzer (PROVEAN), and combined annotation-dependent depletion (CADD). FATHMM and PROVEAN were performed using the ANNOVAR program [206], and CADD was performed using

an online server (https://cadd.gs.washington.edu/snv, version: GRCh38-v1.5). All potentially pathogenic SNVs and indels were manually reviewed using the Single Nucleotide Polymorphism Database (dbSNP) program (https://www.ncbi.nlm.nih.gov/snp/) and Integrative Genome Viewer (IGV) software version 2.4.5. If two MAFs of SNVs (or indels) obtained from ANNOVAR and dbSNP were ambiguous, the MAF obtained from dbSNP was considered true. A candidate mutation was considered true if the mutation identified by GATK4 was confirmed using the Integrative Genomic Viewer (IGV) application.

CHAPTER 3 - EVALUATION OF GENE EXPRESSION DIFFERENCES AND INVESTIGATION OF KEY SIGNALING PATHWAYS IN PATIENTS WITH RBCK1 DEFICIENCY

To approach an understanding of the mechanisms underlying the pathogenesis of this pathology, a comparative analysis of gene expression was conducted between transcription data from patients with RBCK1 deficiency, CINCA/NOMID syndrome, Muckle-Wells syndrome, MVK deficiency, and transcription data from healthy children (Figure 10). Genes with differential expression obtained from the analysis were annotated and functionally enriched, meaning information was obtained about their role in organism functioning, the signaling pathways in which these genes are involved, and the conditions under which they are expressed, based on information obtained by other researchers.

From the dataset GSE40561, which includes a total of 48,803 genes from different individuals, 380 genes with differential expression were detected: 229 genes had increased expression, while 151 genes had decreased expression. Comparative analysis of transcription between samples from healthy individuals and patients with RBCK1 deficiency showed the largest number of differentially expressed genes (DEGs) - 119 genes with significantly reduced expression (Table 2). In addition, when comparing RBCK1 and MWS samples, a significant difference in the relatively high expression of 142 genes was identified in RBCK1 deficiency (Figure 10).

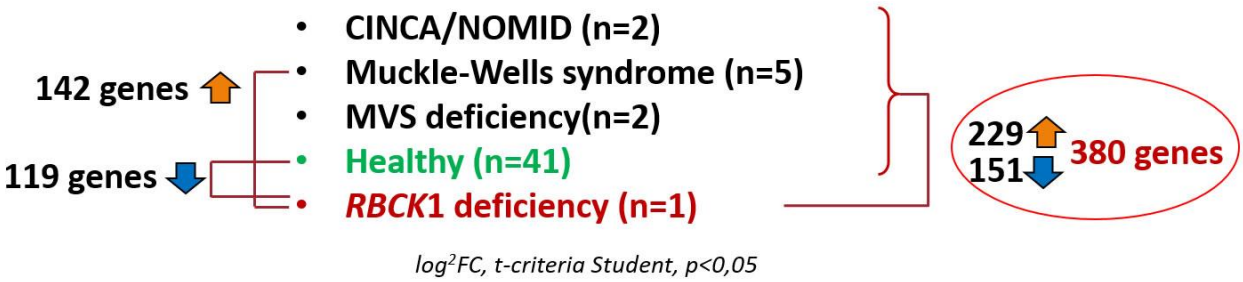


Figure 10 - Significant differences in gene expression in RBCK1 deficiency compared to healthy individuals and patients with other autoimmune syndromes

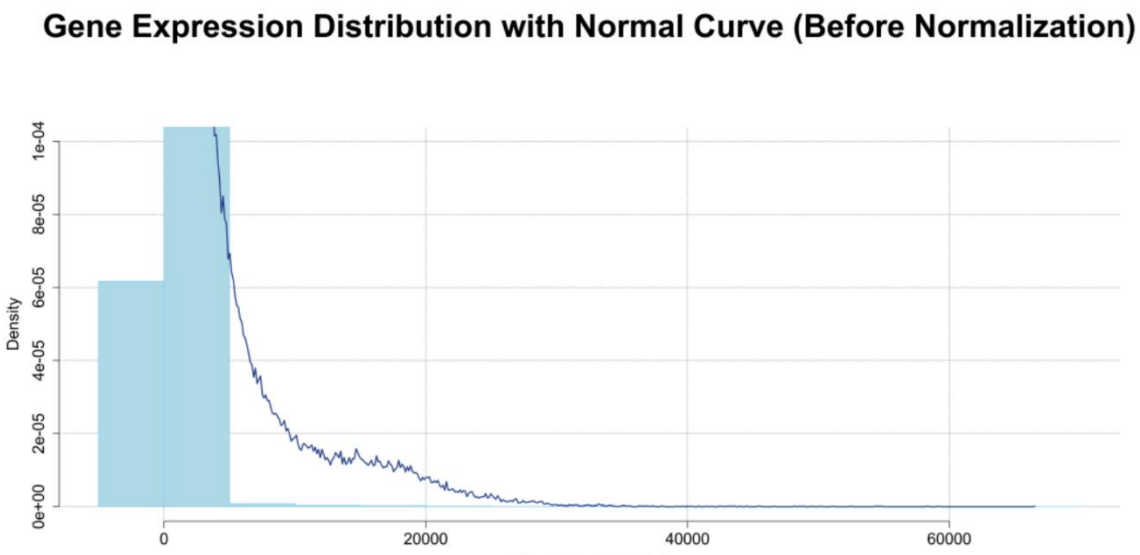
Table 2 - List of the top 10 most down-regulated and up-regulated genes in RBCK1 deficiency compared to healthy samples

Downregulated genes expression				Upregualted genes expression			
№	Genes	p	logFC	№	Genes	p	logFC
1	*CISD2*	7.537936e-18	-0.5969954	1	*HS:551123*	6.458791e-13	3.1687011
2	*EPB41*	4.108343e-16	-0.6256048	2	*HS:552143*	1.583777e-07	1.2937618
3	*LOC253012*	5.739984e-16	-0.6810695	3	*F LJ00312*	3.640115e-07	1.6176159
4	*FAM83A*	1.703811e-13	-0.7337758	4	*HS:19339*	6.034732e-07	0.4643016
5	*NUP98*	5.965986e-13	-0.4364548	5	*ANKMY 2*	6.517125e-06	0.4189648
6	*CHD2*	6.279939e-12	-0.4504437	6	*RPS29*	1.978019e-05	0.4123195
7	*RAP1GAP*	1.338901e-11	-0.9886351	7	*HS:531457*	7.585137e-05	0.4194266
8	*HS:563750*	1.539916e-11	-0.5622475	8	*HS:542923*	0.000385	1.1559277
9	*ABCC13*	1.908913e-11	-0.5476611	9	*HIST1H2BI*	0.000278	0.9167271
10	*MAOA*	3.456357e-11	-0.7114969	10	*PLA2R1*	0.000193	0.8924394

All non-expressed genes were removed, and a new principal component analysis (PCA) diagram was created. The distribution of data before and after normalization can be seen in the histograms (Figure 11) and boxplots (Figure 12). A total of 532 DEGs (genes with altered expression) were obtained in our second dataset GSE31064 after standardization of microarray results, among which 211 genes had decreased expression and 321 genes had increased expression.

The next step was to determine the involvement of differentially expressed genes in RBCK1 deficiency in key signaling pathways and evaluate their impact on biological functions.

A



hysto

B

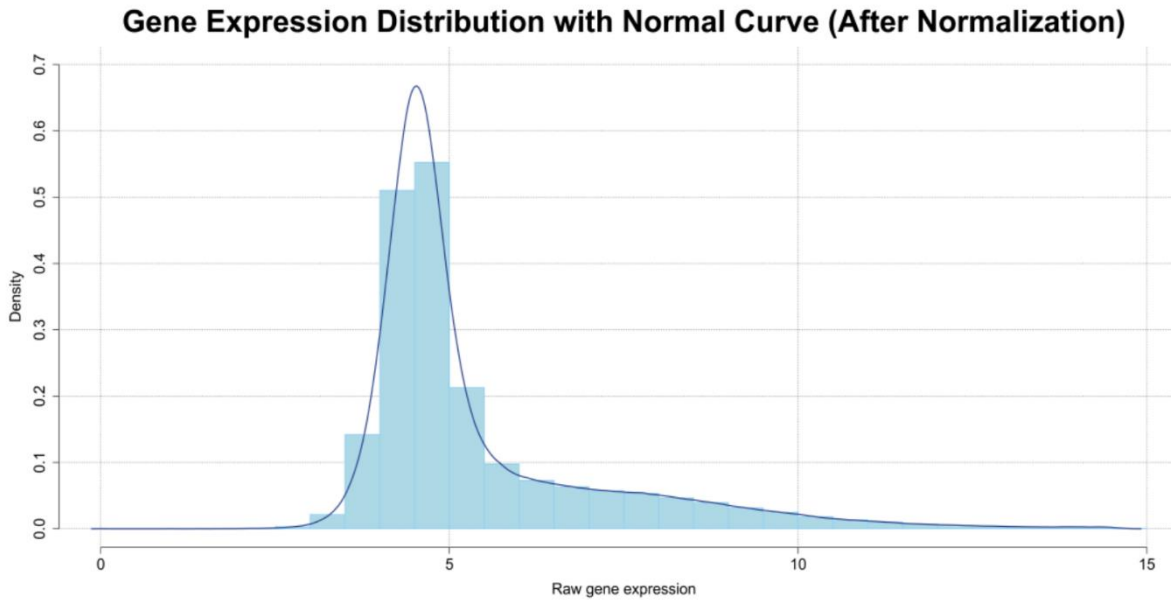
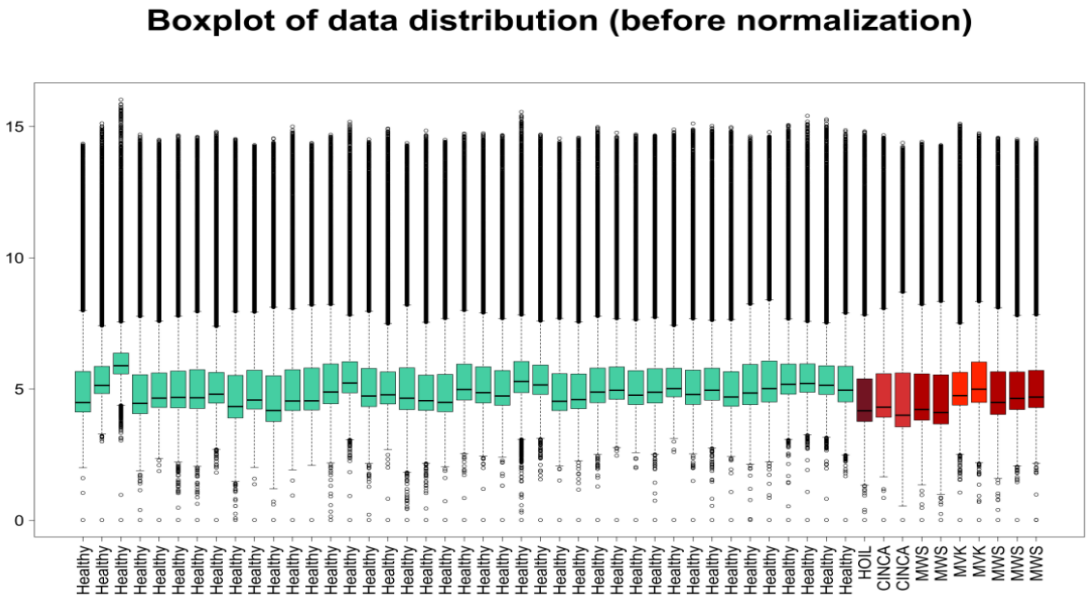


Figure 11 - Histogram of the distribution of unprocessed expression data before normalization (A) and after (B)

Note: The x-axis represents the raw expression values in arbitrary units, while the y-axis represents the number of genes with a certain level of expression.

Immune response, inflammatory response, and protein phosphorylation pathways in the category of biological processes were overrepresented in GO and pathways obtained from co-expressed gene clusters (Table 3, Figure 13).

A



B

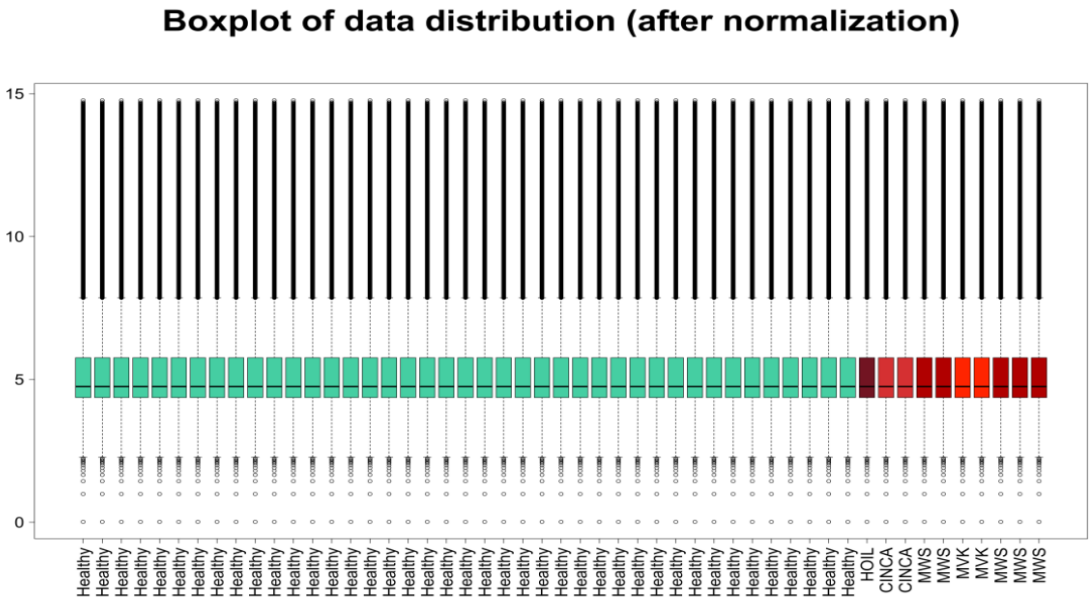


Figure 12 - Boxplots of raw expression data before normalization (A) and after normalization (B)

Note. On the x-axis - transcript samples, on the y-axis - the level of expression.

Protein binding and RNA binding pathways were the most enriched in the category of molecular functions. Finally, cytoplasm, cytosol, and nucleus were the most enriched pathways in the category of cellular components. After uploading identifiers of downregulated genes (with reduced expression) from the comparison between healthy and RBCK1-deficient samples to the WikiPathways database, 425 signaling pathways were selected for analysis. Among those that stood out were signaling pathways associated with the SARS-CoV-2 virus: WP5115, WP5039, WP5098, of which 4 genes from our set - FAM83A, IFI27, NUP98, and TSC1 were identified. Additionally, it was found that the gene HP, which was in the group of downregulated genes in the comparison between healthy and MWS samples, was involved in a pathway related to COVID-19.

Protein binding and RNA binding pathways were the most enriched in the category of molecular functions. Finally, cytoplasm, cytosol, and nucleus were the most enriched pathways in the category of cellular components. After uploading the identifiers of downregulated genes (with decreased expression) from the comparison between healthy samples and RBCK1-deficient samples to the WikiPathways database, 425 pathways were collected. Three most notable pathways were associated with SARS-CoV-2 (COVID-19): WP5115, WP5039, WP5098, among which 4 genes from our set, FAM83A, IFI27, NUP98, and TSC1, were found. In addition, it was found that the HP gene, which was in the downregulated gene group in the comparison between healthy and MWS samples, was involved in a pathway related to COVID-19.

Based on the clusters formed using cemiTool and previously identified DEGs, 30 individual protein-protein interaction graphs were generated. Only one graph, which depicts the interaction between 54 proteins from one of the 14 cemiTool clusters, showed statistical significance (Figure 14).

The subfamilies of lectin-like receptors of killer cells, namely KLRD1, KLRC1, KIR2DL1, KIR2DL2, KIR2DL3, KIR2DL4, KIR3DL2, and KIR3DL3, had the closest interrelation in the protein interaction network. Compared to healthy

individuals and individuals with CINCA syndrome, a decrease in the expression of several genes was detected.

Table 3 - Example of five GO annotations with the smallest false discovery rate coefficient

GO annotation number	Description of GO (gene functions, cellular components, and biological processes)	The proportion of annotated genes	False discovery rate (FDR)
GO:0071799	Cellular response to prostaglandin D stimulation.	2/5	0.0125
GO:0021796	Regionalization of the cerebral cortex.	2/7	0.0162
GO:0030656	Regulation of vitamin metabolism process.	2/12	0.0349
GO:0051712	Positive regulation of killing of cells from another organism.	2/13	0.0376
GO:0001829	Differentiation of trophoectodermal cells.	2/15	0.0456

In addition, chemokine genes (CXCL8 and CXCL10) were particularly highlighted because their activity deficiency can lead to serious errors in cell functioning or cell death due to disruption of chemokine signaling. The p-value for the enrichment of the PPI network created was 1.0e-16 (Figure 14).

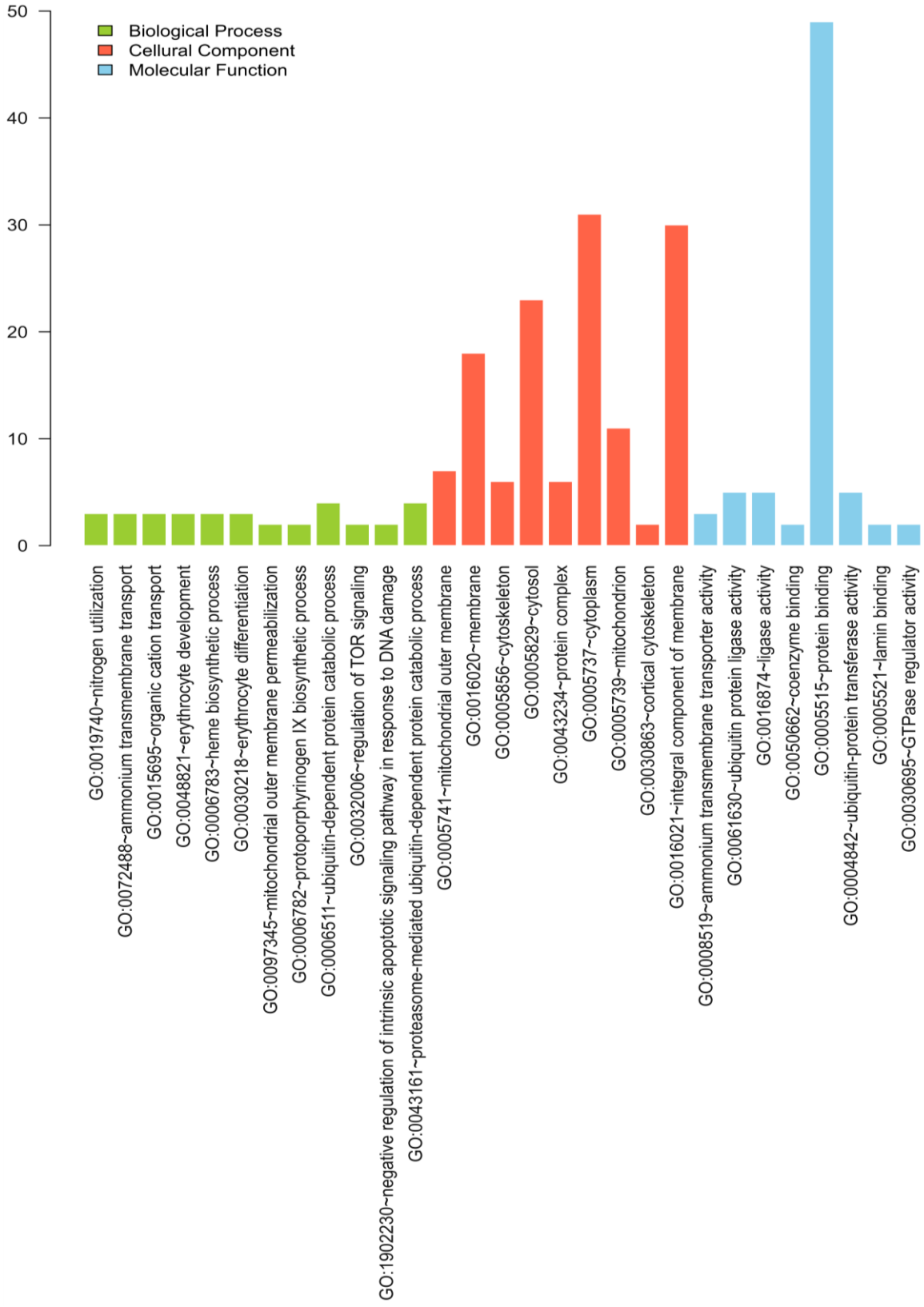


Figure 13 - Histograms of Gene Ontology (GO) representing enriched pathways among genes with decreased expression obtained by comparing healthy individuals and patients with a deficiency RBCK1

Note: The GO terms refer to "molecular function," "cellular component," and "biological process.

In the functional enrichment of differentially expressed genes in RBCK1 deficiency (compared to healthy individuals), involvement of these genes in several significant signaling pathways was detected. In particular, signaling pathways for leishmaniasis development, susceptibility to staphylococcal infection, cholera, NK cell cytotoxicity, and various other pathways affecting the immune response were involved. This does not mean that RBCK1 deficiency increases the probability of the corresponding pathology, but it becomes clearer that the systemic influence of the deficiency of one protein on various processes that somehow affect the immune system and anti-infective defense (Tables 4 and 5).

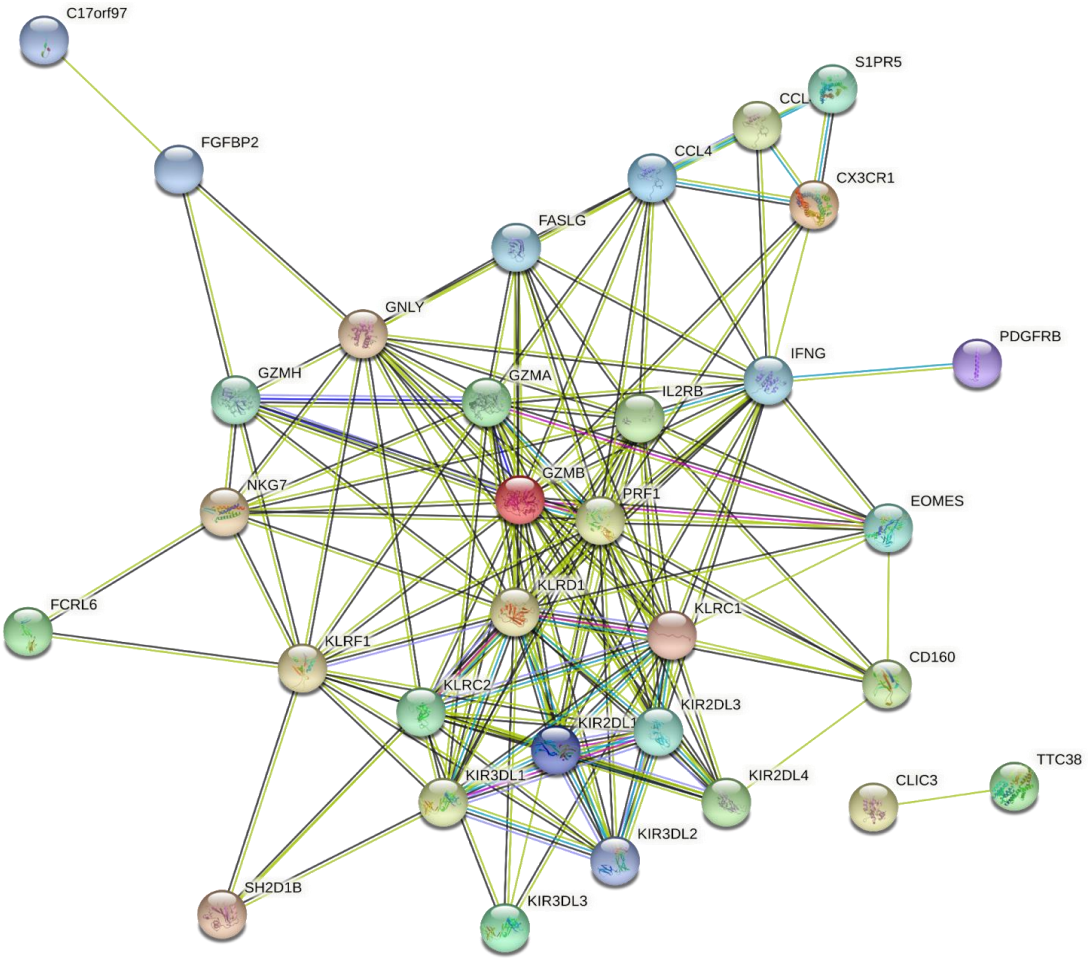


Figure 14 - Statistically significant protein-protein interactions among 54 proteins based on the analysis results of the STRING service

Table 4 - Groups of differentially expressed genes in RBCK1 deficiency (compared to healthy individuals), involved in several significant signaling pathways, obtained by functional enrichment analysis using the KEGG database

Intracellular signaling,	Hematopoietic cell lineage	Dilated cardiomyopathy	Pantothenate and CoA biosynthesis	Vibrio cholerae infection
TIAM1	*MME*	*TPM3*	*VNN1*	*TCIRG1*
FOXO3	*HLA-DRB5*	*ACTG1*	*UPB1*	*ACTG1*
STAT1	*CD8B*	*ITGB1*	*VNN2*	*PRKACB*
GNB4	*IL1R2*	*ITGA4*	*ZNF586*	
GRB2	*CSF3R*	*PRKACB*	*VNN3*	
CXCL5	*CD8A*			
PIK3CG	*IL1B*			
ROCK1				
PIK3R1				
VAV3				

Table 5 - Groups of differentially expressed genes in RBCK1 deficiency (compared to healthy individuals), involved in several significant signaling pathways, obtained by functional enrichment analysis using the KEGG database

NK cell-mediated cytotoxicity.	Processing and presentation of antigen.	Leishmaniasis	Staphylococcus aureus infection	RIG-I-like receptor signaling pathway
KIR2DL3	*KIR2DL3*	*FCGR2A*	*FCGR2A*	*CXCL8*
KIR2DL1	*KIR2DL1*	*PTGS2*	*KRT23*	*MAPK13*
KIR2DS5	*KIR2DS5*	*HLA-DRB5*	*FCGR1A*	*CXCL10*
KIR3DL1	*KIR3DL1*	*FCGR3B*	*FCAR*	*ISG15*
KLRD1	*KLRD1*	*NCF4*	*FPR2*	
KLRC2	*KLRC2*	*TLR2*		
KIR3DL2	*KIR3DL2*	*IL1B*		
KIR2DS3	*KIR2DS3*	*NCF2*		
KIR3DL3	*KIR3DL3*			
KLRC1	*KIR2DL4*			
SH2D1B	*KLRC1*			
GZMB	*IFNG*			
FAS3LG				
PRF1				

Thus, this chapter presents the results of a bioinformatic analysis conducted to obtain information that helps to uncover the pathogenesis of the pathology in patients with RBCK1 deficiency. Our previous studies have shown that 14 new genes involved in ribosome biogenesis, rRNA processing, gene expression, mRNA processing, nuclear lumen, intracellular non-membrane-bound organelles, nucleoplasm, small subunit processomes, antigen processing and presentation pathway, and eukaryotic ribosome biogenesis may play a role in increased susceptibility to viral infections such as influenza or herpes. In this syndrome, these signaling pathways were not involved in the pathological process, which further emphasizes the peculiarity of the RBCK1 deficiency pathogenesis and coincides with observations of increased susceptibility specifically to bacterial infections.

Regarding the antiviral activity of the immune response in RBCK1-deficient patients, according to clinical data, increased susceptibility to them is a rare case. Our study demonstrated the involvement of the signaling pathway responsible for the response to coronavirus infection. However, it was not proven that RBCK1-deficient patients are at risk of COVID-19. This can be explained by the ubiquitination of interferon regulatory factor 3, an important signaling molecule associated with Toll/IL-1R domain-containing adapter inducing IFN and TLR 3.

As a result of the studies, a highly reliable decrease in CISD2 gene expression was detected in this patient with RBCK1 deficiency. It is known that a defect in CISD2 leads to endoplasmic reticulum stress and apoptosis [44], including peripheral blood mononuclear cells. Considering the close functional relationship of this protein with apoptosis and cellular stress processes, it can be assumed that the influence of low expression of this gene on the pathogenesis has a negative effect on the stability of peripheral blood mononuclear cells to apoptosis and cell death.

However, changes in the activity of mTOR, PI3K/AKT, Rho, and Nf-kB signaling pathway genes directly or indirectly affect the expression of genes in the immune system. All the differences in gene expression found do not explain the immediate cause of increased susceptibility to pyogenic infections, but they reveal

some aspects of molecular interactions, allowing us to better understand the pathogenesis of RBCK1 deficiency.

List of work published by the 3rd chapter

1. Is Up-Regulation Gene Expression of the Certain Genes During the Viral Respiratory Tract Infection Would Have Any Influence in Pathogenesis of the SAR-CoV-2 Infection? / K. Shinwari, G. Liu, M. Bolkov, I. Ahmad, M. Daud, I. Tuzankina, V. Chereshnev // Acta Med. Iran. – 2020. – V. 58 (5). – P. 246-248. (Scopus Q4).

2. Identifying Main Genes and Pathways by Using Gene Expression Profiling in Primary Immunodeficiency HOIL-1/RBCK1 Disorder Patients / K. Shinwari, G. Liu, M.A. Bolkov, M. Ullah, I. Tuzankina // Acta Med. Iran. – 2021. –V. 59 (5). – P.265-279 (Scopus, Q4).

3. Gene expression and pathway analysis in patients with inborn error of TLRs and IL-IRs signaling using microarray data / K. Shinwari, G. Liu, M.A. Bolkov, I.A. Tuzankina, V.A. Chereshnev // AIP Conference Proceedings. – 2022. – V. 2390. – 030088 (Scopus).

4. Additional Pathogenic Pathways in RBCK1 Deficiency / E.I. Demicheva, K. Shinwari, K.S. Ushenin, M.A. Bolkov // Mathematical Biology and Bioinformatics. –2022. –V. 17 (2). – P. 174-187. (RSCI).

5. Checking gene expression profile associated with IRF7 and UNC93B deficient patient peripheral blood mononuclear cells infected with pH1N1 influenza virus / K. Shinwari, G. Liu, M.A. Bolkov, I.A. Tuzankina, V.A. Chereshnev // AIP Conference Proceedings. – 2022. – V. 2390. – 030089 (Scopus).

CHAPTER 4 - INVESTIGATION OF THE IMPACT OF IDENTIFIED NON-SYNONYMOUS SINGLE NUCLEOTIDE VARIANTS IN THE ELANE AND TCIRG1 GENES ON THE STRUCTURE AND FUNCTION OF THE ELANE AND TCIRG1 PROTEINS

Congenital neutropenia syndromes are a group of rare diseases that manifest from birth and are characterized by low levels of neutrophils, which are necessary to fight infections. The most common and serious immunodeficiency associated with congenital neutropenia is severe congenital neutropenia, a rare blood disorder that, according to Donadieu J. et al., 2013, affects approximately 1 in 100,000 people of European descent, many cases of which are inherited in an autosomal dominant pattern [65]. Despite several causal genes being identified, the genetic basis of >30% of cases remains unknown.

Approximately half of all cases of severe congenital neutropenia are caused by variants in the ELANE gene. Only a small percentage of cases of this disorder are attributed to other related genes, including TCIRG1.

This study provides data on nsSNPs in the TCIRG1 and ELANE genes obtained from the online NCBI dbSNP database, as well as data on nsSNPs for the TCIRG1 gene from NGS data of one patient, analyzed using bioinformatics methods, including in silico modeling and simulation of molecular dynamics, which allowed for the identification of their potential destabilization of the structure and function of the TCIRG1 and ELANE proteins.

4.1 - Determining the harmfulness of non-synonymous single nucleotide substitutions using SIFT and PolyPhen-2 tools in the TCIRG1 and ELANE genes

The NCBI database reports 5627 SNPs in the TCIRG1 gene. The first step was to select only those polymorphisms that cause amino acid substitutions. It was found that less than 2% of the substitutions, 811 out of 5627, are non-synonymous

coding (missense) substitutions (nsSNPs). In the ELANE gene, only 301 nsSNPs out of 3646 SNPs were identified.

The programs SIFT and PolyPhen-2 calculate the impact of nsSNPs on protein function and evaluate whether the induced amino acid is acceptable at a specific location. SIFT classifies each nsSNP based on scores, and those with scores below a threshold are deemed "tolerated" or benign, while those with scores above the threshold are considered "damaging" or deleterious. For SIFT, the threshold for classification as damaging nsSNP was determined as a score of >0.5.

In the TCIRG1 gene, the SIFT program predicted 118 potentially deleterious nsSNPs, PolyPhen-2 predicted 64, and the mutually intersecting results of the combined analysis allowed only 34 nsSNPs to be selected that resulted in amino acid substitutions out of the total of 811 nsSNPs. Table 6 shows a portion of the obtained analysis results.

For substitutions in the ELANE gene, the SIFT program identified 21 nsSNPs as deleterious polymorphisms, while the combined analysis of SIFT and PolyPhen-2 only indicated 8 nsSNPs as deleterious out of the total of 301 nsSNPs (Table 7).

To confirm the deleteriousness of the polymorphisms selected through SIFT and PolyPhen-2, additional in silico tools were used.

The results of predicting the pathogenicity of significant nsSNPs in the TCIRG1 gene using 17 additional analysis tools are presented in Figure 15 and Table 8.

Table 6 – Deleterious/Damaging non-synonymous single nucleotide polymorphisms (nsSNP) in the TCIRG1 gene based on the results of SIFT and PolyPhen-2 analysis

ID nsSNP	A.A.	SIFT	Score	PolyPhen-2	Score
rs36027301	R56W	Del	0	Pd	0.999
rs368945298	M546V	Del	0	Pd	0.999
rs115854062	P572L	Del	0	Pd	1
rs150260808	I721N	Del	0	Pd	1
rs137853150	G405R	Del	0	Pd	1
rs137853151	R444L	Del	0	Pd	1
rs147580611	F610S	Del	0	Pd	1.00
rs148921764	E722K	Del	0	Pd	1.00
rs140963213	A417T	Del	0.002	Pd	1
rs144775787	A778V	Del	0.46	Pd	0.883
rs145080707	R213W	Low	0.012	Pd	1
rs150648332	R57H	Del	0.001	Pd	1.00
rs150260808	I721N	Del	0	Pd	1
rs201329219	R109W	Del	0.014	Pd	1.00
rs367703865	R191H	Del	0.32	Pd	0.999
rs371214361	S532C	Del	0.001	Pd	1.00
rs199914625	S474W	Del	0	Pd	1
rs200851583	G458S	Del	0	Pd	1
rs371658110	G192S	Del	0.003	Pd	1.00
rs370319355	R50C	Del	0	Pd	1
rs376351835	F529L	Del	0.013	Pd	1.00
rs371004297	G379S	Del	0.011	Pd	1.00
rs200209146	N730S	Del	0.022	Pd	1.00
rs200415611	V375M	Del	0.001	Pd	1.00

Note: nsSNP ID - identifier of non-synonymous single nucleotide polymorphism, A.A. - position of amino acid, Del - high probability of pathogenicity of mutation; Low - low probability of pathogenicity of mutation, Pd - predicted probable pathogenicity of mutation.

Table 7 - Deleterious/Damaging nonsynonymous single nucleotide polymorphisms (nsSNPs) in the ELANE gene based on the analysis results in SIFT and PolyPhen-2

ID nsSNP	A.A	SIFT	Score	PolyPhen-2	Score	Allelic Frequancy
rs201163886	R34W	Del	0.002	Pd	1	
rs28931611	C71R	Del	0	Pd	1	6.076e-06
rs137854449	V101M	Del	0.005	Pd	0.964	
rs137854448	P139L	Del	0	Pd	1	
rs199558534	R143C	Del	0.048	Pd	1	
rs57246956	C151Y	Del	0	Pd	1	
rs201788817	A166T	Del	0.33	Pd	0.976	
rs199891906	A166V	Del	0.23	Pd	0.582	
rs193141883	T175M	Del	0.008	Pd	1	gnomAD_exome 0.0005
rs200449787	R182H	Del	0.015	Pd	1	
rs367663236	V190M	Del	0.047	Pd	1	
rs201723157	R193W	Del	0.006	Pd	1	
rs201139487	G203S	Del	0	Pd	1	4.094e-06
rs137854446	L206F	Del	0	Pd	1	
rs201664319	N209K	Del	0.03	Pd	0.983	
rs140880838	G210R	Del	0.019	Pd	1	
rs137854451	G214R	Del	0.002	Pd	1	
rs200384291	F218L	Del	0.011	Pd	0.998	

Note: nsSNP ID - identifier of non-synonymous single nucleotide polymorphism, A.A. - position of amino acid, Del - high probability of pathogenicity of mutation; Low - low probability of pathogenicity of mutation, Pd - predicted probable pathogenicity of mutation.

All of the listed amino acid substitutions were predicted to be deleterious by the majority of algorithms (FATHMM-MKL, SNP-GO, PHD-SNP, PANTHER,

SNAP2, P-MUT PROVEAN, FATHMM, LRT, M-CAP, CAAD, META SVM, METALR, Mutation Assessor, and Mutation Taster) used in this study. Each of the algorithms used in this study has its unique threshold and evaluation criterion to determine the pathogenicity or tolerability of the substitution.

For the TCIRG1 gene, the combination of SIFT and VEST 3 algorithms identified only 6 nsSNPs (10% of the previously selected ones) as deleterious, while 51 were classified as tolerable. PolyPhen-2, FATHMM, M-CAP, and PANTHER showed the highest percentage of deleterious predictions. When using the SNAP2 method, 41 substitutions were considered deleterious (71%), while 16 predictions had no effect (SNAP2 score of 100). PANTHER was used to predict the impact of 54 (92%) nsSNPs on the TCIRG1 protein, and 48 nsSNPs were likely to have a damaging effect, 6 nsSNPs might have a possibly damaging effect, and 3 nsSNPs were likely to be benign. Specifically, those with a time greater than 450 ms were classified as possibly damaging, those with a time between 450 ms and 200 ms were classified as likely benign, and those with a time less than 200 ms were not classified.

The PROVEAN program, designed to predict the impact of SNPs on protein function, identified 22 (38%) nsSNPs in the TCIRG1 gene as significantly deleterious (with respect to their impact on the structure and function of the protein), while 35 nsSNPs were classified as neutral based on the PROVEAN threshold criteria (> -2.667). Using the threshold (> 0.65, from 5.545 to 5.975), the mutation evaluator classified 24 nsSNPs as deleterious, of which 12 were classified as high, 17 as moderate, 5 as low, and 19 were not detected.

FATHNMM and FATHMM-MKK (<0.5), CADD (>15), DANN (>0.5), Mutation Taster (<0.5), and their respective scores predict more than 75-90% of nsSNPs as deleterious/damaging. P-Mut predicts 45 (75.21%) deleterious, 7 neutral, and data were missing for 5 nsSNPs with the threshold (<0.5). LRT predicts 42 (77%) deleterious nsSNPs with a result (>0.001) and 13 neutral ones. PhD-SNP, SNP-GO, and M-CAP identified 47 (82%), 35 (61%), and 54 (94.73%) nsSNPs as deleterious, respectively. Additionally, MetalR and MTA-SVM identified 10 (17%) and 37 (64%) nsSNPs as deleterious, respectively (Figure 15).

In conclusion, based on the evaluation of substitution positions using PANTHER, PROVEAN score, SIFT score, SNP&GO, FATHMM, LRT, M-CAP, VEST3, CAAD, METALR, Mutation Assessor, Mutation Taster, FATHMM-MKL, PHD-SNP score, and PolyPhen-2, a group of 15 nsSNPs, including P572L, M546V, I721N, F610S, A732T, F51S, A717D, E722K, R57H, R109W, R191H, S532C, G192S, F529L, and H804Q, was found to be significantly deleterious by all modern methods. Only LRT did not confirm the effects of the A717D substitution predicted by other tools. The results obtained using all prediction algorithms were statistically significant and strongly correlated with each other (the p-value for the Student's t-test between the tools was 0.001).

The results of our nsSNP analysis of the ELANE gene showed that 21 nsSNPs were determined to be deleterious using the SIFT algorithm. Of these 21 SIFT-predicted deleterious nsSNPs, 18 were also predicted to be deleterious by the PolyPhen-2 and FATHMM-MKL algorithms. However, other algorithms used in this study did not show 100% agreement (Figure 16).

Among all 21 SIFT-predicted deleterious nsSNPs, the LRT and FATMANH algorithms predicted the fewest matches. Both algorithms predicted only 10 pathogenic nsSNPs for ELANE, while 11 were classified as tolerant, neutral, or of unknown significance.

The PolyPhen-2 platform identified 18 pathogenic nsSNPs; VEST, CADD, and DANN platforms predicted 19; M-Cap and Mutation Taster predicted 20 pathogenic nsSNPs each. Using the SNAP2 approach, 18 damaging mutations were detected, while three had no association with pathology.

For the PANTHER program, 17 nsSNPs were considered as non-synonymous mutations, among which 10 nsSNPs were classified as likely pathogenic, 7 as possibly pathogenic, 2 as likely benign, and 2 as variants of unknown significance. When analyzed using PROVEAN, 14 out of 21 nsSNPs in the ELANE gene were predicted to be strongly deleterious, while 7 were considered neutral.

Mutation Assessor considered 20 nsSNPs to be deleterious, including 3 with high pathogenicity, 6 with medium, and 12 with low, and one with unknown

significance. P-Mut predicted 10 mutations as pathological, 10 with unknown significance, and one without a result. PhD-SNP predicted 13 mutations as pathological, SNP-GO - 10, MetalR - 17, and MTA-SVM - 15.

All modern methods for evaluating the pathogenicity of nsSNPs used together revealed 8 overlapping common mutations in the ELANE gene: C71R, P139L, C151Y, T175M, G203S, G214R, R193W, and F218L (Table 9).

According to the software used, it is known that the allele frequency of C71R in Latin Americans is 3.655e-05, T175M in Africans is 0.0002, in Latin Americans - 0.0023, in East Asians - 5.832e-05, in Europeans - 3.632e-05, and in Latin Americans - 2.979e-05. The results of all prediction algorithms were statistically significant and closely related to each other. The value of the Student's coefficient between the tools has a p-value of 0.001.

Table 8 - Assessment of pathogenicity of identified TCIRG1 substitutions using various prediction tools

Замена	LRT	Mutation Taster	Mutation Accessor	PROVEAN	FATHMM	VEST3	MTA SVM	METALR	M-CAP	CADD	DANN	FATHMM-MKK	PhD-SNP	PANTHER	SNP-GO	P-MUT	SNAP2
P572L	D	D	H	D	D	D	D	D	D	D	D	D	D	D	D	D	D
M546V	D	D	H	D	D	D	D	D	D	D	D	D	D	D	D	D	D
I721N	D	D	H	D	D	D	D	D	D	D	D	D	D	D	D	D	D
F610S	D	D	M	D	D	D	D	D	D	D	D	D	D	D	D	D	D
A732T	D	D	H	D	D	D	D	D	D	D	D	D	D	D	D	D	D
F51S	D	D	M	D	D	D	D	D	D	D	D	D	D	D	D	D	D
A717D	N	D	M	D	D	D	D	D	D	D	D	D	D	D	D	D	D
E722K	D	D	H	D	D	D	D	D	D	D	D	D	D	D	D	D	D
R57H	D	D	H	D	D	D	D	D	D	D	D	D	D	D	D	D	D
R109W	D	D	M	D	D	D	D	D	D	D	D	D	D	D	D	D	D
R191H	D	D	H	D	D	D	D	D	D	D	D	D	D	D	D	D	D
S532C	D	D	H	D	D	D	D	D	D	D	T	D	D	D	D	D	D
G192S	D	D	H	D	D	D	D	D	D	D	D	D	D	D	D	D	D
F529L	D	D	M	D	D	D	D	D	D	D	D	D	D	D	D	D	D
H804Q	D	D	M	D	D	D	D	D	D	D	D	D	D	D	D	D	D
G405R	D	D	-	D	D	D	D	D	D	D	D	D	D	D	D	D	D
S474W	D	D	-	D	D	D	D	D	D	D	D	D	D	D	D	D	D
G458S	D	D	-	D	D	D	D	D	D	D	D	D	D	D	D	D	D
R444L	D	D	-	D	D	D	D	D	D	D	D	D	D	D	D	D	D
R56P	D	D	-	D	D	D	D	D	D	D	D	D	D	D	D	D	D
G379S	D	D	-	D	D	D	D	D	D	D	D	D	D	D	D	D	D
R757C	D	D	M	D	D	D	D	D	D	D	T	D	D	D	D	D	D

Continue table 8

N730S	D	D	M	D	D	T	D	D	D	D	D	D	D	D	D	D
V375M	D	D	-	D	D	D	D	D	D	D	D	D	D	D	D	D
T314M	D	D	-	D	D	D	D	D	D	D	D	D	D	D	D	D
D517N	D	D	H	D	D	T	D	D	D	D	D	D	D	D	D	D
R92W	D	D	M	D	D	T	D	D	D	D	D	D	D	D	D	D
T368M	D	D	-	D	D	D	D	D	D	D	D	D	D	D	D	D
A417T	D	D	H	D	D	D	D	D	D	D	T	D	D	D	D	D
R363C	D	D	-	D	D	D	D	D	D	D	D	D	D	D	D	D
R56W	D	D	H	D	D	T	D	T	-	D	D	D	D	D	D	D
A778V	D	D	M	D	D	D	D	D	D	D	D	D	D	D	N	N	N
R50C	D	D	M	D	D	T	D	D	D	D	D	D	D	D	D	D	-
V52L	D	D	M	T	D	T	D	D	-	D	D	D	D	D	D	D	D

Note. Substitution refers to an amino acid substitution in the molecule; the following columns represent mutation pathogenicity prediction programs. D - damaging substitution, T - tolerated, N - neutral, M - medium probability, L - low, H - high, P - pathogenic, - no data.

Table 9 - Assessment of pathogenicity of identified ELANE substitutions using various prediction tools

Замена	LRT	Mutation Taster	Mutation Accessor	PROVEAN	FATHMM	VEST3	MTA SVM	METALR	M-CAP	CADD	DANN	FATHMM-MKK	PhD-SNP	PANTHER	SNP-GO	P-MUT	SNAP2
C71Y	0.001 -	1 D	4.21 H	-11.22 D	-4.7 D	0.944 D	1.097 D	0.968 D	0.973 D	23.9 D	0.987 D	0.817 D	0.933 D	0.907 D	0.982 D	0.71 D	77 D
P139L	0.002 -	1.00 D	2.72 M	-8.79 D	-3.56 D	0.912 D	0.956 D	0.886 D	0.942 D	26.6 D	0.999 D	0.747 D	0.733 D	0.888 D	0.855 D	0.85 D	59 D
C151Y	0.001 -	1.00 D	2.745 M	-10.41 D	-3.34 D	0.925 D	1.005 D	0.900 D	0.923 D	25.2 D	0.996 D	0.916 D	0.939 D	0.988 D	0.939 D	0.9 D	82 D
T175M	0.002 -	0.940 D	1.98 M	-4.76 D	-2.48 D	0.774 D	0.798 D	0.812 D	0.891 D	33 D	0.999 D	0.685 D	0.242 N	0.833 D	0.719 D	0.7 D	20 D
G203S	0.001 -	0.997 D	3.865 H	-5.4 D	-7.34 D	0.737 D	0.915 D	0.966 D	0.966 D	27 D	0.998 D	0.826 D	0.898 D	0.909 D	0.909 D	0.89 D	77 D
G214R	0.001 -	1.00 D	4.2 H	-6.2 D	-6.13 D	0.965 D	0.997 D	0.989 D	0.949 D	26.7 D	0.999 D	0.934 D	0.924 D	0.98 D	0.904 D	0.9 D	94 D
R193W	0.071 -	1.00 D	1.755 M	-5.69 D	-2.45 D	0.805 D	0.068 D	0.733 D	0.806 D	27 D	0.998 D	0.240 D	0.566 D	0.909 D	0.834 D	0.63 D	55 D
F218L	0.001 D	0.990 D	2.915 M	-4.82 D	-3.25 D	0.766 D	0.898 D	0.861 D	0.807 D	24.7 D	0.998 D	0.904 D	0.769 D	0.531 D	0.899 D	0.84 D	76 D

Note. Substitution refers to an amino acid substitution in the molecule; the following columns represent mutation pathogenicity prediction programs. D - damaging substitution, T - tolerated, N - neutral, M - medium probability, L - low, H - high, P - pathogenic, - no data.

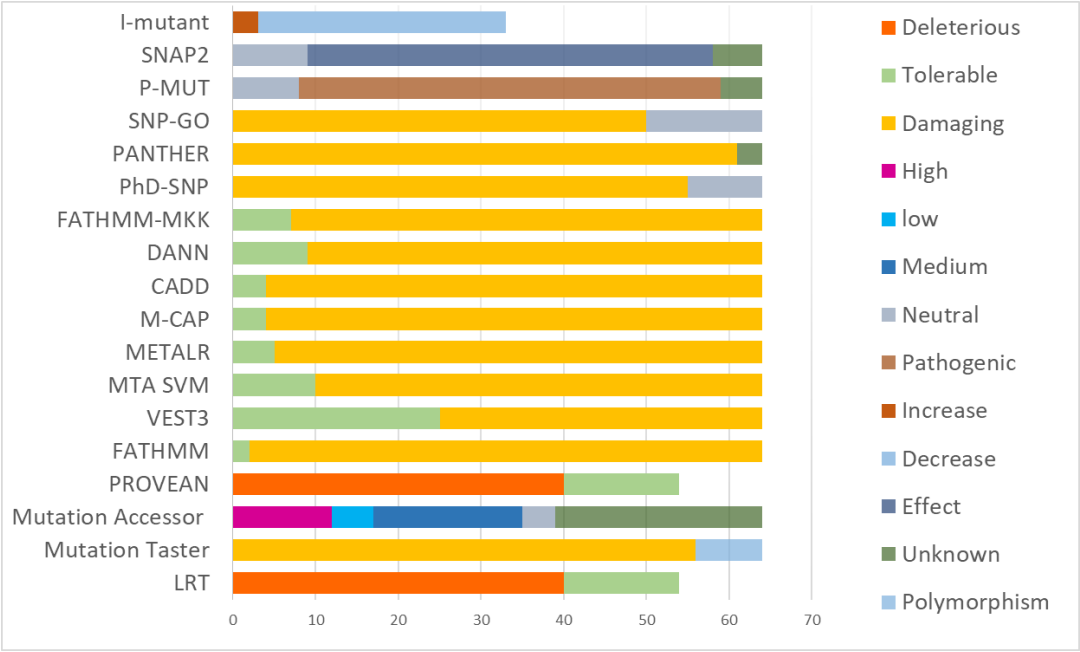


Figure 15 - Results of predicting the impact of 64 nsSNPs in the TCIRG1 gene analyzed by eighteen computational tools

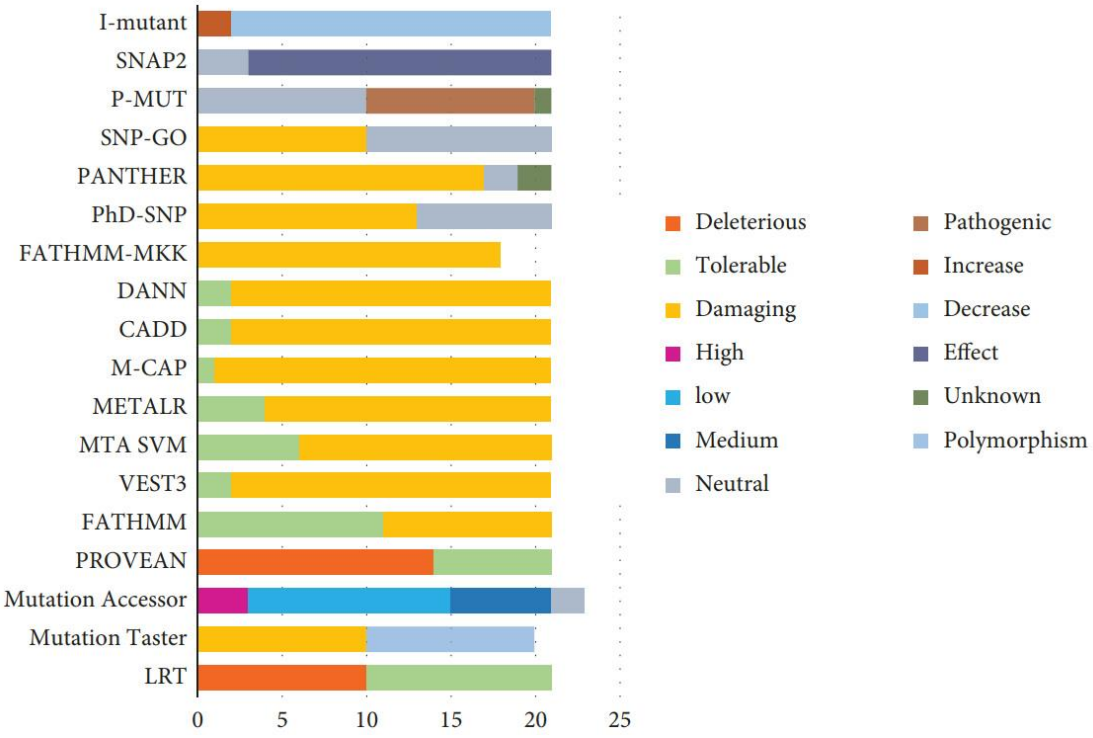


Figure 16 - Results of predicting the impact of 21 nsSNPs in the ELANE gene, analyzed by eighteen computational tools

Thus, following this stage of research, 15 of the most harmful non-synonymous single nucleotide substitutions (and corresponding amino acid

substitutions) were selected, which according to the predictions of the algorithms used, will have the greatest impact on the structure and function of the TCIRG1 protein: rs199902030, rs200149541, rs372499913, rs267605221, rs374941368, rs375717418, rs80008675, rs149792489, rs116675104, rs121908250, rs121908251, rs121908251, rs149792489, rs116675104, rs118141250. One of these substitutions (rs118141250, Val52Leu) was previously identified in a patient from the Sverdlovsk region through whole-genome sequencing.

For the ELANE gene, a total of 8 nsSNPs - rs28931611, rs57246956, rs137854448, rs193141883, rs201723157, rs201139487, rs137854451, and rs20038429 - were selected. These nsSNPs (and corresponding amino acid substitutions) were subsequently analyzed by other methods in order to determine their impact on the 3D structure and function of the proteins.

4.2 - Prediction of nsSNP impact on protein stability using computational tools I-Mutant and MU-pro

The impact of nsSNPs with high pathogenic risk on the stability and function of the TCIRG1 protein was evaluated using the I-Mutant 3.0 web service. The results showed that the amino acid substitutions G405R, S474W, and A778V increase the stability of the protein, while P572L, M546V, I730N, F610S, A732T, F51S, A717D, E722K, R57H, R109W, R191W, S532C, G192S, F529L, H804Q, G458S, R444L, R56P, G379S, R757C, N730S, V375M, T314M, D517N, R92W, T368M, A417T, R363C, R56W, and R50C decrease its calculated stability (Table 10).

The results of the impact of nsSNPs with high pathogenic risk on the stability and function of the ELANE protein showed that the amino acid substitutions V101L and A166V increase the protein stability, while R34W, C71R, V101M, P139L, R143C, C151Y, A166T, T175M, R182H, V190M, R193W, G203S, L206F, N209K, G210R, G214R, F218L, P262S, and P262L R50C decrease its predicted stability. At the same time, the Mu-pro algorithm showed that all the nsSNPs with a high pathogenicity score identified in the previous tests reduce protein stability

Table 10 - Results of analysis of highly deleterious nsSNPs in the TCIRG1 gene using the I-Mutant 3.0 program

A.A.C	"Confidence score	Impact on protein stablity
P572L	-0.35	Decrease
M546V	-0.56	Decrease
I730N	-1.74	Decrease
F610S	-1.43	Decrease
A732T	-0.69	Decrease
F51S	-1.78	Decrease
A717D	-0.51	Decrease
E722K	-0.44	Decrease
R57H	-1.47	Decrease
R109W	-0.06	Decrease
R191W	-0.37	Decrease
S532C	-0.58	Decrease
G192S	-1.00	Decrease
F529L	-0.95	Decrease
H804Q	-0.10	Decrease
G405R	-0.28	Increase
S474W	-0.10	Increase
G458S	-1.26	Decrease
R444L	-0.23	Decrease
R56P	-0.85	Decrease
G379S	-1.41	Decrease
R757C	-1.00	Decrease
N730S	-0.34	Decrease
V375M	-1.06	Decrease
T314M	0.02	Decrease
D517N	-0.98	Decrease
R92W	-0.24	Decrease
T368M	-0.37	Decrease
A417T	-0.78	Decrease
R363C	-1.00	Decrease
R56W	-0.49	Decrease
A778V	-0.15	Increase
R50C	-1.20	Decrease

4.3 - Analysis of the impact of nsSNPs in the TCIRG1 and ELANE genes on protein conserved regions

According to the results of the ConSurf analysis, 22 pathogenic nsSNPs were found in highly conserved regions (7-9 conservation score) of the TCIRG1 protein. Additionally, 16 substitutions - S7K, V52L, G379S, M403I, G405R, G458S, D517N, F529L, S532C, M546V, A640S, D683H, I732N, N730S, A732T, H804Q - were predicted as substitutions in functional and exposed amino acid residues of the protein. Ten substitutions such as A20V, R56P, R57H, R191H, G192C, E321K, R366H, T368M, R444L, and E722K were predicted in the region of functional and exposed residues, while the remaining 16 - S7K, V52L, G379S, M403I, G405R, G458S, D517N, F529L, S532C, M546V, A640S, D683H, I732N, N730S, A732T, and H804Q - were predicted as buried and structural residues. The following 18 substitutions - S3F, R28W, S45A, R50C, R92W, R109W, R166T, T314M, D328M, S340L, R363C, R382H, R467H, S474W, P572L, Y626S, R628W, and R757C - were predicted as substitutions in exposed regions, while the remaining 9 - F51S, V348M, V375M, A417T, T570M, F610S, A717D, A778V, and M783I - were predicted as substitutions in buried amino acids. The results are presented in Figure 17.

According to the ConSurf analysis results for the ELANE protein, 22 dangerous nsSNPs were identified in highly conservative regions of the protein (7-9 on the conservation scale). Among these 22 missense variants, 8 were located in highly conservative positions, 2 - P139L and C71R - were predicted to be functional and exposed residues, and the remaining 3 - G214R, C151Y, and C71Y - were predicted to be buried and structural residues. The following 12 substitutions - R34W, R143C, A166T, A166V, T175M, R182H, V190M, R193W, N209K, G210R, P262S, and P262L - were predicted to be amino acid substitutions in exposed regions of the protein, and F218L, V101L, and V101M - as substitutions in buried residues.

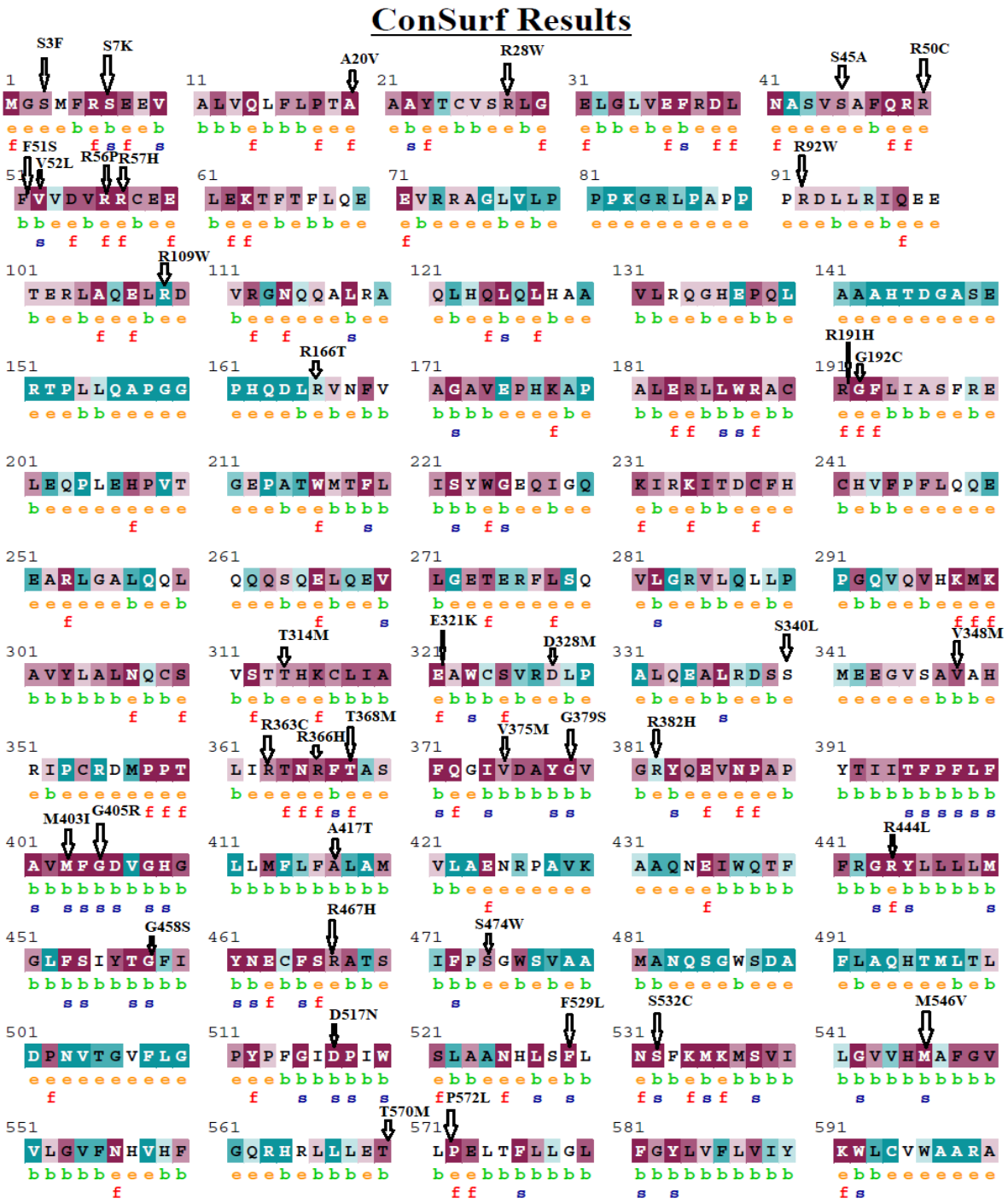


Figure 17 - Location of amino acid substitutions in the TCIRG1 protein considering evolutionary conservation and the location of different regions of the protein according to the ConSurf analysis

Continue figure 17.

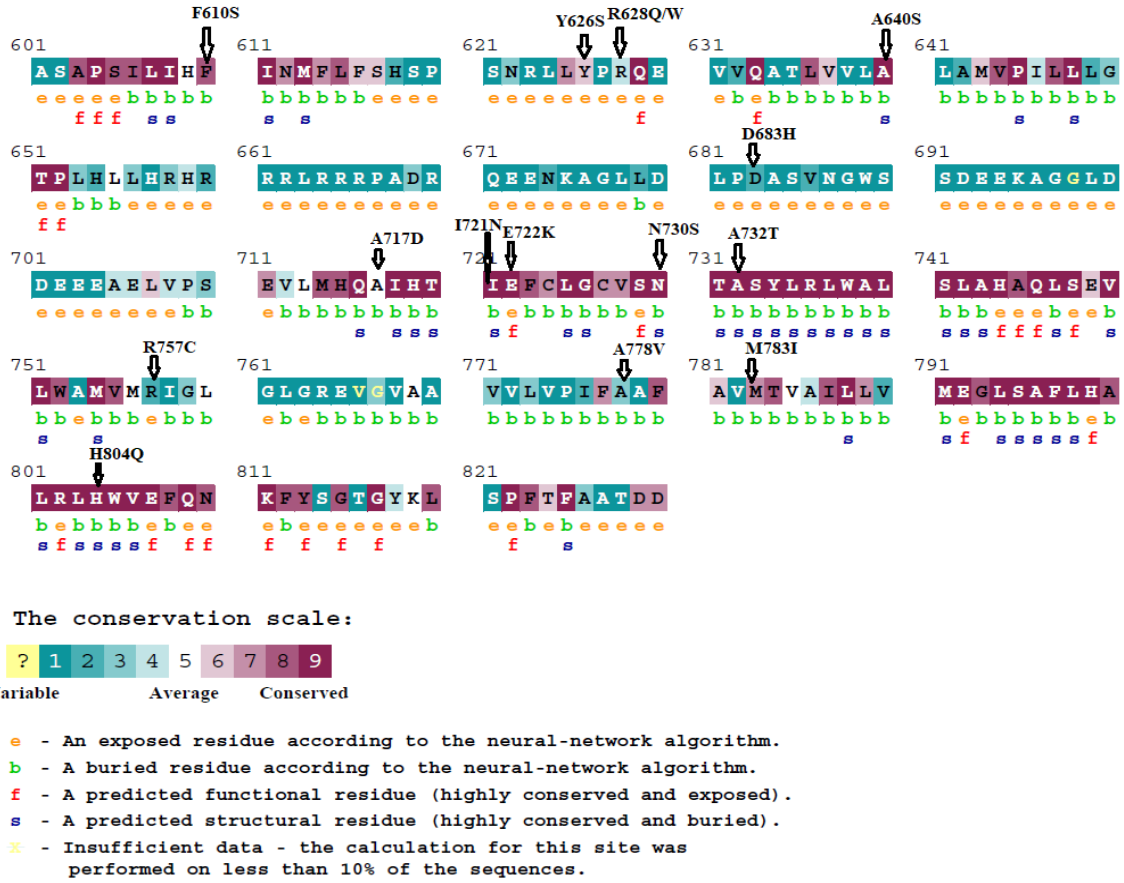


Figure 17 - Location of amino acid substitutions in the TCIRG1 protein considering evolutionary conservation and the location of different regions of the protein according to the ConSurf analysis

Note: Value 1 indicates high variability of the region, while 9 indicates the most conservative region in terms of evolution.

4.4- 3D modeling of protein structures of TCIRG1, ELANE with the identified potentially harmful amino acid substitutions taken into account

The prediction of protein structures, taking into account the selected amino acid substitutions identified in previous stages of the study, was performed using Phyre2, I-Tasser, HHpred, and AlphaFold2. For the protein TCIRG1, there were 15 amino acid substitutions, including the one identified in a patient from the Sverdlovsk region. All of these mutations were included in a single 3D structure of the TCIRG1 protein, as they were located in different regions. Thus, when

overlaying and comparing the models, differences in the 3D structure for each of the regions where a substitution could occur were observed.

The template for predicting the impact of substitutions on TCIRG1 in Phyre2 was the model template C6VQ7A (the template with the highest similarity according to the Phyre2 server data). Phyre2 was used to create 3D structures of the TCIRG1 protein considering its 56 mutations. nsSNP substitutions in the TCIRG1 protein sequence were modeled separately and then passed to Phyre2, which predicted 3D structures of the mutant proteins. However, our comparative studies showed that AlphaFold2 provided much higher quality results for analyzing TCIRG1 than Phyre2. Therefore, further MDS investigations of the TCIRG1 protein were conducted without using Phyre2.

An example of the 3D structure of the TCIRG1 protein in AlphaFold2 with the selected amino acid substitutions included in the study is presented in Figure 18. The wild-type structure was previously predicted by AlphaFold2 and is available for download from UniProt (identifier Q13488).

When comparing the 3D models of wild-type and mutant protein types, metrics for comparing models (TM-score) and root mean square deviation (RMSD) of distances between natural and mutant model carbon atoms (during molecular dynamics simulations for 50 and 100 ns) were determined. Low TM-score and high RMSD values indicated that the mutant structure differed from the wild-type structure. The corresponding analysis of 34 nsSNPs identified as harmful to the TCIRG1 protein during joint analysis using SIFT and PolyPhen2 is presented in Table 11.

The mutant R92W (rs371907380) has the highest RMSD value of 0.89B, followed by R444L (rs137853151), N730S (rs200209146), and S532C (rs371214361) with 0.84B, 0.84B, and 0.81B, respectively. F610S, M546V, and P572L have RMSD values of 0.81B, 0.78B, and 0.78B, respectively, indicating no significant structural differences from the wild type.

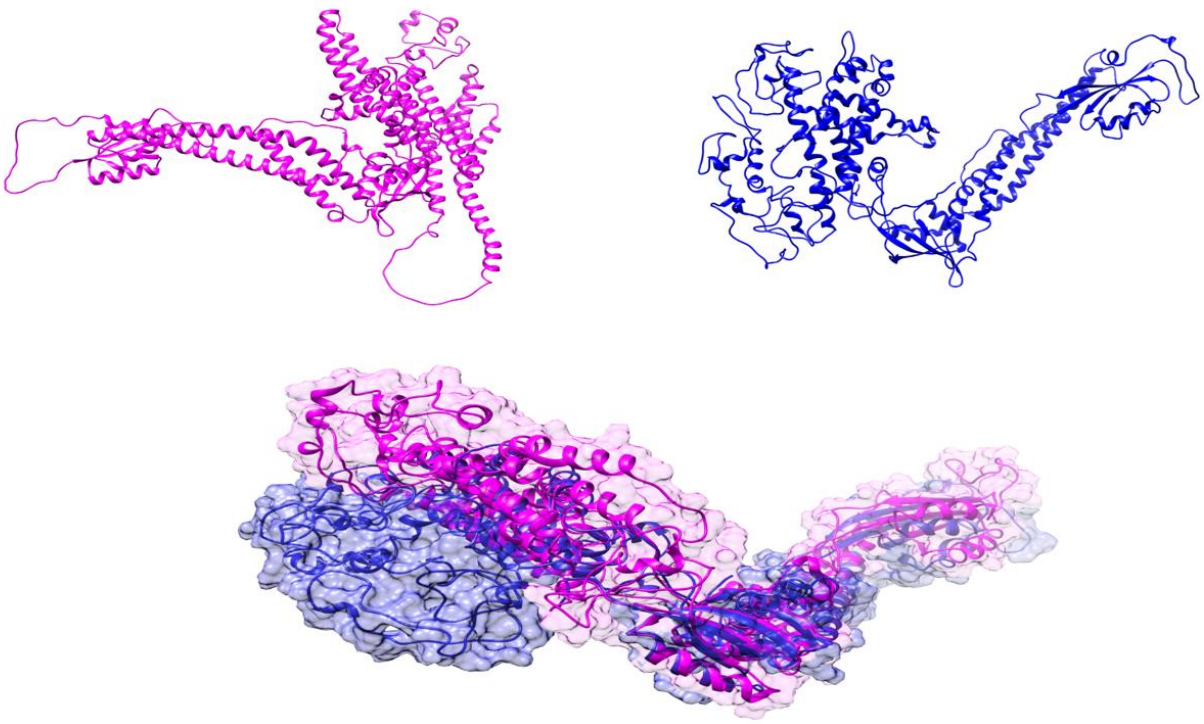


Figure 18 - 3D structure of wild-type and mutant-type TCIRG1 protein predicted by AlphaFold2, and superimposition of the three-dimensional structures (at the bottom)

Other nsSNPs showed minor differences, including I721N (0.53B RMSD), A732T (0.78B RMSD), R51C (0.78B RMSD), A717D (0.73B RMSD), E722K (0.46B RMSD), R57H (0.48B RMSD), R109W (0.78B RMSD), R191H (0.49B RMSD), G192C (0.78B RMSD), F529L (0.58B RMSD), H804Q (0.48B RMSD), G405R (0.48B RMSD), S474W (0.53B RMSD), G458S (0.48B RMSD), R56P (0.48B RMSD), R56W (0.78B RMSD), G379C (0.58B RMSD), R757C (0.48B RMSD), V375M (0.54B RMSD), T314M (0.78B RMSD), D517N (0.49B RMSD), T368M (0.78B RMSD), A417T (0.40B RMSD), R363C (0.78B RMSD), A778V (0.76B RMSD), and R50C (0.78B RMSD).

Four nsSNPs with the highest RMSD values (R92W, R444L, N730S, and S532C) were selected and submitted to I-Tasser for modeling. However, a comparative analysis of the results showed that higher-quality protein modeling results were obtained using HHPred and AlphaFold2. Therefore, below is a comparison of the 3D models of the wild-type and mutant TCIRG1 variants in AlphaFold2 - before starting the molecular dynamics simulation (Figure 19), at 50

nanoseconds of simulation (Figure 20), and at 100 nanoseconds of molecular dynamics simulation (Figure 21).

Table 11 - TMscore and RMSD values for 34 deleterious nsSNPs in TCIRG1

nsSNP	A.AVariants	TM-Score	RMSD
rs371907380	R92W	-	0,89
rs199902030	P572L	0.99626	0.78
rs200149541	M546V	0.99626	0.78
rs372499913	I721N	0.99760	0.53
rs267605221	F610S	0.99312	0.81
rs374941368	A732T	0.99621	0.78
rs375717418	F51S	0.99626	0.78
rs80008675	A717D	0.99661	0.73
rs149792489	E722K	0.99830	0.46
rs116675104	R57H	0.99790	0.48
rs121908250	R109W	0.99626	0.78
rs121908251	R191H	0.99785	0.49
rs121908251	S532C	0.99092	0.81
rs149792489	G192C	0.99626	0.78
rs116675104	F529L	0.99435	0.58
rs121908251	G405R	0.99674	0.62
rs116675104	G458S	0.99674	0.48
rs121908251	R56P	0.99657	0.48
rs121908252	R56W	0.99621	0.78
rs121908254	G379C	0.99435	0.58
rs147974432	R757C	0.99790	0.48
rs192224843	N730S	0.99275	0.84
rs115982879	V375M	0.99743	0.54
rs139059968	T314M	0.99626	0.78
rs141125426	D517N	0.99785	0.49
rs147208835	R92W	0.96213	0.89
rs147681552	T368M	0.99626	0.78
rs148498685	A417T	0.99790	0.48
rs149531418	R363C	0.99626	0.78
rs149531418	A778V	0.99661	0.76
rs147208835	R50C	0.99621	0.78
rs121908250	H804Q	0.99790	0.48
rs149792489	S474W	0.99760	0.53
rs121908250	R444L	0.99270	0.84

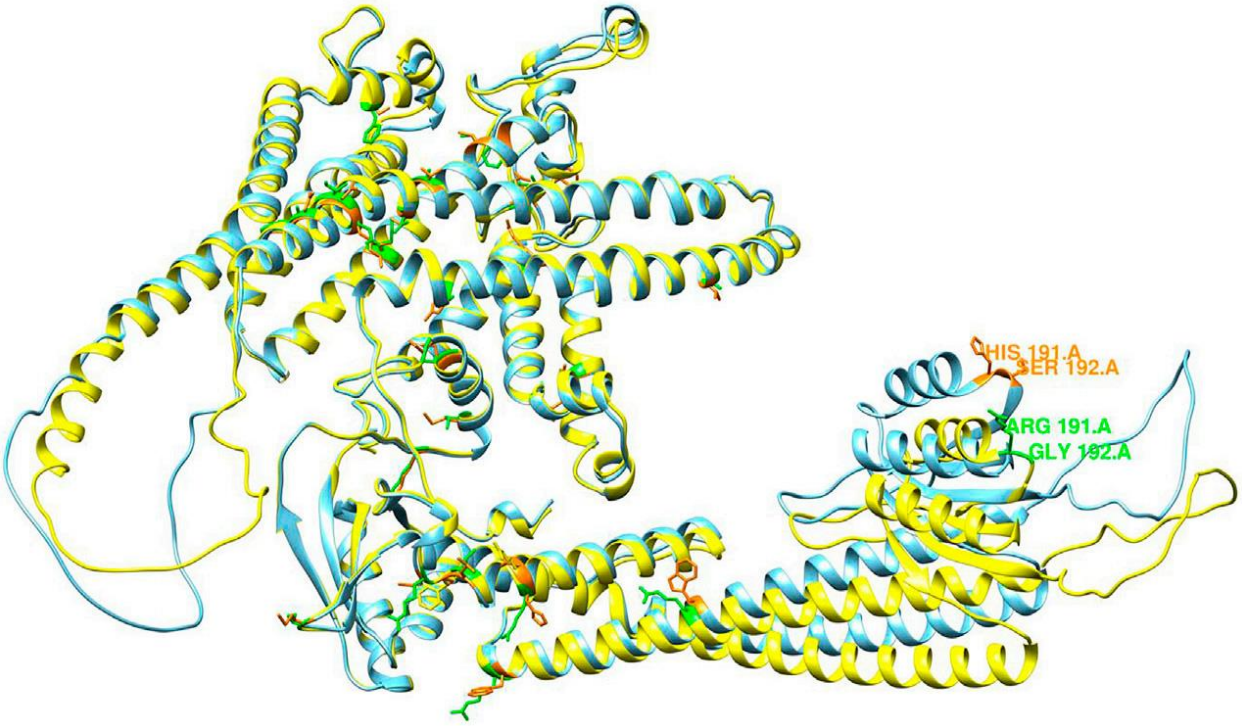


Figure 19 - Overlay of the 3D structures of wild-type (yellow) and mutant type (blue) TCIRG1 protein prior to the start of molecular dynamics simulation

Note: The most deleterious substitutions incorporated into the model are highlighted in orange, and the corresponding regions on the wild-type model are highlighted in green.

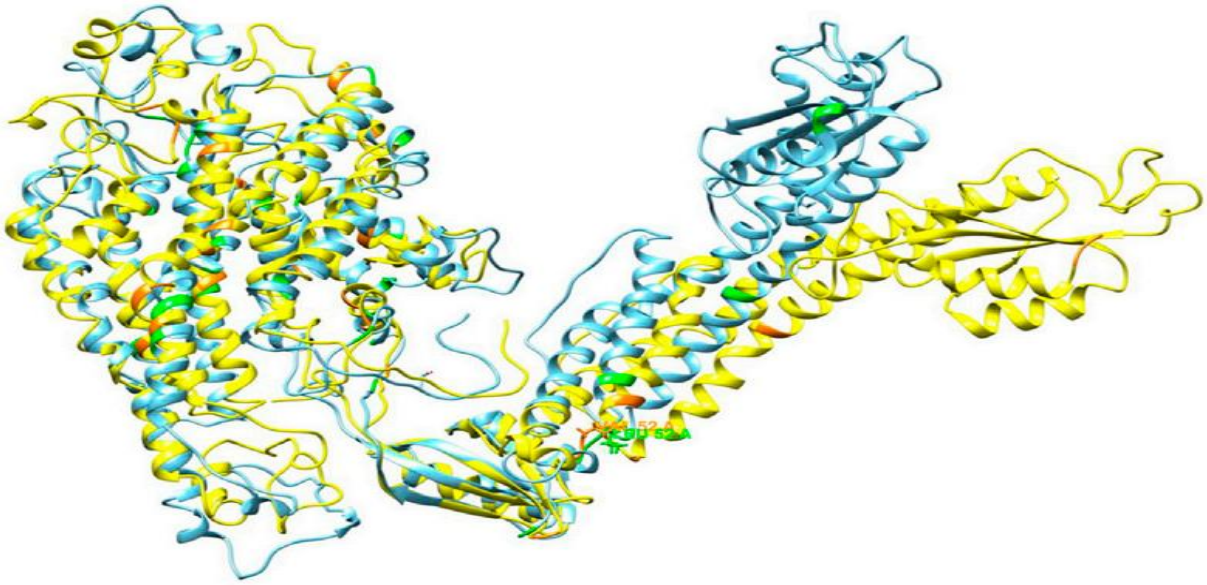


Figure 20 - Overlay of 3D structures of wild-type (yellow) and mutant-type (blue) TCIRG1 protein after 50 nanoseconds of molecular dynamics simulation

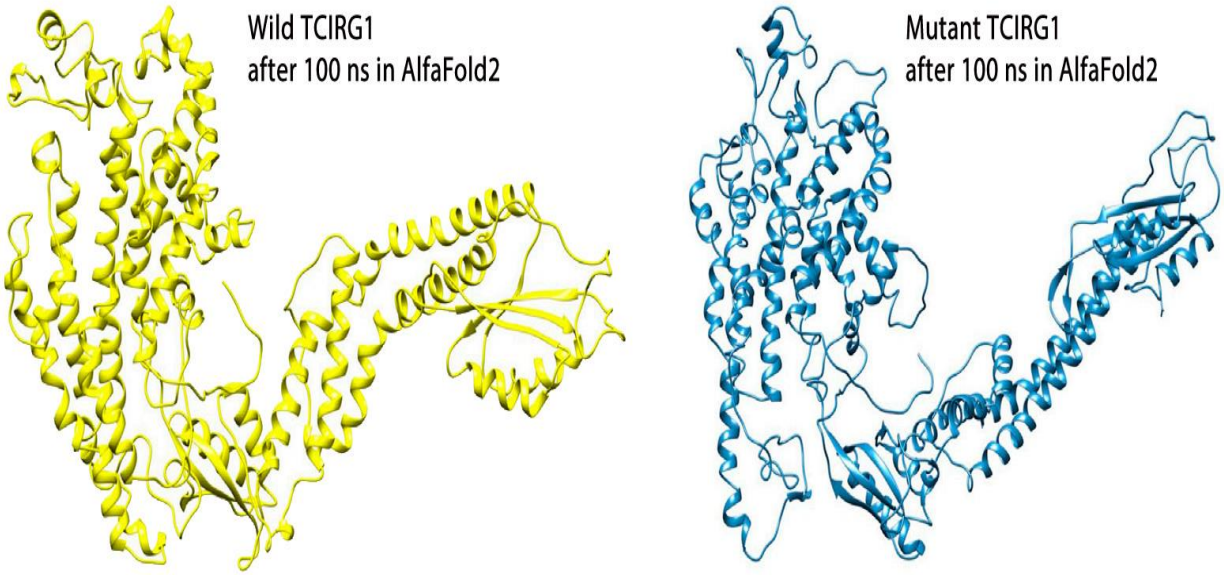


Figure 21 - 3D structures of wild-type (yellow) and mutant (blue) TCIRG1 protein during a 100-nanosecond molecular dynamics simulation

Subsequently, these selected mutant types of TCIRG1 were evaluated using Schrodinger packages in molecular dynamics simulations. Phyre2 was used to model the 3D structures of both the wild-type and mutant types of the ELANE protein. The c6o1gA model was chosen as the template for predicting the 3D model of ELANE in Phyre2 (Figure 22). The predicted 3D structures of the mutant proteins are shown in Figure 23

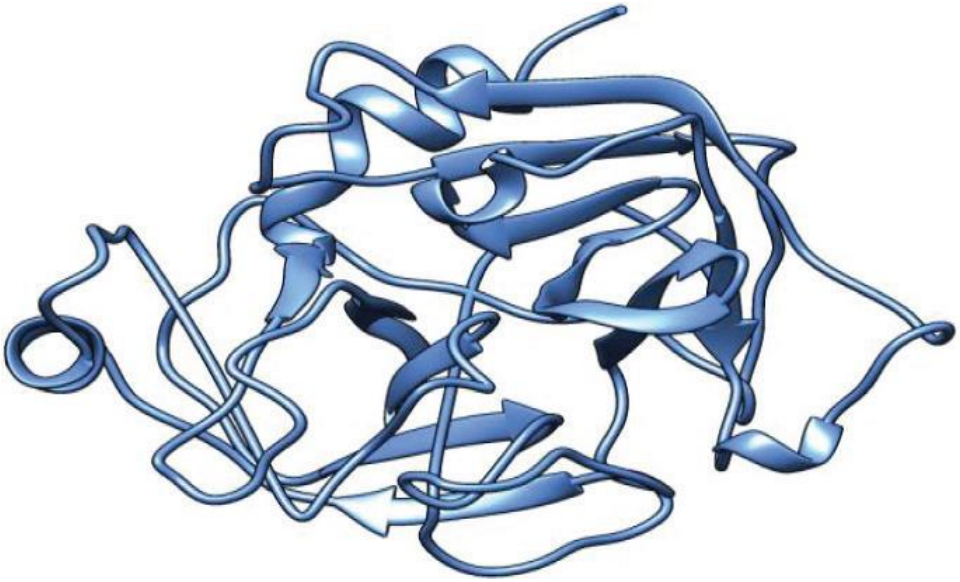


Figure 22 - Wild-type ELANE protein model generated by Phyre2

The I-Tasser program was used for three-dimensional modeling of the ELANE protein. The resulting 3D models from I-Tasser were then uploaded to the Zhanggroup online service, which provided metrics for comparing the models, including the TM-score and root-mean-square deviation (RMSD).

The mutant model C71Y (based on nsSNP rs28931611) had the highest deviation from the wild-type ELANE template, with an RMSD value of 2.05Å. This was followed by R34W (rs201163886), F218L (rs200384291), and G214R (rs137854451), with RMSD values of 1.98Å, 1.96Å, and 1.12Å, respectively. P139L, G203S, and R193W had RMSD values of 0.04Å, 0.49Å, and 0.96Å, respectively, indicating no significant structural differences from the wild-type. Table 12 shows the TM-scores and RMSD values for the ELANE mutant types.

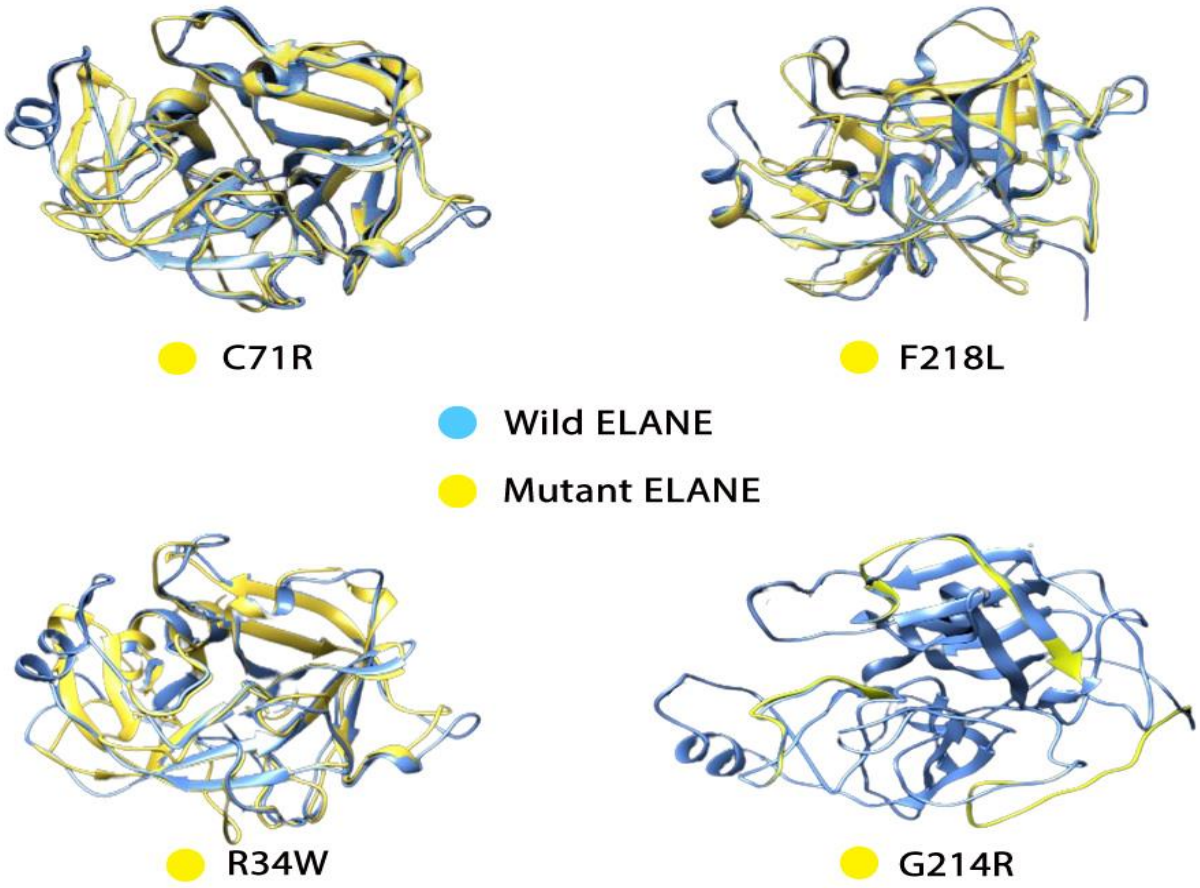


Figure 23 - Overlay of wild-type (blue) and mutant (yellow) ELANE proteins including 4 most significant amino acid substitutions: C71R, F218L, R34W, and G214R

The four nsSNPs with the highest RMSD values (C191Y, G214R, R34W, and F218L) were selected and passed into I-Tasser for remodelling and comparison with the wild-type structure (Figure 23). The verification results for the wild-type and mutant 3D models were satisfactory. These selected ELANE mutant types were subsequently used in an in silico experiment for molecular docking screening.

Table 12 - TM-score and RMSD values for 7 selected nsSNPs in ELANE

nsSNP	A.A.S	TM-Score	RMSD
rs28931611	C71Y	0.85993	2.05
rs201163886	R34W	0.86482	1.98
rs200384291	F218L	0.87828	1.96
rs137854451	G214R	0.96114	1.12
rs201723157	R193W	0.95176	0.96
rs201139487	G203S	0.99524	0.49
rs137854448	P139L	0.99994	0.04

Note: A.A.S - Amino acid substitutions

4.5 - Evaluation of the interaction of mutated ELANE types by docking

Docking was visually evaluated using Discovery Studio and PyMol, and docking interactions were calculated to identify binding strengths, which were decisive in stabilizing the formation of receptor-ligand complexes. The ANH ligand was docked into the active site of the wild-type protein as well as four mutant proteins. The docking score for the wild-type was -8.4 kJ/mol, and 2D interaction showed that the wild-type had two hydrogen bonds with SER202, as well as seven van der Waals and seven hydrophobic contacts (Figure 25). The docking score for the G214R, R34W, C71Y, and F218L mutations was -9.2, -7.5, -7.1, and -6.8 kJ/mol, respectively. The 2D interaction for G214R showed two hydrogen bonds

with SER202, seven van der Waals, and five hydrophobic interactions. The R34W substitution showed two hydrogen bonds with SER202 and VAL219, six van der Waals interactions, and eight hydrophobic interactions. One hydrogen bond was present in the C71Y mutation with ARG81, four van der Waals, and eight hydrophobic interactions. Figure 26 shows the interactions for ELANE with the C71Y substitution. Similarly, F218L showed one hydrogen bond with ASN74, six van der Waals, and four hydrophobic interactions (Figure 27). Two mutations, G214R and R34W, have interactions quite similar to the wild-type. All of them are involved in a hydrogen bond with SER202. The other two substitutions, C71Y and F218L, have fewer hydrogen bonds, indicating that these two mutations may affect the stability and energy of the protein (Figures 24-26).

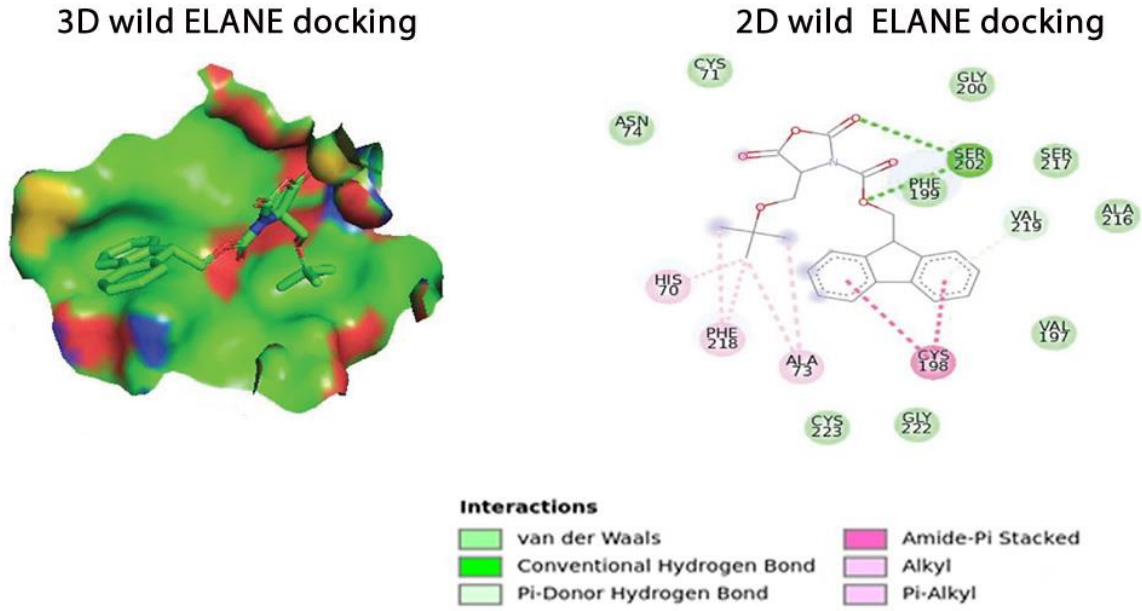


Figure 24 - 2D and 3D surface plots of wild-type ELANE with a ligand inside the active pocket

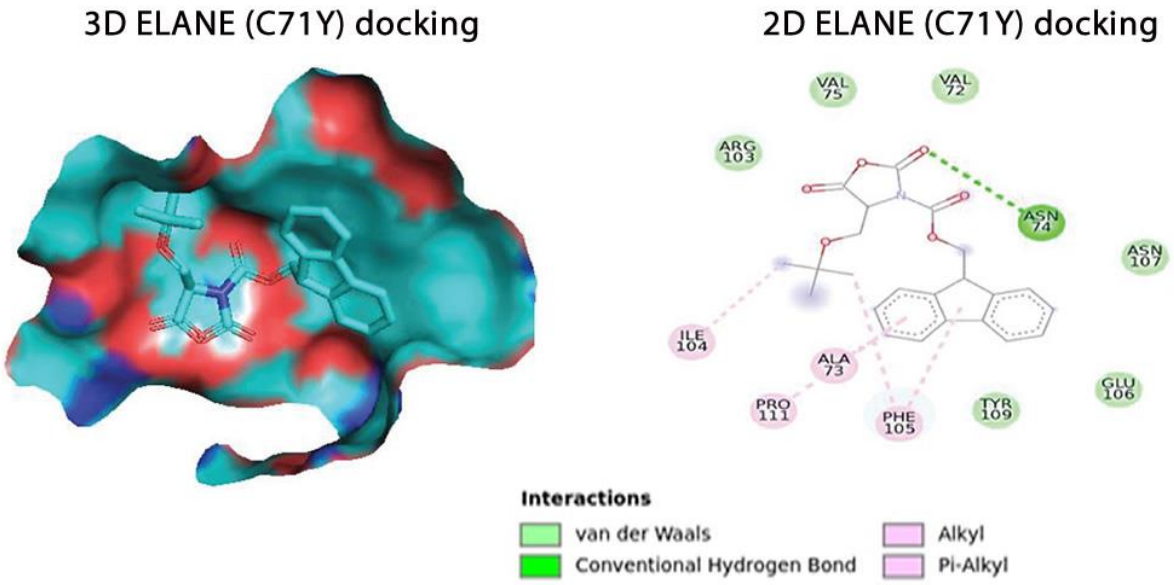


Figure 25 - 2D and 3D surface graphs of ELANE with C71Y substitution and ligand inside the active pocket

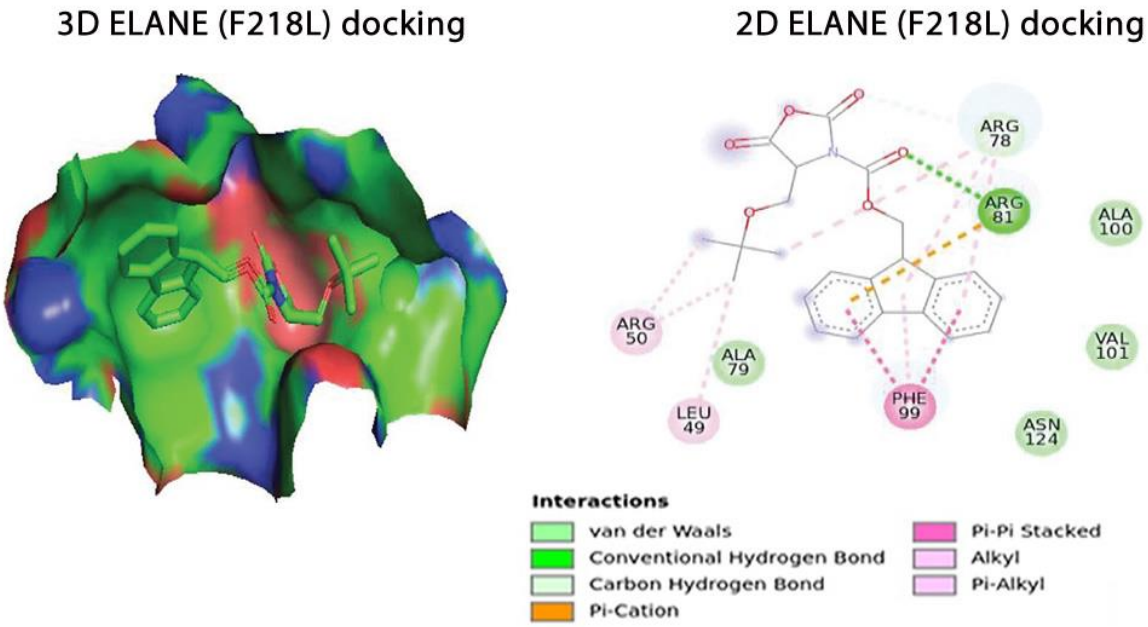


Figure 26 - 2D and 3D surface graphs of ELANE with F218L substitution and ligand inside the active pocket

4.6 - Molecular dynamic simulation of wild-type and mutant TCIRG1

During the molecular dynamics simulations in the HHPred program (Figure 27) and AlfaFold2 (Figure 28), the evolution of the root mean square deviation (RMSD) of the alpha-carbon atoms (Cα) in the protein molecule over time was

generated. The graph in Figure 28, obtained from HHPred, showed that the protein reached stability at 20,000 ps. Subsequently, throughout the simulation time, the fluctuation of the RMSD values for the wild type remained within 2.0 angstroms, which is acceptable [78]. The RMSD values for the mutant protein fluctuated within 3.5 angstroms after they had been equilibrated. These results indicate that the mutant protein has a higher RMSD throughout the simulation period. On the RMSF plot, peaks represent protein parts that oscillate the most during the simulation (Figure 29).

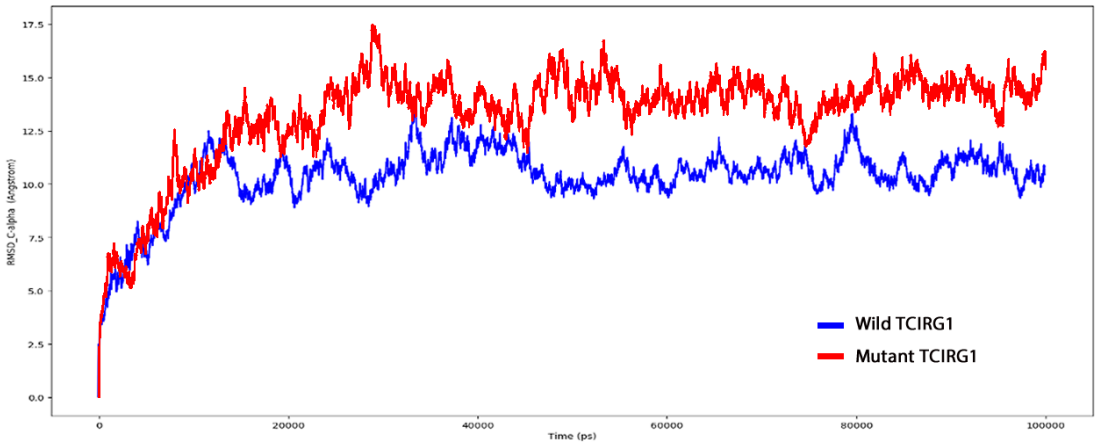


Figure 27 - Root-mean-square deviation (RMSD) of wild type and mutant Cα atoms over time (100 ns) based on HHpred data

Note: The x-axis represents time in picoseconds (ps), and the y-axis represents RMSD in angstroms (Å).

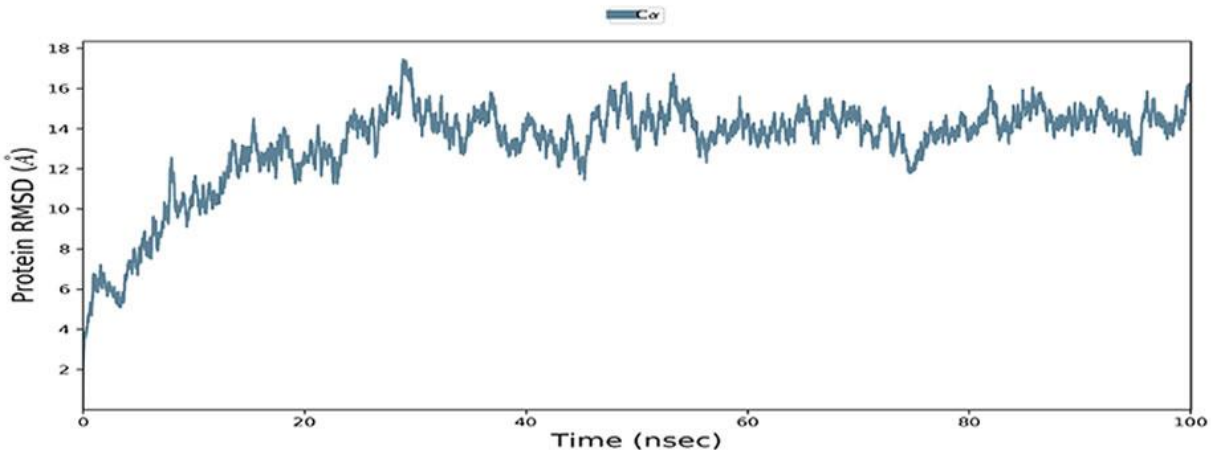


Figure 28 - Root mean square deviation (RMSD) of wild-type and mutant TCIRG1's Cα atoms over time (100 ns) according to AlfaFold2 data

Note: The x-axis represents time in picoseconds (ps), and the y-axis represents RMSD in angstroms (Å).

Figure 30 shows the total energy of the mutant and wild-type TCIRG1 protein, and Figure 31 shows the Van der Waals energy of the wild-type and mutant TCIRG1. Protein tails (both N- and C-terminal) usually undergo more significant changes than any other part of the protein. Alpha helices and beta sheets, for example, are usually more rigid than the unstructured part of the protein and oscillate less than loop parts. According to the calculated MD trajectories, residues with large peaks belong to loop regions or N- and C-terminal zones. Alpha helices and beta sheets are tracked as secondary structure elements (SSE) during modeling. Figure 32 shows the distribution of secondary structures by residue index for all protein structures, and Figure 33 shows the distribution of secondary structure elements over the simulated time of 100 ns. All of these results indicate that the stability of the mutant TCIRG1 molecule is reduced relative to the wild-type protein.

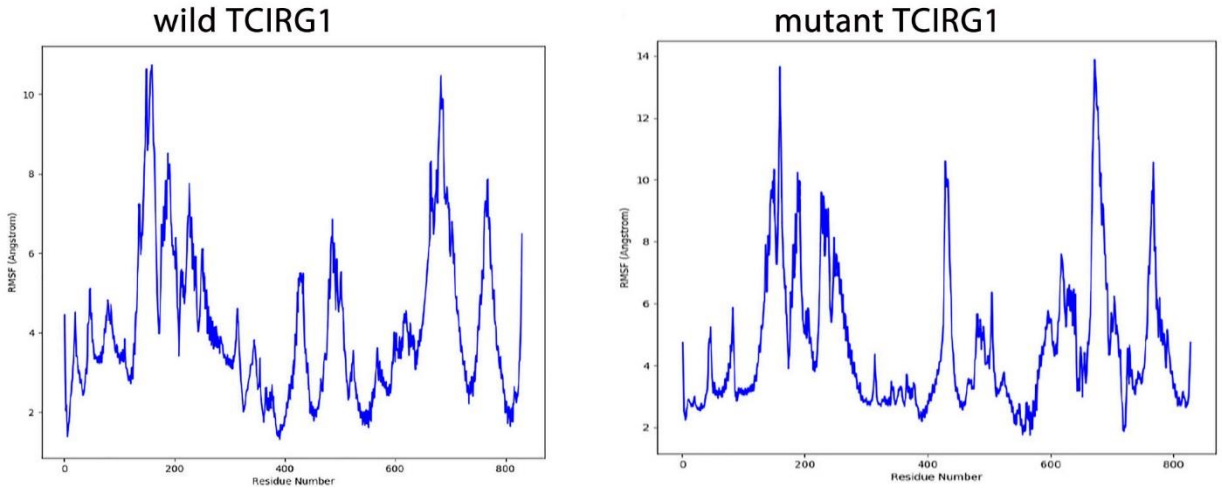


Figure 29 - Root Mean Square Fluctuation (RMSF) of wild-type TCIRG1 protein (left) and mutant TCIRG1 protein (right)

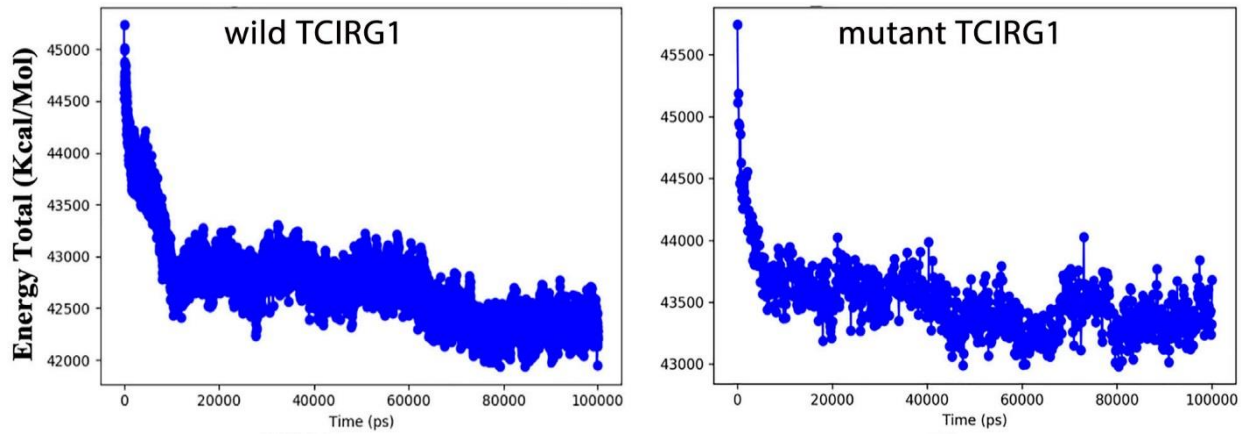


Figure 30 - Total energy of wild-type and mutant TCIRG1 protein

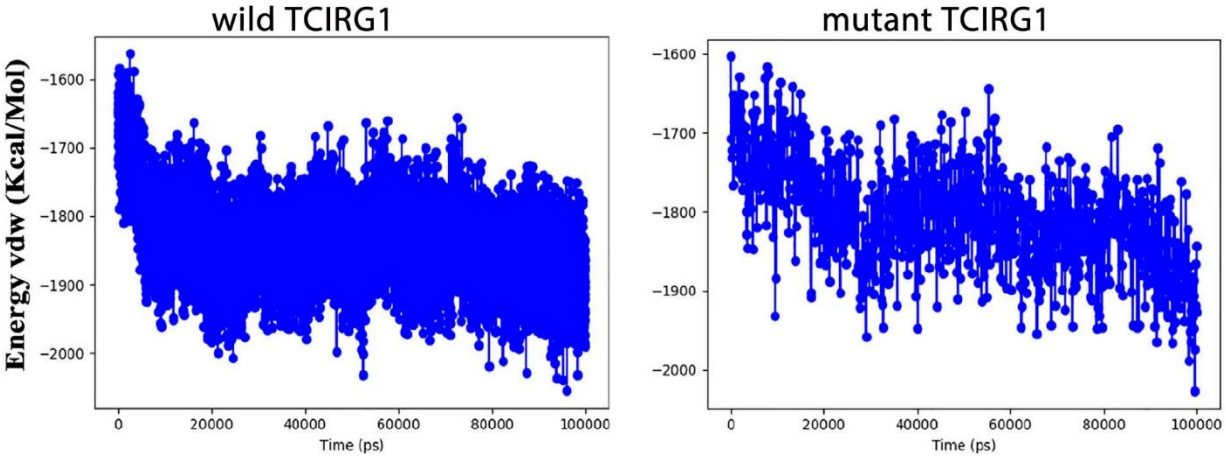


Figure 31 - Van der Waals energy of wild-type and mutant TCIRG1 protein

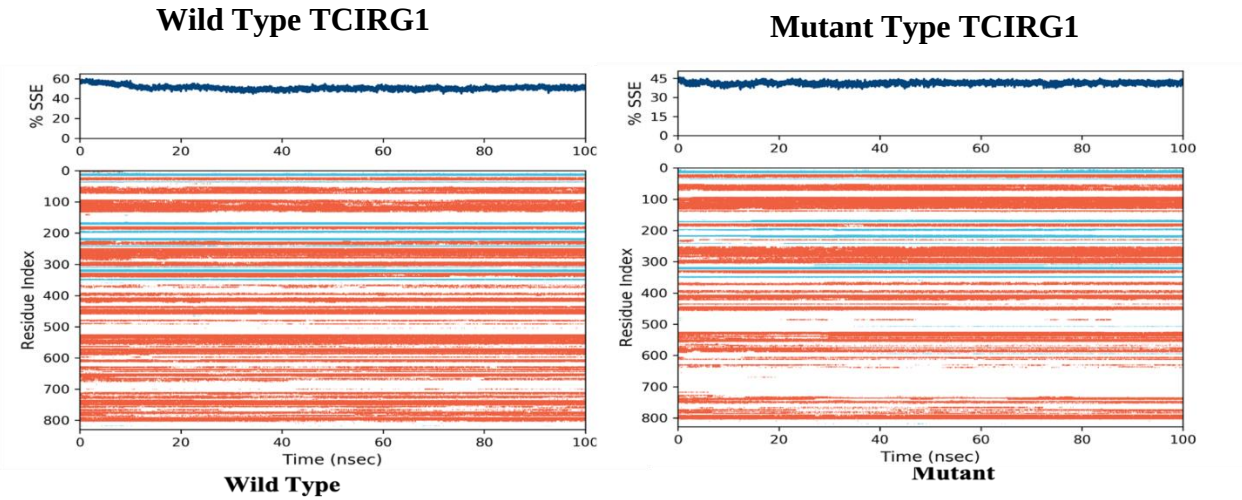


Figure 32 - Percentage of secondary structure elements in wild-type and mutant TCIRG1 protein

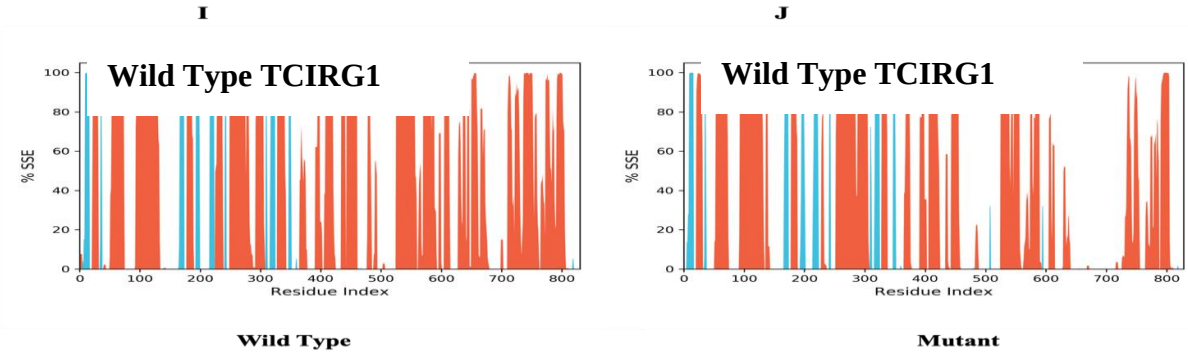


Figure 33 - Distribution of secondary structure elements during the simulated time of 100 ns

The majority of significant intramolecular interactions detected using molecular dynamics simulations are hydrogen bonds (Figure 34). The time scale shows the interactions and contacts. The distribution of atoms in a protein around its axis is known as the radius of gyration (Rg). The folding speed of a protein is directly related to its compactness, which can be tracked using an advanced computational approach to determine the radius of gyration (Figure 35).

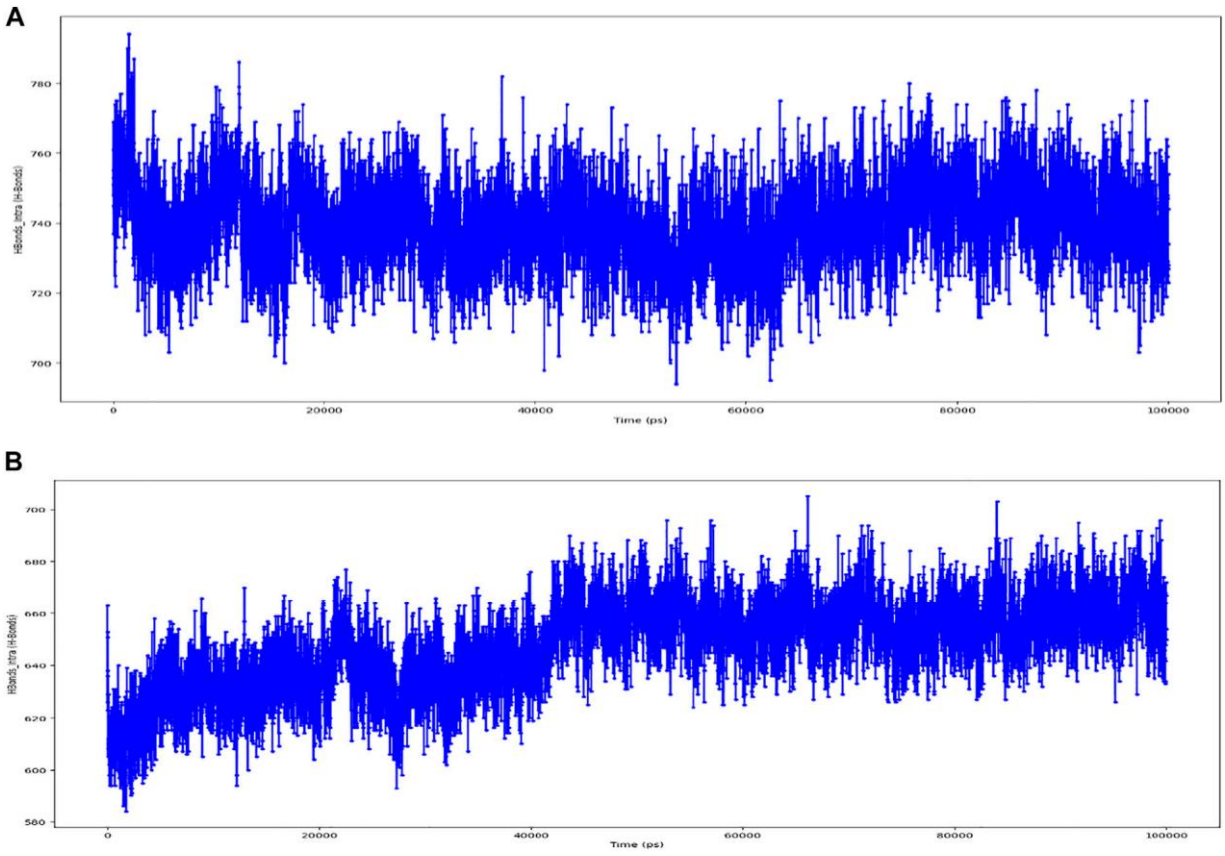


Figure 34 - Temporal representation of hydrogen bond interactions and contacts in wild-type (A) and mutant (B) TCIRG1 protein

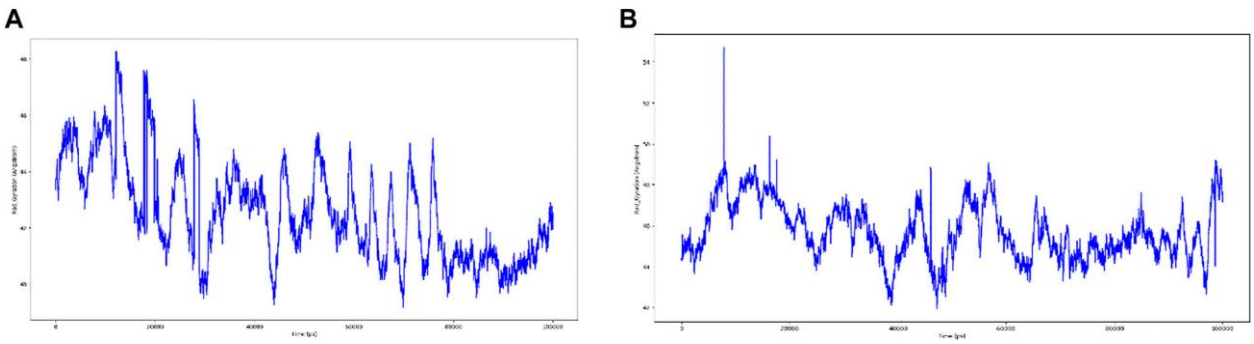


Figure 35 - Radius of gyration of wild-type (A) and mutant (B) TCIRG1 protein

We did not perform molecular dynamics simulation and corresponding amino acid substitution analysis for the ELANE protein. Only docking was conducted for ELANE. This is because the relevance of studying TCIRG1 was justified by having the patient's whole-genome data with phenotypic manifestations of congenital neutropenia, but with clinical features that raised doubts about the accuracy of the diagnosis. After whole-genome sequencing was performed in a commercial laboratory, the diagnosis was not confirmed. Therefore, the decision was made to apply new methods to identify mutant genes related to the patient's phenotype and confirm the diagnosis.

Analysis of the whole-genome sequencing data revealed several potentially significant mutations, but only one was related to neutropenia. Specifically, a non-synonymous single nucleotide substitution g. 68041789G >C was identified in the TCIRG1 gene (amino acid substitution V52L). This substitution was included in the list of substitutions analyzed above, allowing for a more justifiable assumption that the variant gene found in the patient may have clinical significance.

Thus, non-synonymous single nucleotide substitutions in the TCIRG1 (rs199902030, rs200149541, rs372499913, rs267605221, rs374941368, rs375717418, rs80008675, rs149792489, rs116675104, rs121908250, rs121908251, rs121908251, rs149792489, and rs116675104) and ELANE (rs200384291, rs201163886, rs193141883, rs201139487, and rs201723157) genes destabilize the protein structure and function.

4.7 - Investigation of candidate genes in congenital neutropenia

The first step in searching for or predicting new candidate genes for a congenital disease is a review analysis of published information, analysis of information in genome and inherited disease databases, as well as a review of genetic studies related to the specific disease.

Simple information search in genetic databases such as OMIM (Online Mendelian Inheritance in Man) and HGMD (Human Gene Mutation Database) helped identify previously registered mutations leading to diseases, and analysis of publications in PubMed suggested the direction of further research.

Analysis of protein-protein interactions (PPI) of known genes in congenital neutropenia in the human genome was the key to understanding the multigenic nature of congenital neutropenia and further identification of candidate genes.

Using the STRING database, information on protein-protein interactions (PPI) was extracted for all known genes in primary immunodeficiencies (PID). Its visual representation using Cytoscape software is shown in Figure 36, where genes in congenital neutropenia are shifted to the center of the network of interactions. This suggested that genes in congenital neutropenia interact more often than random PID genes, which is logical, since despite the different genetic cause of congenital neutropenias, the phenotypes of different disease variants are similar, and common signaling pathways are involved in providing a similar pathogenesis. In congenital neutropenia, mechanisms related to regulating the number and functions of neutrophils are primarily disrupted.

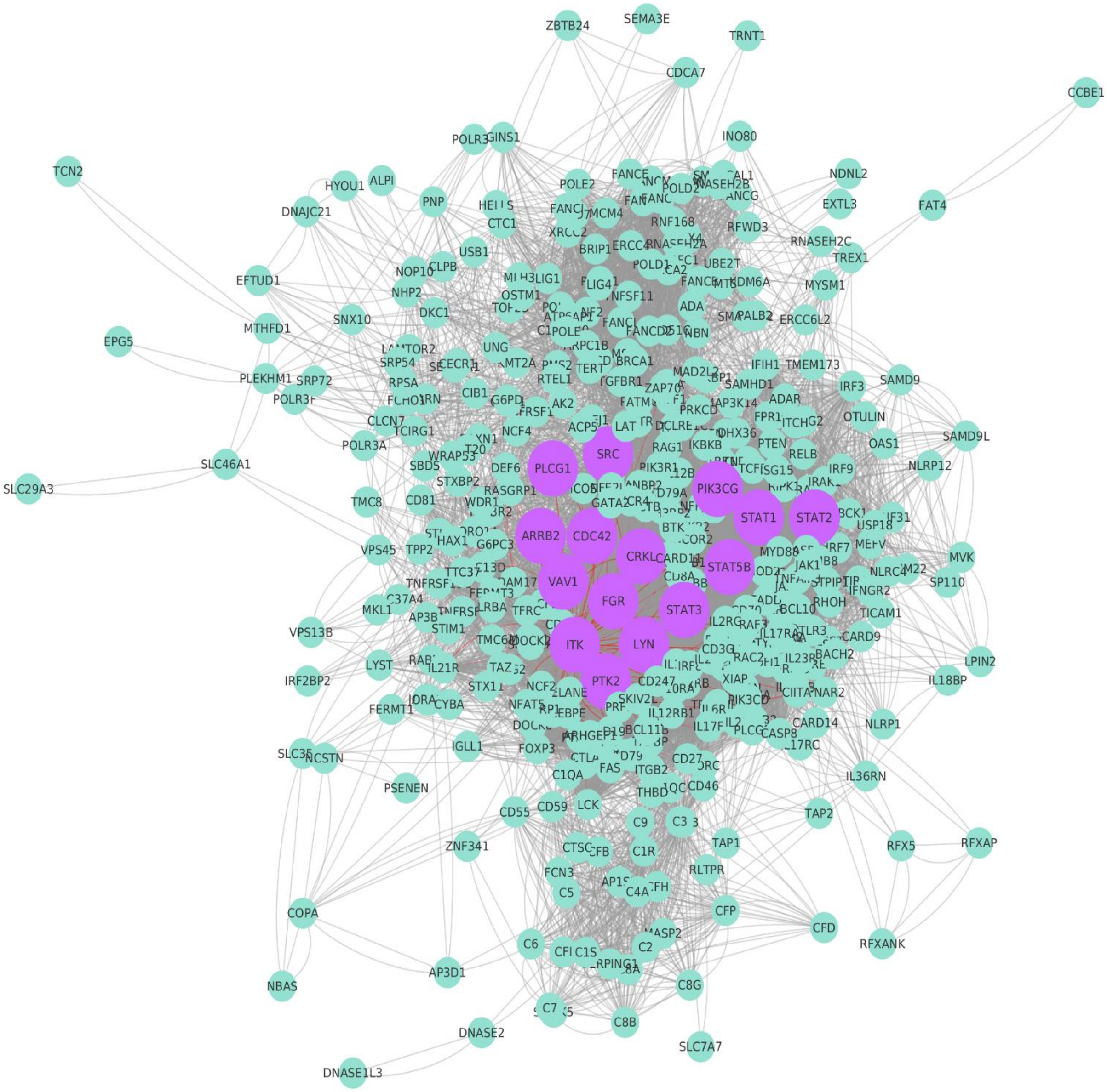


Figure 36 - Protein-protein interaction network of PID genes (Cytoscape was used to visualize the data extracted from the STRING database)

Note: The known genes of congenital neutropenia are represented by purple nodes, while the PID genes are represented by green nodes in the network.

The figure 37 shows a plot of the functional relationships between known genes associated with congenital neutropenia.

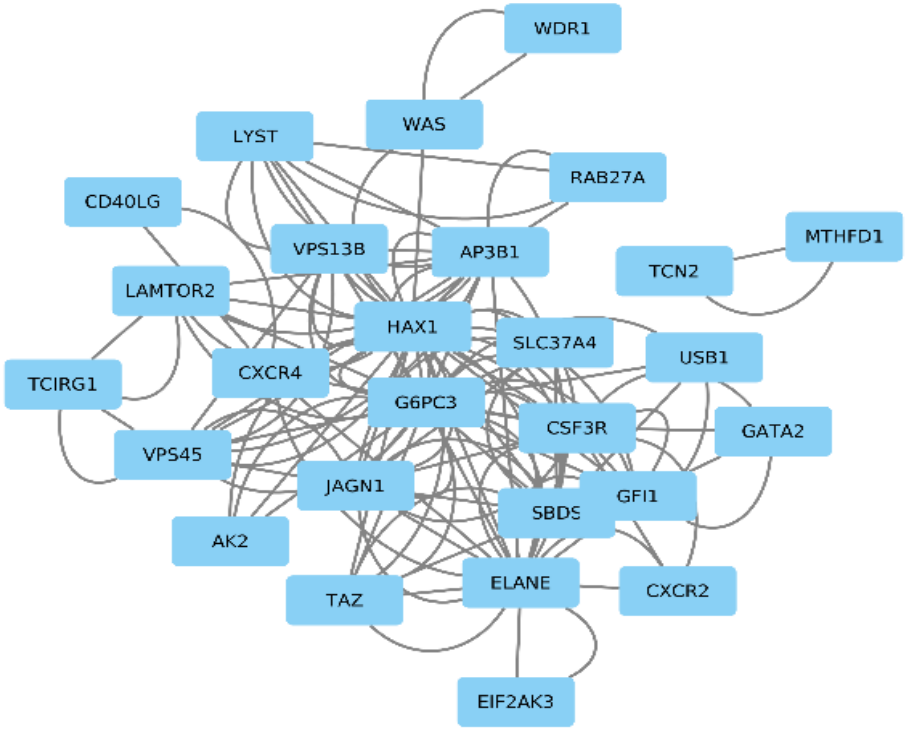


Figure 37 - A visualized network of protein-protein interactions between known genes of congenital neutropenia, extracted from the STRING database (result of analysis in Cytoscape)

To further investigate the complex gene interactions in congenital neutropenia, the network density of a group of 31 congenital neutropenia genes was evaluated and compared with ten random PID groups, each consisting of 41 genes. The connectivity and network density of PPI networks in each group were then measured and compared using the network density estimation method (network D), and our results showed a higher network density in the congenital neutropenia group compared to the 10 random groups. These results indicated a strong interaction between congenital neutropenia genes (Figure 38).

We also analyzed the distribution of biological distance between the group of known genes associated with congenital neutropenia and two random groups, taking into account that a smaller biological distance indicates a stronger association between the genes in the group (Figure 39).

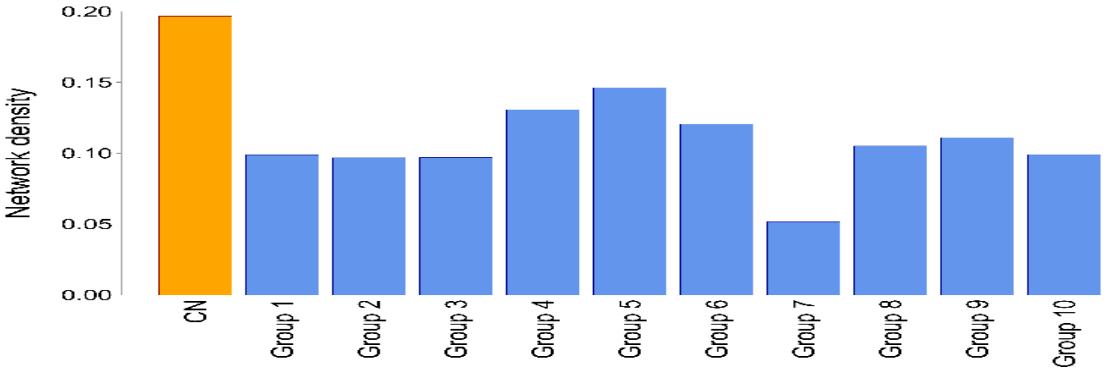


Figure 38 - Comparison of network density between a group of known congenital neutropenia genes and ten random PID gene groups

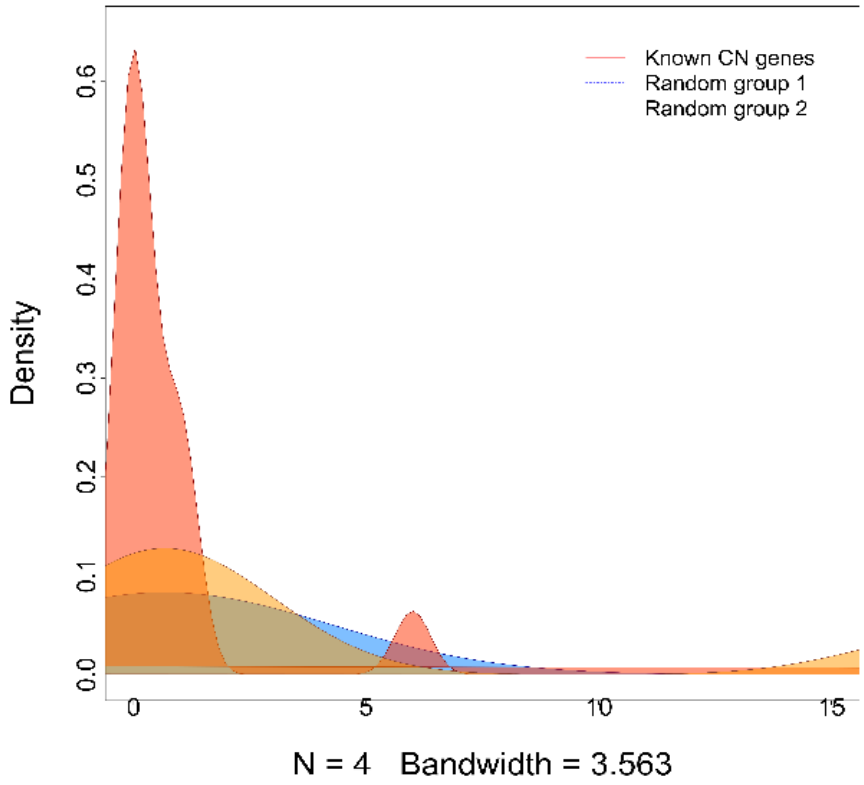


Figure 39 - Comparison of biological distance in the group of known genes for congenital neutropenia and 2 random PID groups

The results showed that the median density of the distribution of the group of known genes associated with congenital neutropenia is 1.067, whereas the median density of the distribution of random groups 1 and 2 is around 2.6, indicating a stronger functional relationship between the known genes associated with congenital neutropenia (Figure 39).

We also studied the distribution of biological distance between the group of known genes associated with congenital neutropenia and two random groups (a smaller biological distance indicates a stronger association between genes in the group). The results showed that the median density of the distribution of the group of known genes associated with congenital neutropenia was 1.067, while the median density of the distribution of random group 1 and random group 2 was about 2.6, indicating a closer functional relationship between the known genes associated with congenital neutropenia (Figure 39).

Based on Pearson correlation analysis (PCC) and protein-protein interactions provided by Cheng F., et al. (2018) [203], we obtained 4,613 specific gene interactions functionally related to congenital neutropenia and 177 candidate genes. Using KEGG data, we conducted functional enrichment analysis of known congenital neutropenia genes by linking the genes in the list to their biological functions. Our KEGG pathway analysis revealed five statistically significant signaling pathways ($p<0.05$), such as cytokine-cytokine receptor interaction, chemokine signaling pathways, and others (Figure 40).

We searched for specific candidate genes that are functionally similar to known congenital neutropenia genes and enriched in at least one of the aforementioned five KEGG pathways. Thus, we identified 15 new candidate genes for congenital neutropenia: STAT1, STAT2, STAT3, STAT5B, LYN, FGR, SRC, PIK3CG, ITK, VAV1, CDC42, PTK2, CRKL, PLCG1, and ARRB2.

Figure 41 shows the PPI network of known congenital neutropenia genes and candidate genes. Functional enrichment analysis of congenital neutropenia genes, including the 15 candidate genes, showed a total of 15 statistically significant signaling pathways described in the KEGG database (e.g., Epstein-Barr virus infection, cytokine-cytokine receptor interaction, and B-cell receptor signaling pathway).

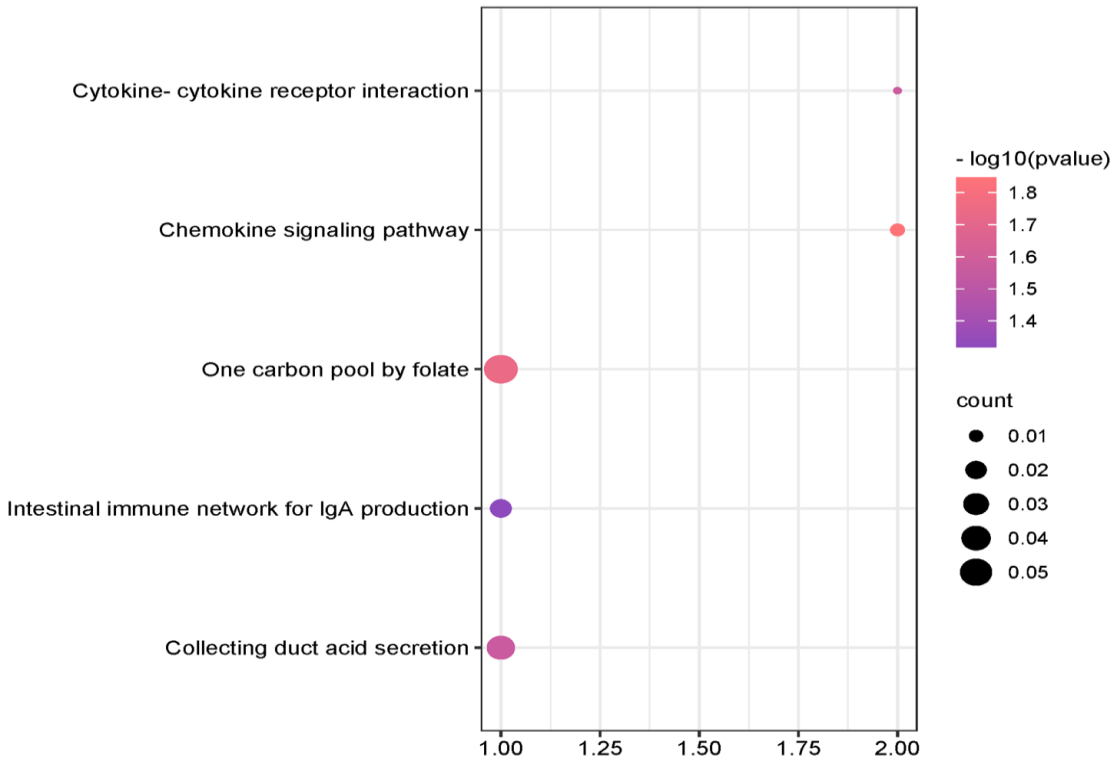


Figure 40 - Analysis of functional enrichment of candidate genes for congenital neutropenia based on the KEGG database

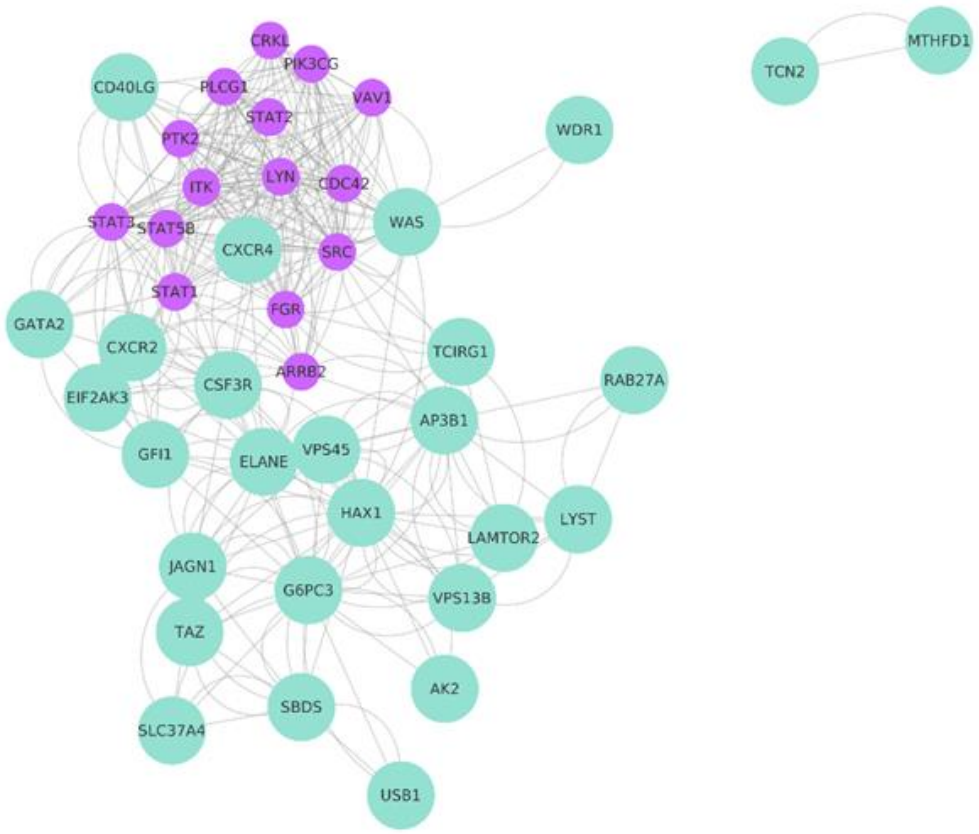


Figure 41 - Protein-protein interaction network of known and candidate genes associated with congenital neutropenia (Cytoscape)

Biological distances between the 15 candidate genes involved in congenital neutropenia were assessed and compared to the biological distances of 31 known genes involved in congenital neutropenia. As a result, the mean biological distance of the candidate genes was 6.08, which was lower (or equivalent) than that of the known genes involved in congenital neutropenia. This indicates that the candidate genes for congenital neutropenia have comparable strong biological connections (Figure 42).

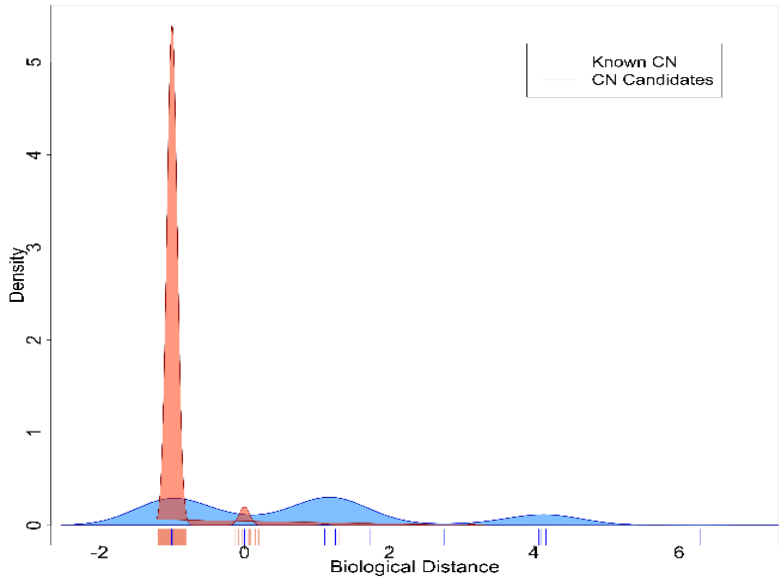


Figure 42 - Density plot of biological distances between known genes involved in congenital neutropenia and predicted candidate genes

Then the candidate gene for congenital neutropenia was mixed with known congenital neutropenia genes, and the biological distance of the mixed gene was determined again. Then the mixed genes were subjected to phylogenetic analysis FGA to determine the biological relatedness between the congenital neutropenia genome and the candidate congenital neutropenia genome. The results showed that the candidate genes for congenital neutropenia were evenly distributed across the range of known congenital neutropenia genes, implying their close association with known congenital neutropenia genes (Figure 43).

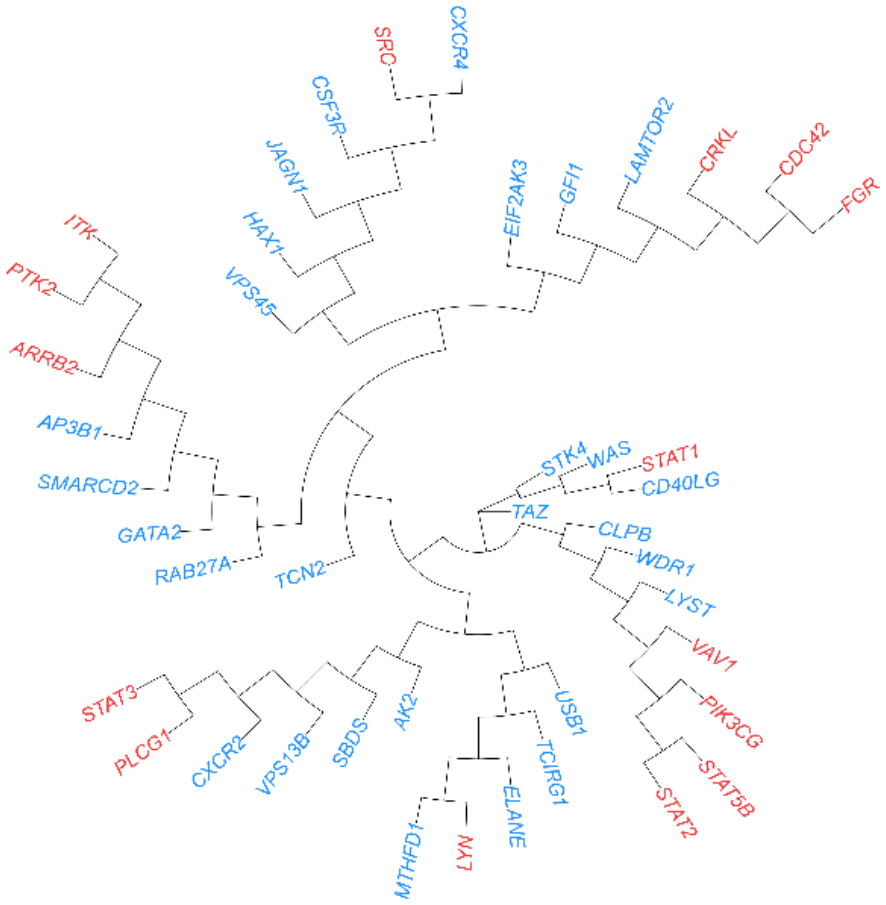


Figure 43 - Phylogenetic tree of biological distances generated by FGA showing hierarchical clustering of all known congenital neutropenia genes (blue) and predicted congenital neutropenia genes (red)

Note: the length of the branch indicates the strength of the separation between subjects

In addition, a diagram of the interrelationships between the candidate genes for congenital neutropenia and their associated signaling pathways was formed using KEGG through functional enrichment analysis (Figure 44).

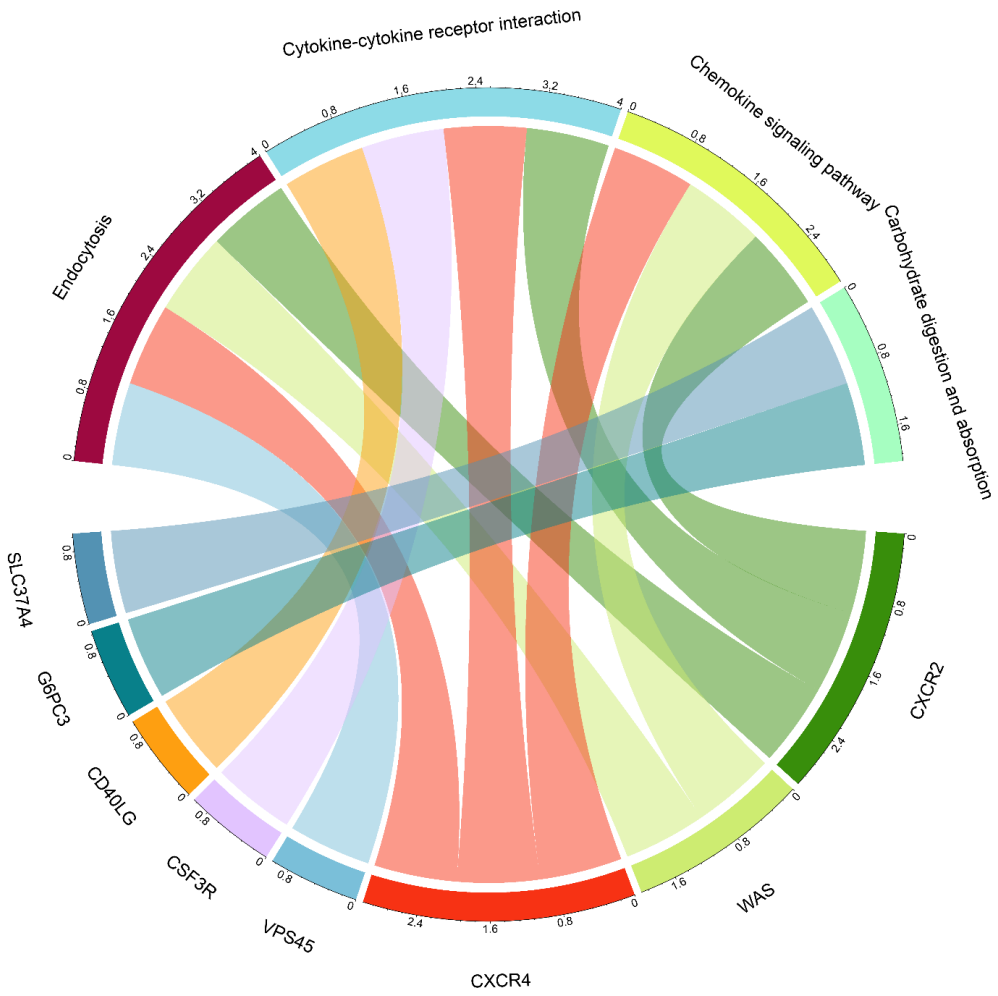


Figure 44 - Chord diagram of candidate genes for congenital neutropenia and their associated signaling pathways (based on KEGG data)

Assessment of gene expression differences in peripheral blood neutrophils of patients allowed for a search for new candidate genes from a different perspective, confirming our preliminary findings.

In the GSE142347 dataset, the expression of 1327 genes was significantly different in peripheral blood neutrophils of patients with congenital neutropenia compared to healthy controls, with 739 genes upregulated and 558 genes downregulated in expression. In the GSE6233 dataset, 573 genes were found to have significant differential expression in B-cells of patients with congenital neutropenia compared to control samples, with 274 genes upregulated and 299 genes downregulated in expression (Figure 45).

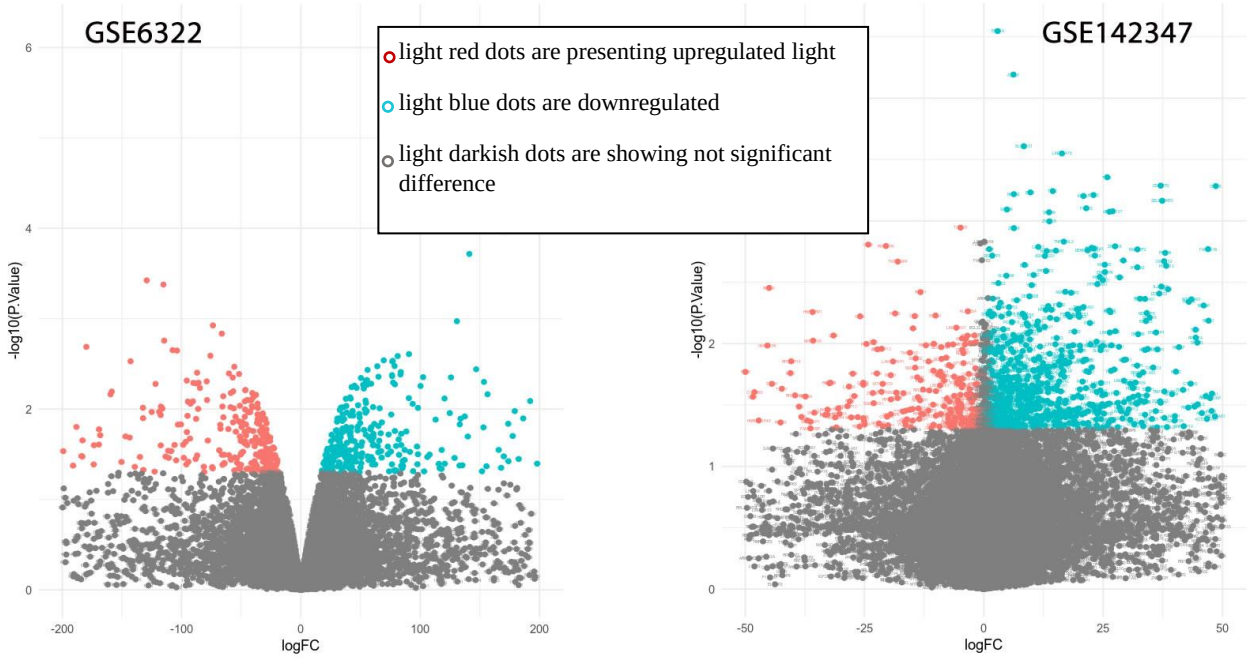


Figure 45 - Volcano plot of differentially expressed genes in samples from the GSE6322 (left) and GSE142347 (right) datasets
Note: light red dots are presenting upregulated light blue are downregulated while light darkish dots are showing not significant difference

In addition, comparison of genes with increased expression in neutrophils from peripheral blood and B cells from patients with congenital neutropenia identified 1 common gene, while comparison of genes with decreased expression in neutrophils from peripheral blood and B cells from patients with congenital neutropenia identified 7 common genes with reduced expression relative to control samples (Figure 46). This effectively indicated the identification of common transcriptomic features of neutrophils and B cells in patients with congenital neutropenia.

Some of the known PID genes also showed significant differences in expression. In the GSE6233 dataset, 10 genes had increased expression and 7 had decreased expression. In the GSE142347 dataset, 3 genes were increased and 18 were decreased in expression. The genes with increased expression in GSE6233 were LAMTOR2, SmarCD2, CD81, ZBTB24, ACTB, CASP10, APOL1, PARN, ITGB2, and IRF3. The genes with increased expression in the GSE142347 dataset were SEC61A1, MASP2, and RAD51. Among the genes with decreased expression in the GSE6233 dataset were SEC61A1, MTHFD1, STIM1, EXTL3, TGFBR1,

CEBPE, and HAX1. The genes with decreased expression in the GSE142347 dataset were PTPRC, RAC2, BRCA1, PRF1, FCGR3A, ACTB, COPA, IL2RG, MSN, IKZF1, KDM6A, CD55, AP1S3, NFKB1, WDR1, JAK1, IFIH1, and RAD51C. All known and candidate genes for congenital neutropenia, except CXCR4, LAMTOR2, STAT1, and STAT2, were almost more highly expressed in patients with congenital neutropenia than in control samples (Figure 47).

Figure 46. Venn diagrams of overlapping highly and lowly expressed genes in congenital neutropenia in different cell types

We conducted a quality control study by examining some studies after identifying new candidate genes. In particular, ten candidate genes for congenital neutropenia (STAT1, STAT2, STAT3, STAT5B, LYN, FGR, SRC, PIK3CG, ITK, VAV1, CDC42) that were not included in our initial list of congenital neutropenia genes obtained from ESID, but predicted by us, were found in clinical cases of congenital neutropenia. This demonstrates the significance of the identified candidate genes for congenital neutropenia (Table 13).

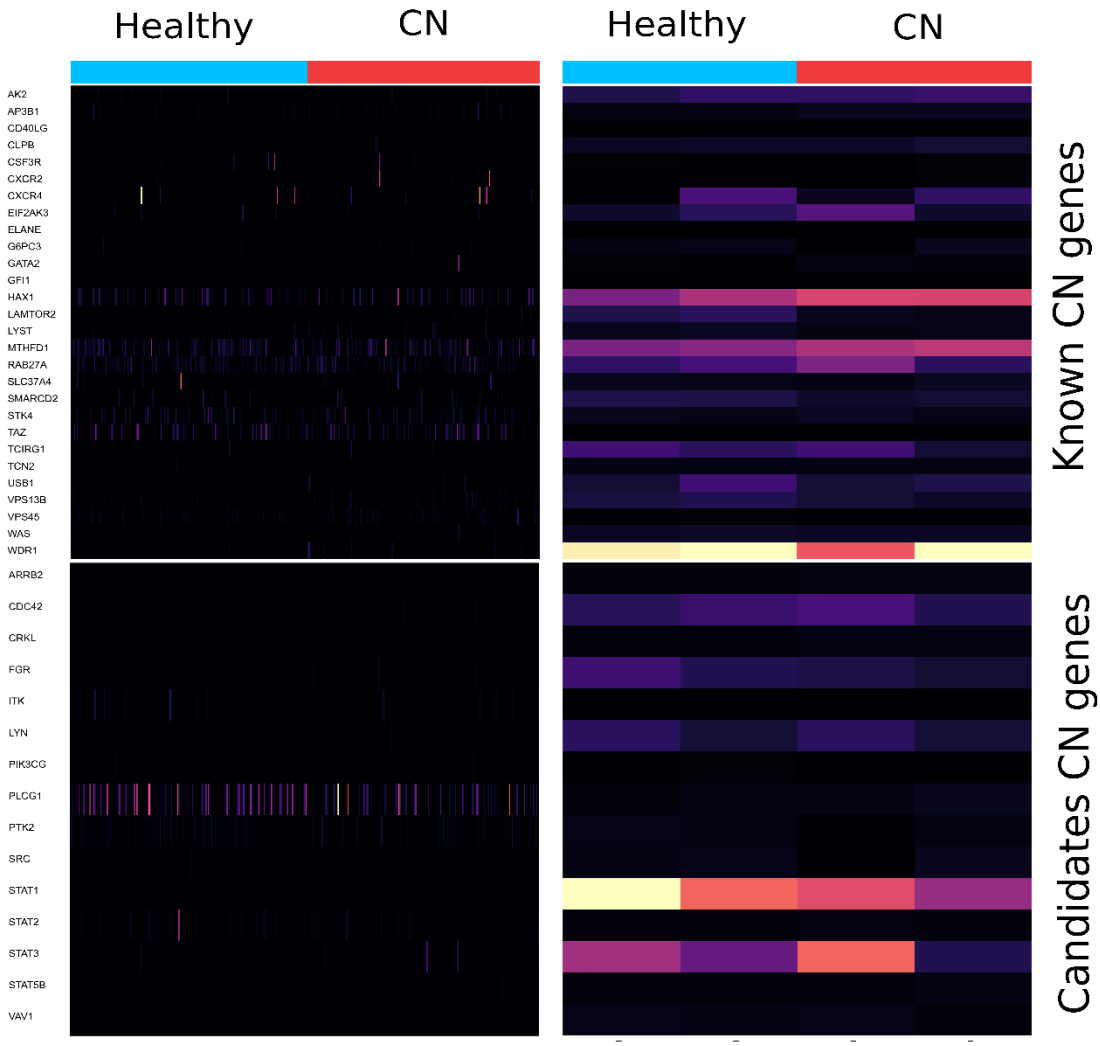


Figure 47 - Heatmap of differentially expressed known and candidate genes in congenital neutropenia

Note: High expression is shown in light yellow color, low expression is indicated by black-purple color.

Table 13 - Candidate genes for congenital neutropenia with recently reported associations with congenital neutropenia

Gene Symbol	Description	Common pathways and main role	Main effect on neutrophils due the	Ref.
CDC42	Cell Division Cycle 42 protein, responsible for cell morphology, cell cycle, and in particularly actin polymerization in N-WASP	ATP-binding component of the Arp2/3 complex through the WASP	Actin polimerisation and phagocytosis	PMID: 19082760 PMID: 21178275 PMID: 10360578 PMID: 34425130
CRKL	Crk Like Proto-Oncogene, Adaptor Protein. CrkL binds to WASP protein	ATP-binding component of the Arp2/3 complex through the WAVE	Actin polimerisation and phagocytosis	PMID: 11313252 PMID: 22837718 PMID: 12504004 PMID: 23934128
FGR	Src family of protein tyrosine kinase	Tyrosine Kinases / Adaptors and Regulatio n of actin dynamics for phagocytic cup formation.	G-CSF Neutrophil regulation	PMID: 1895577 PMID: 8634424

Continue table 13

SRC	proto-oncogene tyrosine-protein kinase Src	Cytokine Signaling in Immune system and PEDF Induced Signaling.	G-CSF	PMID: 16772601
LYN	Src Family Tyrosine Kinase involved in the regulation of cell activation	Antigen sygnaling transduction	Initiation of the B-cell response, B-cell differentiation	PMID: 10643150 PMID: 23001182 PMID: 19201855
PLCG1	phospholipase C gamma 1, plays an important role in the intracellular transduction of receptor-mediated tyrosine kinase activators, has role in neutrophil extracellular trap formation	NGF Pathway and CCR5 Pathway in Macrophages.		PMID: 29543328 (?)
ARRB2	Arrestin beta 2	Cytokine Signaling in Immune system and Tyrosine Kinases / Adaptors.	IL8-mediated granule release in neutrophils	PMID: 24657625
PIK3CG	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Gamma	NF-kappaB Pathway, Immune response CCR3 signaling in eosinophils.	Together with PIK3CD participates in neutrophil respiratory burst. Together with PIK3CD is involved in neutrophil chemotaxis and extravasation	PMID: 29233821 PMID: 29191916 PMID: 31964785
PTK2	Protein Tyrosine Kinase 2	NF-kappaB Pathway and Cytokine Signaling in Immune system	-	A question about glycogen storage disease 1b
STAT1	Signal transducer and activator of transcription 1	Peginterferon alpha-2a/Peginterferon alpha-2b Pathway (Hepatocyte), Pharmacodynamics Cytokine Signaling in Immune system	Role in immune responses	PMID: 27879260 PMID: 29202461 PMID: 33344614 PMID: 27222657
STAT2	Signal transducer and activator of transcription 2	Peginterferon alpha-2a/Peginterferon alpha-2b Pathway (Hepatocyte), Pharmacodynamics Immune response IFN gamma signaling pathway	Act as transcription activators	PMID: 27881648 PMID: 27713294
STAT3	Signal transducer and activator of transcription 3 (acute-phase response factor)	Cytokine Signaling in Immune system IL-4 Signaling Pathways.	G-CSF, Maturation of immune system cells, especially T cells and B cells	PMID: 29330115 PMID: 28253502
STAT5B	Signal Transducer And Activator Of Transcription 5b	Cytokine Signaling in Immune system and IL-4 Signaling Pathways.	Granulocyties differentiation	PMID: 29160632 PMID: 33255665 PMID: 24512550 PMID: 31585621
VAV1	Vav Guanine Nucleotide Exchange Factor 1	Cytokine Signaling in Immune system and Developme nt Dopamine D2 receptor transactivation of EGFR	Cell differentiation T-cell and B-cell development and activation	PMID: 12874226 PMID: 31456807 PMID: 10879282
ITK	IL2 Inducible T Cell Kinase	Tyrosine Kinases / Adaptors and T-Cell Receptor and Co-stimulatory Signaling.	Regulates the development, function and differentiation of conventional T-cells and nonconventional NKT-cells	PMID: 32306816 PMID: 34365077 PMID: 34368657 PMID: 33007409 PMID: 32049330

In conclusion, of this cheapter, 15 candidate genes for congenital neutropenia have been identified that may influence neutrophil functions: STAT1, STAT2, STAT3, STAT5B, LYN, FGR, SRC, PIK3CG, ITK, VAV1, CDC42, PTK2, CRKL, PLCG1, ARRB2. The identified missense variants for TCIRG1 and Elane gene contain scientific and clinincal importans.

Publication of work published by 4th chapter

1. Novel Disease-Associated Missense Single-Nucleotide Polymorphisms Variants Predication by Algorithms Tools and Molecular Dynamics Simulation of Human TCIRG1 Gene Causing Congenital Neutropenia and Osteopetrosis / K. Shinwari, H.M. Rehman, G. Liu, M.A. Bolkov, I.A. Tuzankina, V.A. Chereshnev // Front. Mol. Biosci. 2022. 9. 879875. (WoS Q2, Scopus Q1).

2. In Silico Analysis Revealed Five Novel High-Risk Single-Nucleotide Polymorphisms (rs200384291, rs201163886, rs193141883, rs201139487, and rs201723157) in ELANE Gene Causing Autosomal Dominant Severe Congenital Neutropenia 1 and Cyclic Hematopoiesis / K. Shinwari, M.A. Bolkov, M. Yasir Akbar, L. Guojun, S.S. Deryabina, I.A. Tuzankina, V.A. Chereshnev // Scientific World Journal. 2022. V. 2022. 3356835. (Scopus Q1).

CHAPTER 5 - IDENTIFICATION OF NEW MISSENSE MUTATIONS IN THE CCBE1, FAT4, AND ADAMTS3 GENES LEADING TO HENNEKAM SYNDROME

The aim of the study was to investigate the potential pathogenicity of novel missense substitutions in the CCBE1, FAT4, and ADAMTS3 genes identified in the NCBI dbSNP databases, as well as single nucleotide non-synonymous substitutions in the FAT4 gene found in a patient diagnosed with Hennekam syndrome, on the structure and function of the proteins. We then selected the most probable deleterious substitutions in these genes and assessed their impact on protein structure and function by incorporating the substitutions into the wild-type protein structure using molecular dynamics simulations.

5.1 - Identification of deleterious nsSNPs in the FAT4, ADAMTS3, and CCBE1 genes leading to the development of Hennekam syndrome

In total, 407 nsSNPs in the CCBE1 gene were assessed for their impact on protein structure and function. Of the 407 nsSNPs, 23 were identified as deleterious by both SIFT and PolyPhen-2 programs. Information on the minor allele frequency (MAF) was available for 11 nsSNPs. With the exception of T153N, G107D, P249S, S19N, C75S, C102S, G327R, C174R, D397Y, R125W, P87W, and G330E, the calculated frequency of other nsSNPs in the population was less than 1% (Table 14). Subsequently, all 23 selected nsSNPs were analyzed using an additional 16 bioinformatics tools for predicting the deleteriousness of substitutions on protein structure and function (Table 15, Figure 48).

Table 14 - Non-synonymous single nucleotide substitutions in the CCBE1 gene assessed by SIFT and PolyPhen2 as deleterious

nsSNP	A.A	SIFT	Score	PolyPhen-2	Score	MAF
rs199902030	D336N	Del	0.003	Prob damage	1	< 0.001 (T)
rs200149541	T153N	Del	0.001	Prob damage	1	

Continue table 14

nsSNP	A.A	SIFT	Score	PolyPhen-2	Score	MAF
rs372499913	G107D	Del	0	Prob damage	1	
rs267605221	P249S	Del	0.007	Prob damage	1	
rs374941368	S19N	Del	0.004	Prob damage	0.981	
rs375717418	R301W	Del	0.004	Prob damage	1	< 0.001 (T)
rs80008675	D41E	Del L	0.016	Prob damage	0.982	0.017 (T)
rs116596858	P181S	Del L	0.007	Prob damage	0.906	< 0.001 (A)
rs116675104	R167W	Del L	0.017	Prob damage	0.990	0.003 (A)
rs121908250	C75S	Del L	0.002	Prob damage	0.981	
rs121908251	C102S	Del L	0	Prob damage	0.999	
rs121908252	G327R	Del	0	Prob damage	1	
rs121908254	C174R	Del	0.001	Prob damage	0.984	
rs147974432	T144M	Del L	0.002	Prob damage	1	< 0.001 (A)
rs192224843	Q353R	Del	0.011	Prob damage	0.993	< 0.001 (C)
rs115982879	R118L	Del L	0.001	Prob damage	0.910	< 0.001 (T)
rs139059968	K355T	Del	0.002	Prob damage	0.883	< 0.001 (G)
rs141125426	D397Y	Del L	0.002	Prob damage	0.828	
rs147208835	R125W	Del L	0	Prob damage	0.995	
rs147681552	P290L	Del	0.005	Prob damage	1	< 0.001 (A)
rs148498685	P87S	Del L	0.002	Prob damage	1	
rs149531418	G330E	Del	0	Prob damage	0.999	
rs149792489	A96G	Del L	0.004	Prob damage	1	< 0.001 (C)

Note. Substitution - amino acid substitution in a protein; Del - damaging substitution, Del L - likely less damaging substitution, Prob damage - probably damaging substitution. MAF - minor allele frequency.

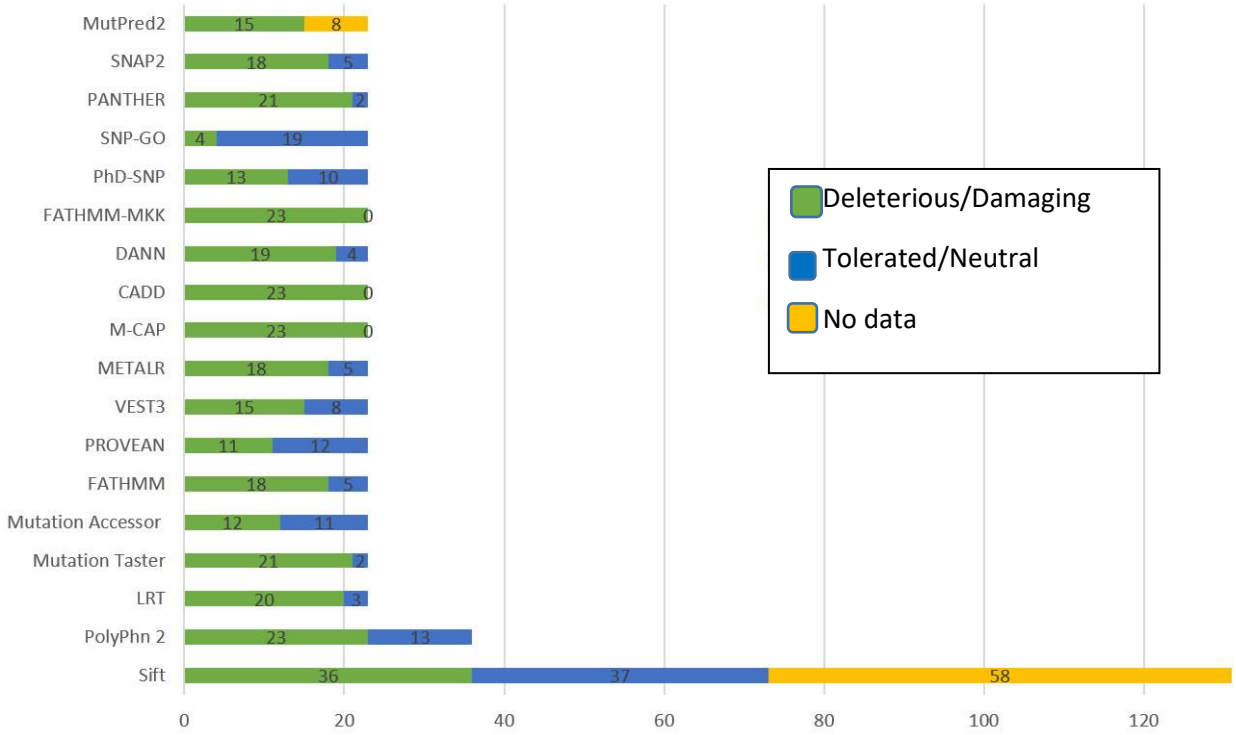


Figure 48 - Results of predicting the effects of identified nsSNPs in the CCBE1 gene analyzed by eighteen computational tools

For the ADAMTS3 gene, 919 nsSNPs were tested. Only 50 out of the 919 nsSNPs were selected by the SIFT program as fully deleterious, and these 50 were then analyzed by several tools (Figure 48). Information on minor allele frequency (MAF) is available for 16 of them, while MAF of other nsSNPs may be less than 1%.

The visual representation of the results of filtering ADAMTS3 gene nsSNPs through 19 bioinformatics tools for predicting the pathogenicity of substitutions (including SIFT and Polyphen-2) is presented in Figure 49. All prediction methods provided statistically significant results. The p-value for the Student's t-test was 0.001 for all tools.

Table 15 - Verification of pathogenicity of 23 identified nsSNPs in the CCBE1 gene by other tools

A.A	LRT	Mutation Taster	Mutation Accessor	PROVEAN	FATHMM	VEST3	MetaLR	M-CAP	CADD	DANN	FATHMM-MKK	PhD-SNP	PANTHER	SNP-GO	SNAP2
G330E	D	D	H	D	D	D	D	D	D	D	D	D	D	D	E
C102S	D	D	M	D	D	D	D	D	D	D	D	D	D	D	E
C174R	D	D	H	D	D	D	D	D	D	D	D	D	D	D	E
G107D	D	D	L	D	D	D	D	D	D	D	D	D	D	D	E
R125W	D	D	L	D	D	T	D	D	D	D	D	D	D	N	E
G327R	D	D	H	D	D	D	D	D	D	D	D	N	D	N	E
P290L	D	D	M	D	T	D	D	D	D	D	D	N	D	N	E
K355T	D	D	M	N	D	D	D	D	D	D	D	D	D	N	E
Q353R	D	D	M	N	D	D	D	D	D	D	D	D	D	N	E
D336N	D	D	M	N	D	T	D	D	D	D	D	D	D	N	E
T153N	D	D	M	N	D	T	D	D	D	D	D	D	D	N	E
C75S	D	D	L	D	D	D	D	D	D	T	D	N	D	N	E
P87S	D	D	L	N	D	D	D	D	D	D	D	D	D	N	E
T144M	D	D	L	N	D	D	D	D	D	D	D	N	D	N	E
R118L	D	D	L	D	D	D	T	D	D	D	D	D	D	N	E
D397Y	N	D	M	D	D	T	D	D	D	T	D	D	D	N	E
R301W	D	D	M	D	T	D	T	D	D	D	D	N	D	N	E
P249S	D	D	M	N	T	T	D	D	D	D	D	N	D	N	N
D41E	D	P	L	N	D	T	T	D	D	T	D	D	D	N	N
S19N	N	P	L	N	D	T	D	D	D	D	D	N	D	N	N
R167W	N	D	L	N	D	T	D	D	D	D	D	N	N	N	E
A96G	D	D	L	N	T	D	T	D	D	D	D	N	D	N	N
P181S	N	D	L	N	T	D	T	D	D	T	D	N	N	N	N

Note: A.A – amino acid substitution in the molecule; the following columns – mutation pathogenicity prediction programs. D – damaging substitution, T – tolerant; N – neutral; L – low, M – moderate, H – high probability of pathogenicity; P – pathogenic, E – effect.

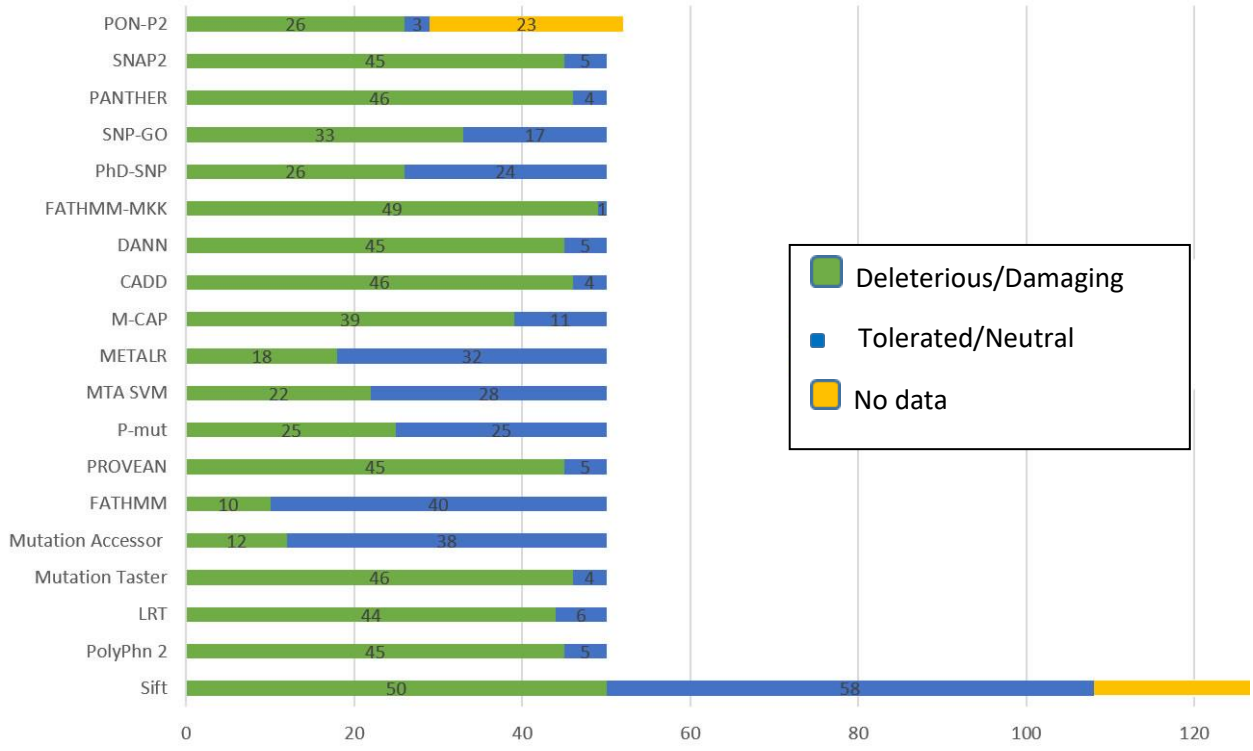


Figure 49 - Results of predicting the impact of identified nsSNPs in the ADAMTS3 gene analyzed by 19 bioinformatics tools for predicting the pathogenicity of mutations (including SIFT and Polyphen-2)

For the FAT4 gene, out of 3434 nsSNPs, SIFT and PolyPhen-2 predicted a total of 298 harmful or damaging nsSNPs. Only 70 nsSNPs had minor allele frequency (MAF) information available. Except for the substitutions G4361, S4710, A785, G1822, D1124N, R2285, R3128, R4726, H1513, S2098, A2959, T1914, V543I, R2285, K294R, and G412, other MAFs for nsSNPs in the FAT4 gene had a value of less than 1%. After applying 18 different bioinformatics tools to predict the pathogenicity of substitutions, only 11 nsSNPs - D2978G, V986D, Y1912C, R4799C, D1022G, G4786R, D2439E, E2426Q, R4643C, N1309I, Y2909H - were considered high-risk polymorphisms that could affect the structure and function of FAT4, even though SIFT considered all 11 substitutions to be damaging with low probability (above >0.5, but below <0.8 on the SIFT scale) (Table 16-17, Figure 50).

Table 16 - Verification of the pathogenicity of 11 identified nsSNPs in the FAT4 gene using other in-silico tools

AAS	Mutation Taster	Mutation Accessor	FATHMM	PROVEAN	VEST3	MTASVM	METALR	M-CAP	CADD	DANN	FATHMM-MKK	PhD-SNP	PANTHER	SNP-GO	SNAP2	P-Mut
D2978G	D	M	T	D	D	D	D	D	D	D	D	D	D	D	D	D
V986D	D	H	D	D	D	D	D	D	D	D	D	D	B	D	D	D
Y1912C	D	M	T	D	D	D	D	D	D	D	D	D	D	D	D	D
R4799C	D	L	D	D	D	D	D	D	D	D	D	D	D	N	D	D
D1022G	D	H	T	D	D	D	D	D	D	D	D	D	D	D	N	D
G4786R	D	L	D	D	D	D	D	D	D	D	D	D	D	N	D	D
D2439E	D	M	T	D	D	D	D	D	D	D	D	D	D	N	D	D
E2426Q	D	H	T	D	D	D	D	D	D	D	D	D	D	D	N	D
R4643C	D	L	D	D	D	D	D	D	D	D	D	D	D	D	D	N
N1309I	D	H	T	D	D	D	D	D	D	D	D	D	D	D	N	D
Y2909H	D	H	T	D	D	D	D	D	D	D	D	D	D	D	N	D

Note. Substitution - amino acid substitution in the protein; D - damaging substitution; N - neutral; L - low, M - medium, H - high probability of pathogenicity; B - benign substitution.

Table 17 - Assessment of the filtered 11 nsSNPs in the FAT4 gene and their minor allele frequency in the population.

nsSNP	AAS	SIFT	Score	PolyPhen-2	Score	MAF
rs147663284	D2978G	Del-Low	0.005	Prob-Damaging	0.99	
rs192514171	V986D	Del-Low	0	Prob-Damaging	1.00	
rs138137489	Y1912C	Del-Low	0.001	Prob-Damaging	1.00	
rs199895179	R4799C	Del-Low	0	Prob-Damaging	1.00	<0.001 (T)
rs372060616	D1022G	Del-Low	0	Prob-Damaging	1.00	
rs138173652	G4786R	Del-Low	0	Prob-Damaging	1.00	<0.001(A)
rs142184187	D2439E	Del-Low	0	Prob-Damaging	0.99	
rs147633644	E2426Q	Del-Low	0	Prob-Damaging	1.00	
rs181607904	R4643C	Del-Low	0	Prob-Damaging	1.00	<0.001 (T)
rs184971791	N1309I	Del-Low	0	Prob-Damaging	0.99	
rs148655455	Y2909H	Del-Low	0	Prob-Damaging	1.00	

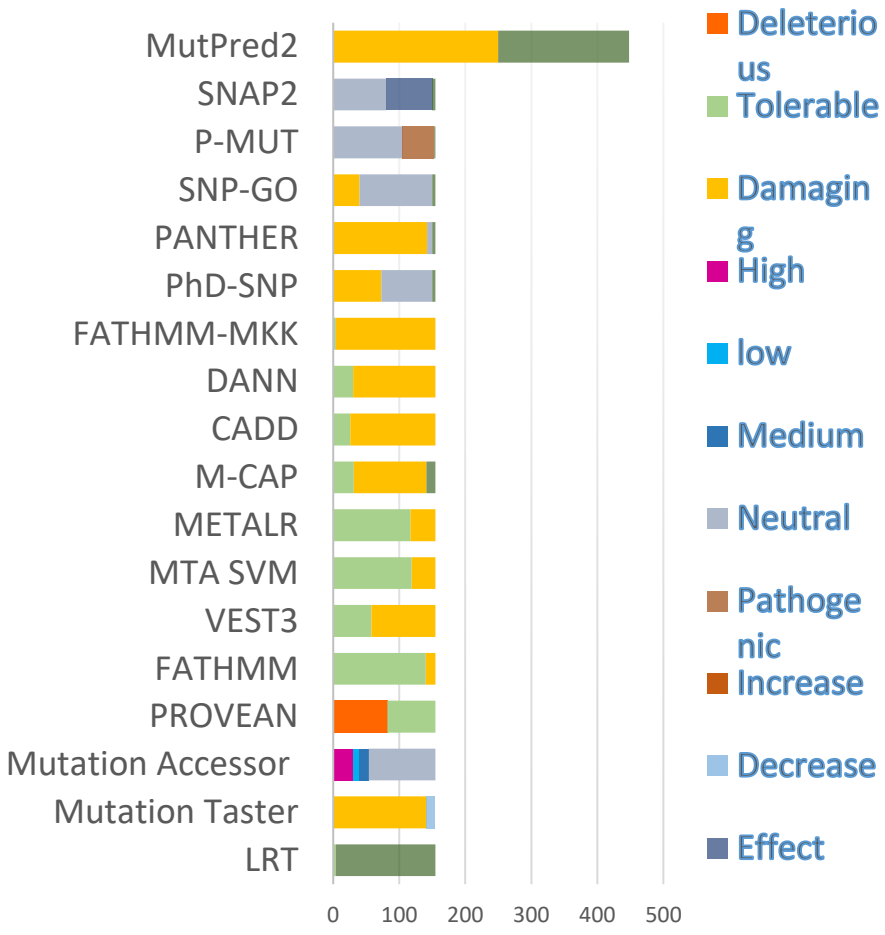


Figure 50 - Results of predicting the impact of identified nsSNPs in the FAT4 gene, analyzed by 18 bioinformatics tools for predicting the pathogenicity of mutations (after filtering by SIFT and Polyphen-2)

5.2 - Prediction of protein stability including non-synonymous substitutions in the genes FAT4, ADAMTS3, and CCBE1, identified in the previous stage

To analyze the stability prediction of CCBE1, the web tool iStable 2.0 was used. This web tool consists of 11 sequence- and structure-based prediction tools, and machine learning approach is used for all results. The results showed that substitutions G330E, C174R, G327R, P290L, D41E, A96G, T114M, D397Y, S19N, and Q359RT increase stability, while amino acid substitutions P249S, R167W, R301W, C75S, P87S, R118L, T153N, D336N, R125W, K355T, G107D, and C102S decrease the stability of the CCBE1 protein. No data could be obtained for the substitution P181S in the iStable 2.0 program.

Table 18 - Predictions of iStable 2.0 on the stability of the CCBE1 protein, taking into account the identified amino acid substitutions

AAS	Realibality score	Impact on Protien
G330E	-0.002680719	Increase
C174R	0.021838337	Increase
C102S	-1.2213084	Decrease
G107D	-0.86388123	Decrease
R125W	-0.85255766	Decrease
G327R	0.0042461157	Decrease
P290L	0.2298831	Decrease
K355T	-0.052274585	Increase
Q353R	0.8725257	Increase
D336N	-1.2082165	Decrease
T153N	-0.546193	Decrease
C75S	-1.0542232	Decrease
P87S	-1.9976869	Decrease
T144M	0.23297998	Increase
R118L	-0.5704589	Decrease
D397Y	0.071232796	Decrease
R301W	-0.3441298	Decrease
P249S	-1.1325055	Decrease
D41E	0.4703572	Increase
S19N	0.77003396	Increase
R167W	-0.4350294	Decrease
A96G	-0.041893244	Increase

The programs I-Mutant 3.0 and MUpro were used to evaluate 50 nsSNPs with high risk of affecting the stability of the ADAMTS3 protein. The protein stability disruption prediction ($\Delta\Delta G$) in I-Mutant 3.0 showed that 47 nsSNPs decrease stability ($\Delta\Delta G < 0$) while 3 nsSNPs increase stability ($\Delta\Delta G > 0$). MUpro identified 48 nsSNPs that individually decrease protein stability. Variants with substitutions of

S1038F, S58F, and D791V (as per I-Mutant) as well as R576L, R954H, and G412S (as per MUpro) were identified as increasing protein stability. Calculations showed that the structure and function of the protein would be disrupted by 19 variations, which included V395I, A336V, G298R, Q616H, Q927H, S1038F, G374S, D815Y, R94L, G983S, Q588H, G25H, R565W, R817C, R713L, R55L, N98S, Y636S, R576L, R1053C, D791V, G412S, and L801F. All of these variants showed ΔΔG values less than -1 kcal/mol as determined by these two tools.

Using the same tools, I-Mutant 3.0 and MUpro (by comparing free energies), the impact of the 11 nsSNPs identified in the FAT4 gene on the stability of the corresponding protein was evaluated (Table 19).

Table 19 - Prediction of the impact of identified amino acid substitutions on the stability of the FAT4 protein (using I-Mutant 3.0 and MUpro)

AAS	Stablity on Protien	AAS	Stablity on Protien
D2978G	Decrease	D2439E	Decrease
V986D	Decrease	E2426Q	No data
Y1912C	Decrease	R4643C	Decrease
R4799C	Decrease	N1309I	Decrease
D1022G	No data	Y2909H	Decrease
G4786R	Decrease		

5.3 - Analysis of the preservation of identified substitutions in conservative regions of CCBE1, ADAMTS3, and FAT4 proteins

An investigation of the effect of 23 substitutions in the CCBE1 gene on the CCBE1 protein using the ConSurf service showed that 13 substitutions were located in highly conserved regions of the protein. Eleven of them (C75S, P87S, P290L, A96G, G107D, R118L, G330E, D336N, R125W, Q353R, and T153N) were predicted to be functional and exposed residues, while the other two, C102S and C174R, were predicted to be buried and structural amino acid residues. The substitution S19N was predicted to be a conservative and buried residue, while the

remaining eight (T144M, R167W, P249S, R301W, G327R, K355T, D397Y, and D41E) were predicted to be exposed amino acid residues (on the surface of the protein). The results are shown in Figure 51.

Figure 51 - Location of amino acid substitutions in the CCBE1 protein with consideration of evolutionary conservation and the location of different protein regions according to the ConSurf service

Note: A value of 1 indicates a highly variable region, while 9 indicates the most evolutionarily conserved region.

A similar study was conducted for ADAMTS3 and the identified 50 nsSNPs. Twenty-six out of the 50 missense variants were identified as located in highly conserved regions. Nineteen out of 26 (Q927R, G298R, C567Y, C567R, Q616H,

R565W, R565Q, P371S, P513T, R248H, T668M, R435H, N98S, R883C, G412S, L801F, S1038F, G983S, R959W) were expected to be functional and exposed residues, while the remaining 7 (I291T, V395I, A336V, G374S, S58F, I287F, and A370T) were expected to be buried and structural residues. In addition, conserved and buried residues were shown to have substitutions at F81L, Y148C, R435H, Y536C, M731T, F777L, R94L, R270H, P510A, R572C, R572H, Q588H, R713L, R817C, R943H, and R954H. Furthermore, eight substitutions were located on the surface (G25V, R55L, P77T, R137W, Y636C, D791V, D815Y, and R1053C). Among the 11 nsSNPs with high risk of pathogenicity, 7 (D1022G, N1309I, D2439E, E2426Q, R4799C, G4786R, and R4643C) were predicted to be functional and exposed residues, while the remaining 3 (V986D, Y1912C, Y2909H) were expected to be buried (Figure 52).

For the FAT4 protein, ConSurf showed that many of the amino acid substitutions previously identified as having a high impact risk on the protein were located in highly conserved regions. Seven out of 11 nsSNPs (at positions 1022, 1309, 2439, 2426, 4799, 4786, and 4643) were expected to be functional and exposed residues, while the remaining were considered to be buried. The substitution of G4786R was considered as structurally buried amino acid residue, while the substitution of G298 affected an exposed region of the protein. Due to its size, no image is provided for the FAT4 protein.

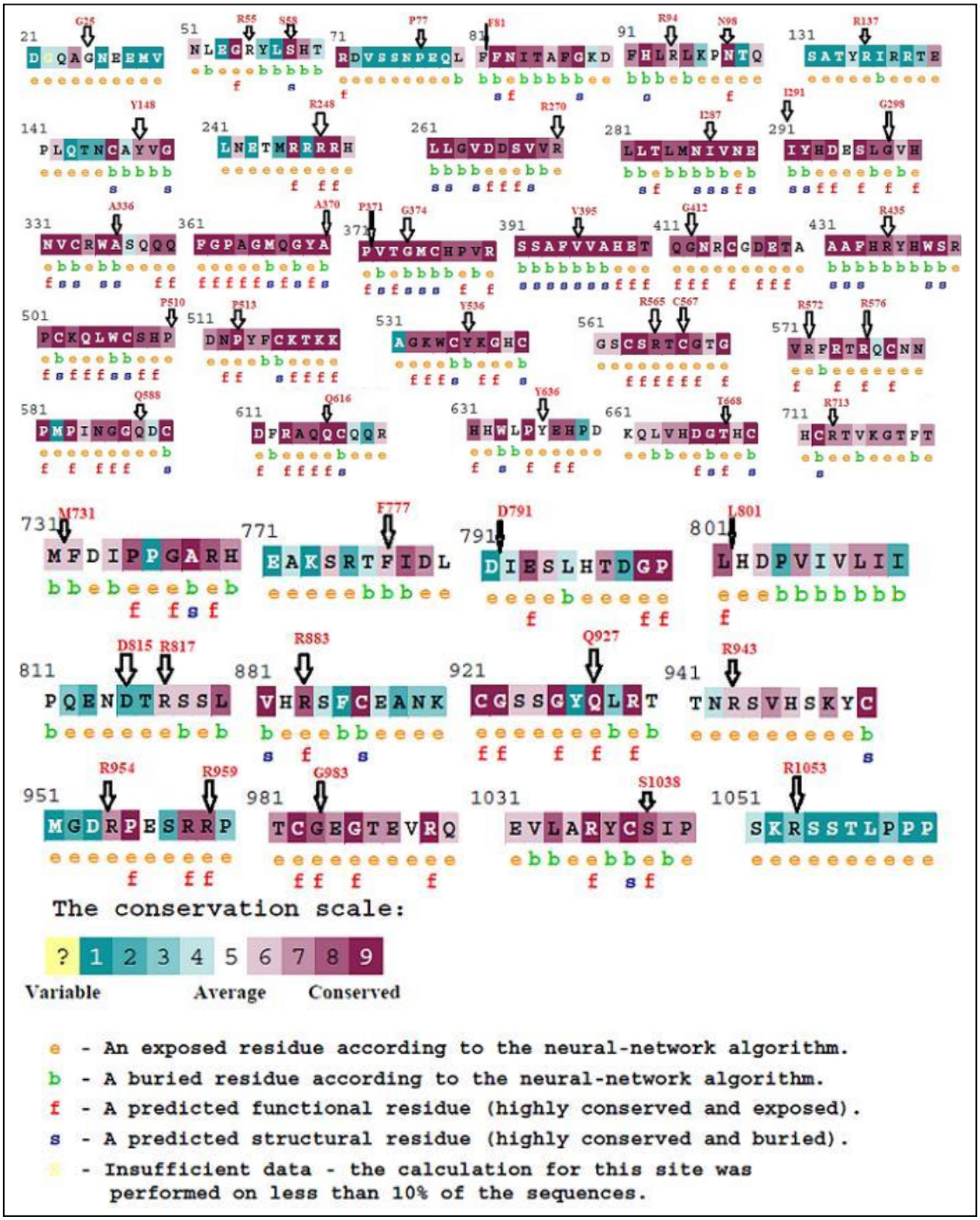


Figure 52 - Location of amino acid substitutions in the ADAMTS3 protein with consideration of evolutionary conservation and location of different protein regions according to the ConSurf service

Note: a value of 1 indicates a high variability region, while 9 indicates the most evolutionarily conserved region.

5.4 - 3D modeling of wild-type and mutant CCBE1 protein structures

To model the 3D structures of the wild-type CCBE1 protein and 22 mutant types, Phyre2 was used to predict the 3D structures of the mutant proteins. The model c5to3B was chosen as a template for predicting the 3D model of CCBE1 in Phyre2. The model for the R118L (rs115982879) mutant showed the greatest deviation, with an RMSD value of 1.56B, followed by A96G (rs149792489), S19N (rs374941368), and C174R (rs121908254) with RMSD values of 1.50B, 1.44B, and 1.46B, respectively. R125W, C75S, and T153N showed RMSD values of 0.89B, 0.90B, and 0.85B, respectively, indicating no structural changes compared to the wild-type. Other amino acid substitutions showed little effect on the 3D structure of CCBE1. These were G327R (1.36B RMSD), P290L (1.3.6B RMSD), Q353T (1.3.2B RMSD), P290L (1.25B RMSD), D336N (1.25B RMSD), C102R (1.22B RMSD), R167W (1.16B RMSD), P87L (1.14B RMSD), G107D (1.13B RMSD), T144M (1.13B RMSD), G330R (1.12B RMSD), D41E (1.12B RMSD), D297Y (1.06B RMSD), R301W (1.02B RMSD), and K355T (1.01B RMSD). The TM coefficients and RMSD values are presented in Table 16. The four nsSNPs (R118L, A96G, S19N, and C174R) with the highest RMSD values were selected and submitted to I-TASSER for remodelling. The protein structure obtained using I-TASSER is the most reliable as it is the most modern modelling tool. Each of these three mutants was studied and superimposed on the wild-type CCBE1 protein using Chimera 1.11, as shown in Figure 53.

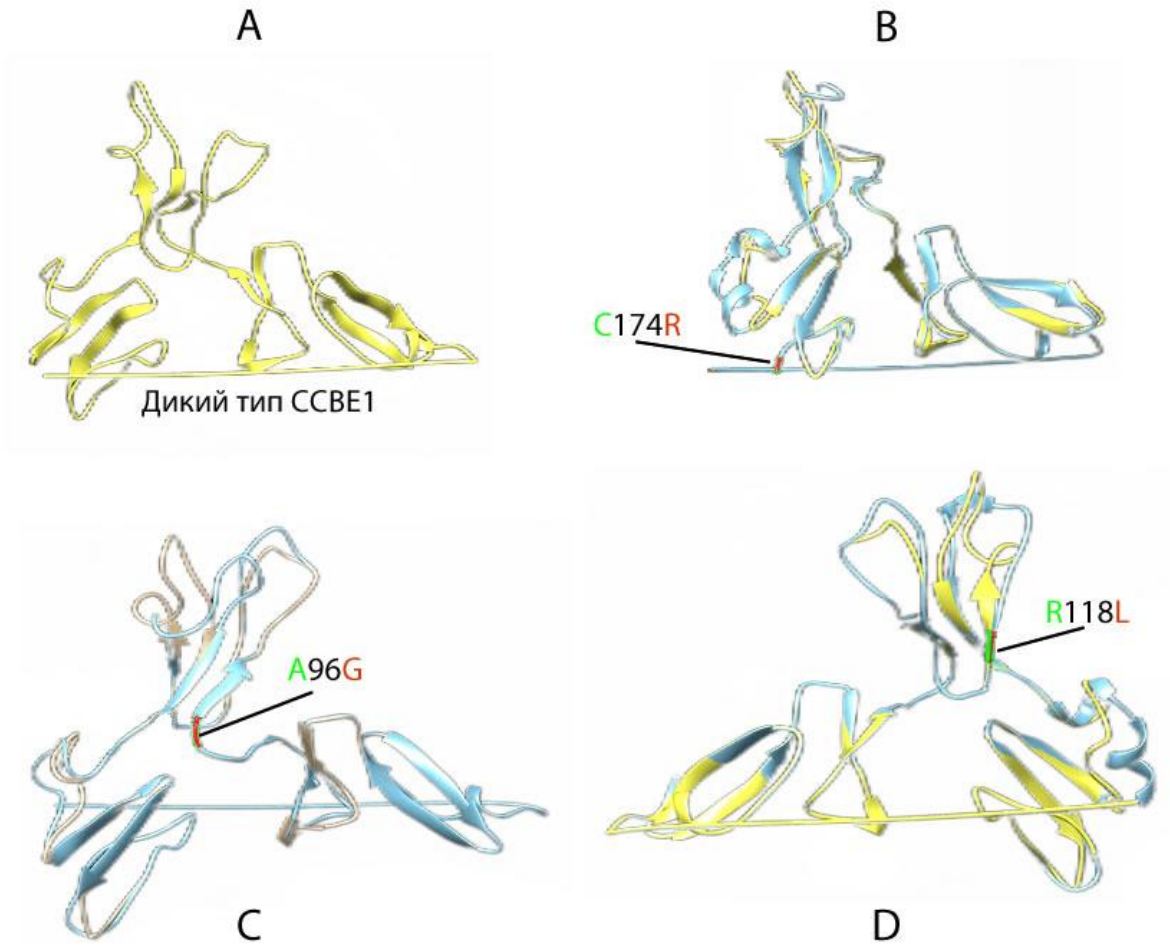


Figure 53 - (A) Structure of wild-type CCBE1 protein. (B) Superimposed structure of CCBE1 and its C174R mutant. (C) Superimposed structure of CCBE1 and its A96G mutant. (D) Superimposed structure of CCBE1 and its R118L mutant. Visualization of Phyre2 model results in Chimera 1.11 program

We conducted an assessment of ligand binding sites using FTSite, i.e., the analysis of protein CCBE1 docking and the evaluation of the impact of identified single nucleotide polymorphisms (SNPs) on docking. Ligand binding sites were predicted using FTSite algorithms, visualized, and further analyzed using PyMOL. With this tool, three ligand binding sites were identified in human CCBE1 protein. Site 1 consisted of 14 amino acid residues; site 2 and site 3 consisted of 7 and 5 residues, respectively. Some of the 22 amino acid substitutions predicted by the SIFT server as potentially deleterious were localized in the presumed ligand binding

sites (T153N and R167W). These results were not further utilized but were published and provided in additional files to the publication on this topic.

A set of various software programs was used to predict post-translational modifications. Specifically, GPS-MSP 3.0 showed the absence of methylation sites in CCBE1. Programs GPS 3.0 and NetPhos 3.1 predicted phosphorylation sites in CCBE1, regions with potential for phosphorylation. BDM-PUB and UbPred were used to predict ubiquitination. In particular, BDM-PUB predicted the ubiquitination of 11 lysine residues. The NetOGlyc4.0 program was used to predict potential glycosylation sites and loss of glycosylation in certain regions due to the described substitutions. These results were not further utilized but were published and provided in additional files to the publication on this topic.

5.5 - 3D Modeling of ADAMTS3 Protein Structures of Wild-Type and Mutant Types

The structures of wild-type and mutant types of ADAMTS3 were predicted using AlphaFold 2. Visualization was performed using Chimera 1.3. In modeling the mutant structure, 25 mutations were included. 21 of these mutations, including S58F, I291T, G298R, A336V, A370T, P371S, G374S, G412S, R435H, Y536C, R565W, C567R, R572C, R576L, Q616H, Y636C, T668M, R883C, R954H, R959W, and G983S, were confirmed by several programs as deleterious (C567Y was not included because it occupies the same position as C567R), and four of them (R138K, R574C, C578L, and Q606H) were found to be clinically relevant. The wild-type and mutant models were validated by Ramachandran plots and analysis of all-atom contacts using the MolProbity program. The wild-type model shows 1032 residues (85.8%) in the favored region, 77 (6.4%) in the allowed region, and 94 (7.8%) in the outlier region, with a total of 1109 residues (92.2%) in the favored and allowed regions. The mutant model shows 1008 residues (83.8%) in the favored region, 110 (9.1%) in the allowed region, and 85 (7.1%) in the outlier region, with a total of 1118 residues (92.9%) in the favored and allowed regions.

For all-atom contact analysis using MolProbity, the wild-type protein showed a score of 3.61, while the mutant protein showed a score of 1.88, which is an acceptable value.

The structures can be divided into three segments (segment 1: Met1-Pro466; segment 2: Lys467-Val831; segment 3: Pro832-Arg1205), which are connected by loops (Figure 54). Segment 3 of both proteins consists mainly of loops without many secondary structures, so we consider it an inaccurate prediction and ignore it for further analysis.

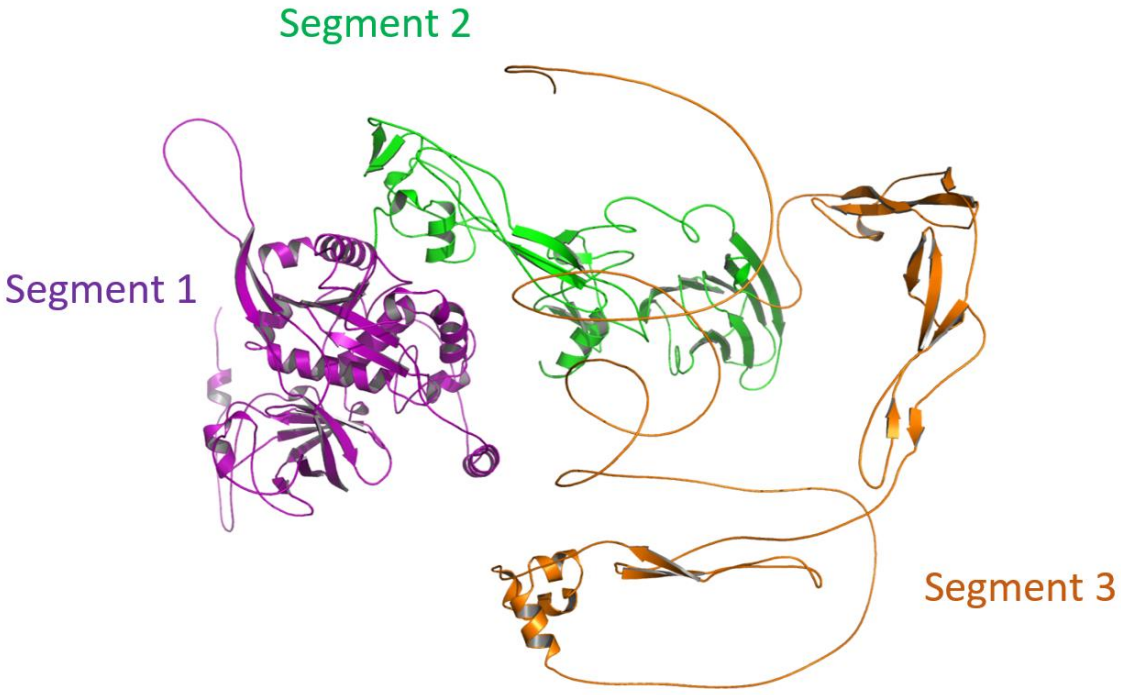


Figure 54 - Segmented structure of the ADAMTS3 protein using the example of the wild-type molecule: segment 1 (residues 1-466), segment 2 (residues 467-831), and segment 3 (residues 832-1205). These three segments are connected by loops. Segment 3 consists mainly of loops (result of the AlphaFold2 model visualization in the Chimera 1.3 software)

We mainly focused on segments 1 and 2, which contain extensive secondary structures. We assume that there are minor interactions between segments, so mutations in one segment will not have a significant impact on the other. The superimposition of the wild type and mutant ADAMTS3 structures (Figure 55) shows an RMSD value of 30.367 Å.

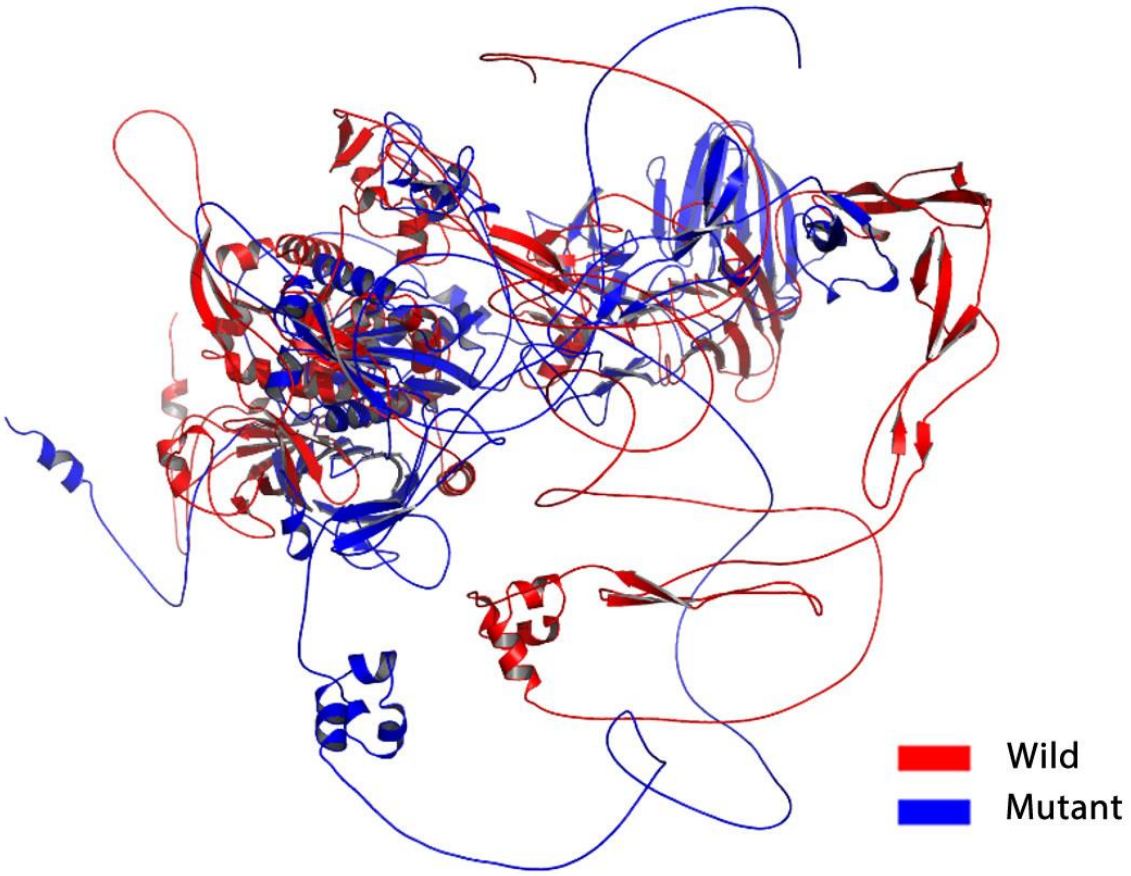


Figure 55 - Overlapping structures of wild type and mutant ADAMTS3 (visualization result of the model from AlphaFold2 in the Chimera 1.3 program)

In order to assess the impact of high-risk pathogenic amino acid substitutions on ligand binding sites, docking analysis of mutant types of ADAMTS3 was performed. Binding sites of ADAMTS3 protein were predicted using the RaptorX Binding server (with a pocket multiplicity value of more than 40) and the COACH ligand binding site prediction server. The RaptorX Binding analysis determined a pocket multiplicity of 151, which is the highest value, and linked it to residues that are subject to G365, M366, Q367, G368, Y369, V395, H398, E399, H402, H408, A426, P427, L428 and V429 substitutions, with the expected Zn2+ cation ligand. The COACH server predicted a Zn2+ cation binding site with a C-score of 0.15 located on residues H398, H402 and H408. Second-ranked sites identified by COACH were associated with Co2+ cation on residues E259, L334, 351, 355 and 356.

Additionally, we studied the effects of each mutation and how they impact neighboring structures. In the Project HOPE, the effects of 50 selected non-synonymous single nucleotide polymorphisms (nsSNPs) in ADAMTS3 on amino acid sizes, charges and hydrophobicity were analyzed. Among these nsSNPs, 26 led to a decrease in amino acid size, while 22 led to an increase. Charge was altered in 23 regions, with 20 changing from positive to neutral, one changing from neutral to positive, and two changing from negative to neutral. Hydrophobicity decreased in seven mutations, while 22 others led to its increase. These results suggest that changes in amino acid properties at these positions may affect the protein structure and its interaction with other molecules, ultimately affecting the protein's function. Local 3D structures of the aforementioned 25 mutations included in AlphaFold protein models were also investigated. The results show that most mutations do not have a significant impact on the sequence structure in the vicinity of the amino acid substitution position in the 25 mutations. Only the Y536C substitution has a significant disruption in the secondary structure compared to other mutations (Figure 56). The remaining 3D structure images altered after amino acid substitutions are presented in the article's appendix and are not included in this dissertation.

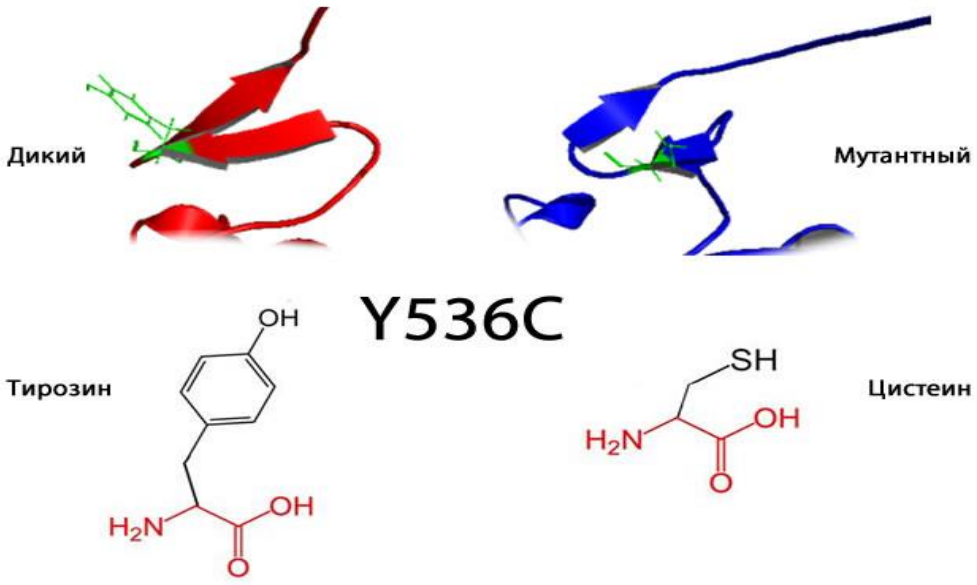


Figure 56 - Alteration in the three-dimensional structure of ADAMTS3 due to the Y536C amino acid substitution. The mutation sites are colored in green. Results from Project HOPE

A calculation of the overall post-translational modifications, including methylation, phosphorylation, ubiquitination, and glycosylation, was also conducted for the wild-type ADAMTS3 structure and mutant. GPS-MSP predicted the absence of methylated sites in ADAMTS3. The predicted serine, threonine, and tyrosine phosphorylation sites by different kinases differ between NetPhos 3.1 and GPS 6.0, with GPS 6.0 predicting more phosphorylation sites than NetPhos 3.1 for both structures. Interestingly, some phosphorylation sites appear and disappear after mutation. GPS 6.0 shows the disappearance of sites on Ser58, Tyr536, Tyr636, and Thr668, and the appearance of new sites on Ile291, Ala370, Pro371, Gly374, Gly412, and Gly983, while NetPhos 3.1 shows the disappearance of sites on Tyr56, Ser58, and Ser957 and the appearance of new sites on Ile291, Pro371, Gly374, Gly412, and Gly983. Most of these changes are in mutation sites involving serine, threonine, and tyrosine. More changes in phosphorylation sites are observed in segment 1. For ubiquitination, UbPred detected 9 lysine residue ubiquitination sites in both the wild-type and mutant structures, while BDM-PUB detected 37 and 36 ubiquitinated lysine residues in the wild-type and mutant proteins, respectively, and after mutation, there are several new and disappeared ubiquitination sites, most of which are in segment 3. Analysis using NetOGlyc4.0 predicted all possible O-glycosylation sites in both proteins, and some mutants lost or gained glycosylation at certain positions, most of which are located in segment 3. These results were not further utilized but are published and provided in additional files accompanying the publication on this topic.

5.6 - 3D modeling of protein structures of wild-type and mutant types of the FAT4 protein

We created five models of the FAT4 protein using I-Tasser and evaluated their quality and the impact of mutations on the structure of the mutant protein. Due to the large size of FAT4, we only modeled those protein sequences where mutations were detected in our clinical case (already published) as well as the most deleterious mutations obtained from the above in-silico study.

Then, using the QMEAN score, Prosa Z-score, and Ramachandran plot analysis, each structure was evaluated for reliability. A more positive QMEAN score indicated the best protein model, while Prosa scanned the structure models and compared them to PDB crystal structures to determine their quality. Based on the QMEAN score, Prosa Z-score, and Ramachandran plot analysis, the selected models were optimized by energy minimization using UCSF Chimera (Figure 57). The 3D models of sequences 2-5 regions are presented in the supplemental materials of this article.

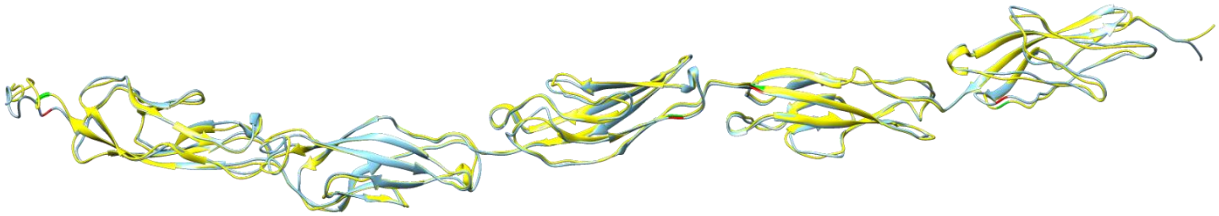


Figure 57 - Overlay of 3D models of FAT4, sequence 1, containing substitutions A807V, V986D, D1022G, and N1309I. Yellow represents the wild-type FAT4, blue represents the mutant variant of FAT4 with red highlighting of the mutations

The consequences of 11 investigated amino acid substitutions in FAT4 on amino acid size, charge, and hydrophobicity were analyzed in Project HOPE. Four mutant amino acids were larger than their wild-type counterparts, while six mutant amino acids were smaller. Charge was altered at eight different sites: 2 from positive to neutral, 1 from neutral to positive, 3 from neutral to negative, and 2 from negative to neutral. Analysis showed that hydrophobicity was decreased in five mutations and increased in four others. These results suggest that amino acid mutations affect protein function by altering protein structure and interaction with other components. The most deleterious nsSNPs, which were found to have a possible model template, were Y1912C (5DZY), D2439E (1L3W), E2426Q (1L3W), D2978G (5W1D), and Y2909H (1L3W). They provide a unique conformation to the central axis of the molecule. However, these results were not further used and were published as supplementary data to the article on this topic.

Before constructing the 3D model of FAT4, we predicted the secondary structure of FAT4 using the SOPMA program, which helped to refine the distribution of alpha helices, beta sheets, and random coils. Analysis of the secondary structure showed the presence of 49.31% random coils (1148), followed by 36.90% extended strands (859), 8.98% alpha helices (209), and 4.81% beta sheets (111). The distribution of amino acid substitutions in secondary structures was not further considered, but the data were published as supplementary material to the article on this topic.

In analyzing the possible impact of amino acid substitutions on post-translational modifications in the FAT4 protein, the GPSMSP 3.0 program did not provide information on methylation in this protein. NetPhos 3.1 predicted a phosphorylation site for 579 residues. The UbPred tool predicted that none of the lysine residues could be ubiquitinated. In contrast, BDMPUB predicted that 101 lysine residues could be ubiquitinated, but none of them were included in the list of analyzed amino acid substitutions. Sites of glycosylation were also evaluated using SUMOylation. These results were not further used, but were published and provided in additional files to the publication on this topic.

5.7 - Molecular Dynamics Modeling of Wild-Type and Mutant ADAMTS3

The change in the root-mean-square deviation (RMSD) values of the Cα atoms of wild-type and mutant ADAMTS3 is presented in Figure 58. For segment 1 of both the wild-type and mutant structures, equilibrium is reached after 130 ns, after which the RMSD values of the two structures do not differ significantly, indicating that mutations in segment 1 do not greatly affect the structure (wild-type: mean 10.830 Å, SD 0.169 Å; mutant: mean 11.109 Å, SD 0.157 Å). However, for segment 2, the wild-type protein reaches stability in just under 10 ns. After this, the system equilibrates, and the modeling converges throughout the entire runtime, but the RMSD values of the mutant protein fluctuate more compared to the wild-type structure throughout the modeling. The mutant structure has a higher RMSD,

indicating that mutations in segment 2 destabilize this part of the protein more (wild-type: mean 5.31 Å, SD 0.344 Å; mutant: mean 14.312 Å, SD 0.584 Å).

The regions of the proteins that fluctuate the most during modeling are shown as peaks on the RMSF graphs (Figure 59). β-sheets and α-helices are often more rigid and less variable than the unstructured component of the protein. In segment 1, although the peak on residues Asn119-Pro129 is higher for the wild-type structure, the RMSF of the wild-type and mutant structures are overall similar. This shows that mutations in segment 1 do not significantly stabilize or destabilize the structure. In segment 2, the overall RMSF of the mutant structure is higher than that of the wild-type, indicating that mutations have destabilized the structure in this segment. There is a large difference in RMSF in residues Met478-Pro523, indicating that this region is the most destabilized.

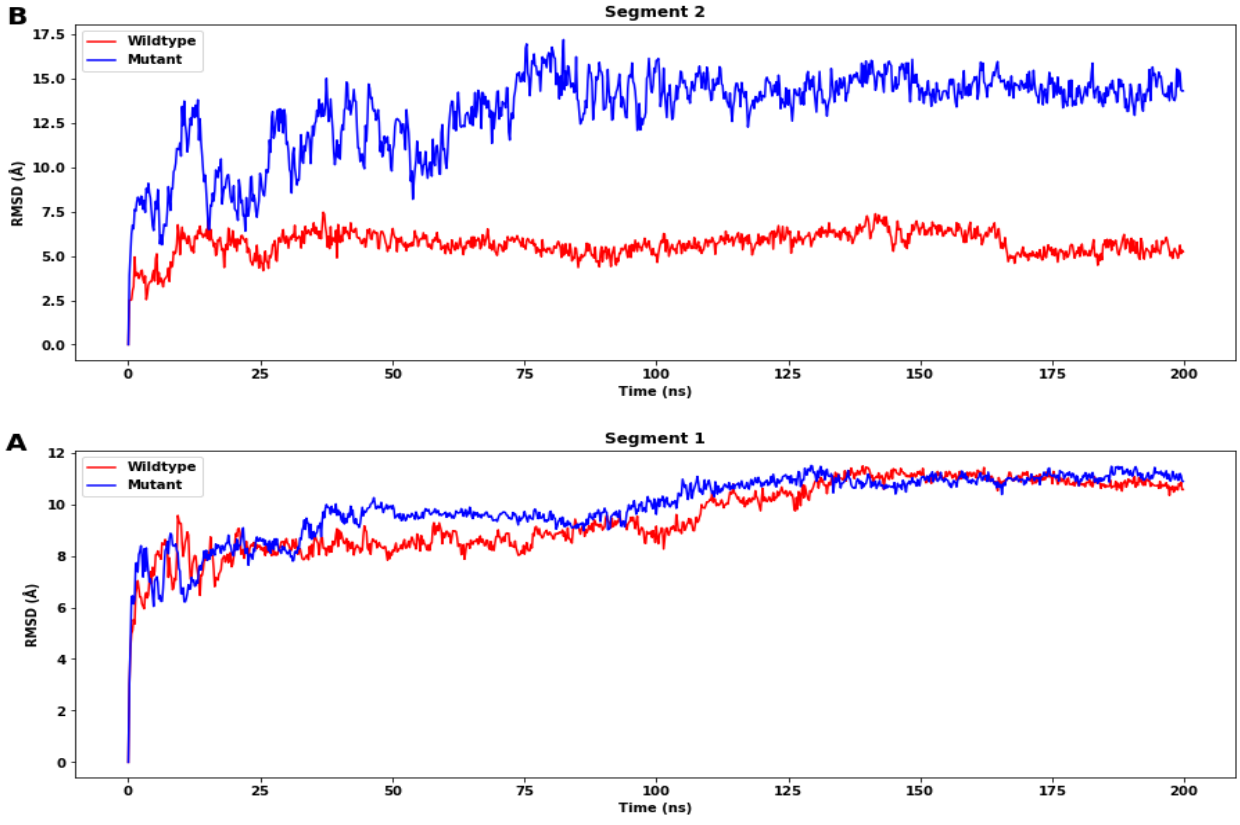


Figure 58 - Root mean square deviation (RMSD) of the Cα atoms of wild-type (red) and mutant (blue) ADAMTS3 protein segments 1 (A) and 2 (B) over time. For segment 1 (A), there is little difference in the equilibrium RMSD between the wild-type and mutant structures. For segment 2 (B), there is a significant difference between the RMSD values. All mean values and SD were calculated from values after 170 ns

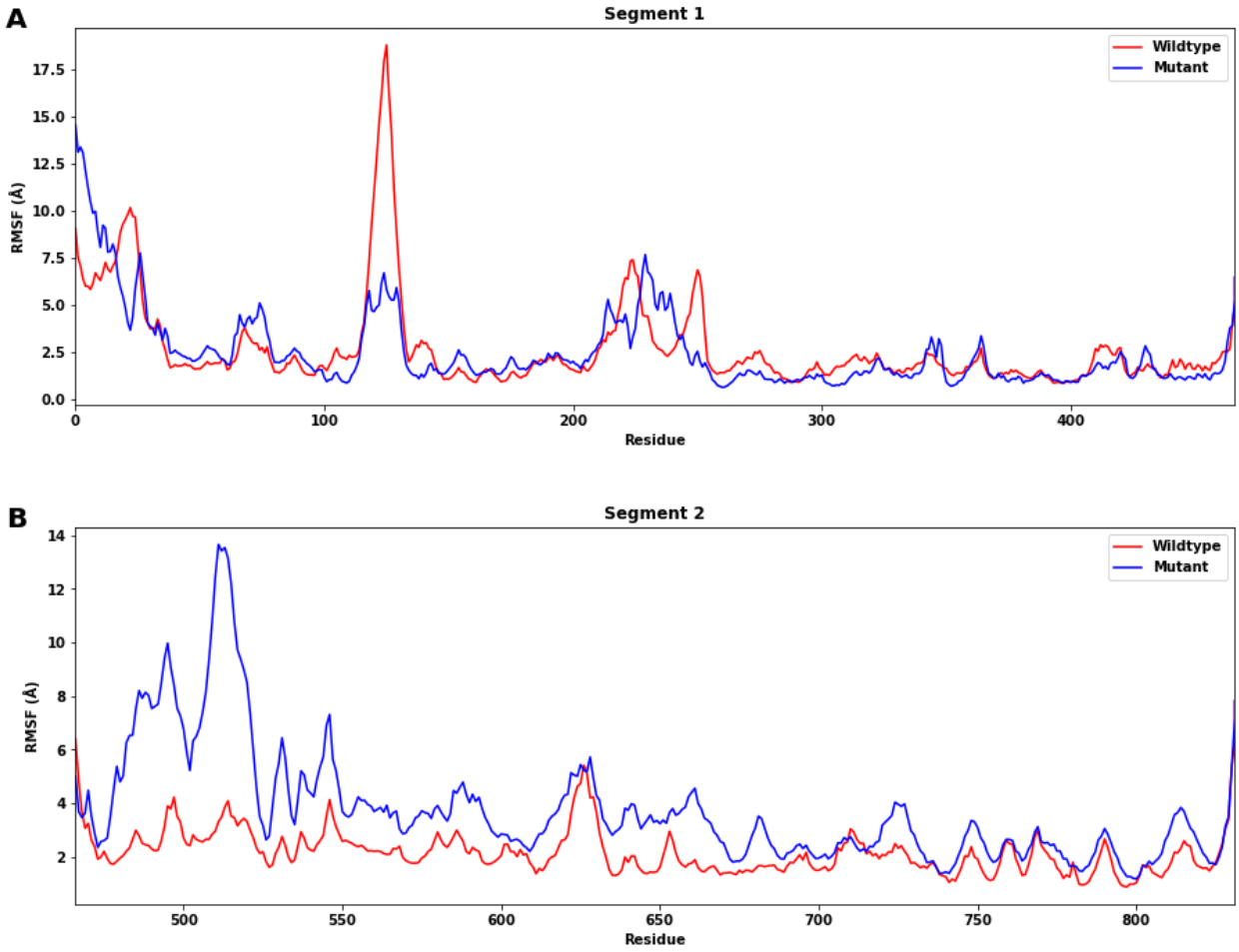


Figure 59 - Root Mean Square Fluctuation (RMSF) for each segment of wild-type (red) and mutant (blue) ADAMTS3 protein. Panel A shows the RMSF for segment 1, and the RMSF values for wild-type and mutant structures are similar

In panel B, the overall RMSF for the mutant structure is higher than that of the wild-type structure for segment 2, indicating that mutations destabilized the structure in this segment. There is a large difference in RMSF for residues Met478-Pro523, indicating that this region is the most destabilized.

In addition, the average distributions of protein secondary structure elements are calculated during the simulation at 170 ns. For segment 1 (Figure 60), the percentage of average α-helix secondary structure decreased by 5.15% in the mutant structure compared to the wild type, but an increase of 3.86% in the percentage of 310-helices was observed, which may stabilize the mutant structure and counteract the destabilizing effect of α-helix disruption by mutations, along with an increase of

1.50% in the percentage of turns. For segment 2 (Figure 61), a decrease of 5.48% in the percentage of β-sheets and an increase of 4.11% in the percentage of turns were observed, which may destabilize the overall structure of the mutant protein in this segment. Additionally, the percentage of α-helices also increased by 1.10%.

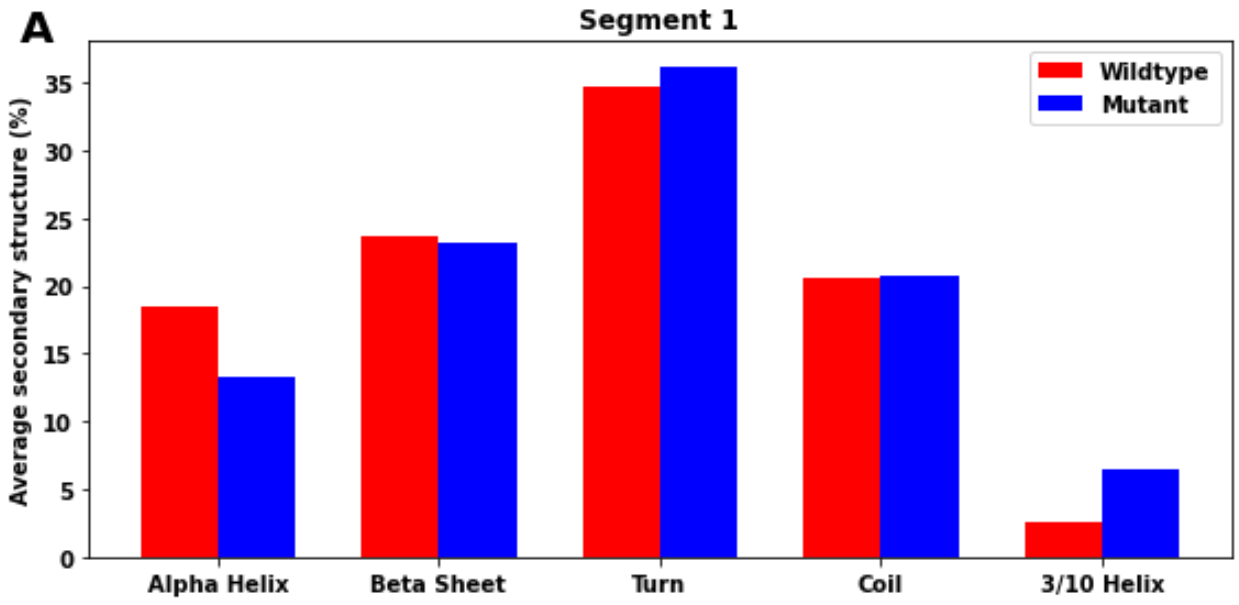


Figure 60 - Distribution of secondary structures in segment 1 after 170 ns of molecular dynamics simulation

Figure 62 shows the distribution of secondary structure elements in segment 1 of both wild-type and mutant ADAMTS3. Examining the distribution of changes in secondary structure in different residues in segment 2 (Figure 63), we observe that β-sheets are disrupted in residues Lys491-Met492, Trp506-His509, and Asn512-Thr518, which may be the cause of the increased RMSFs in residues Met478-Pro523 in the mutant structure.

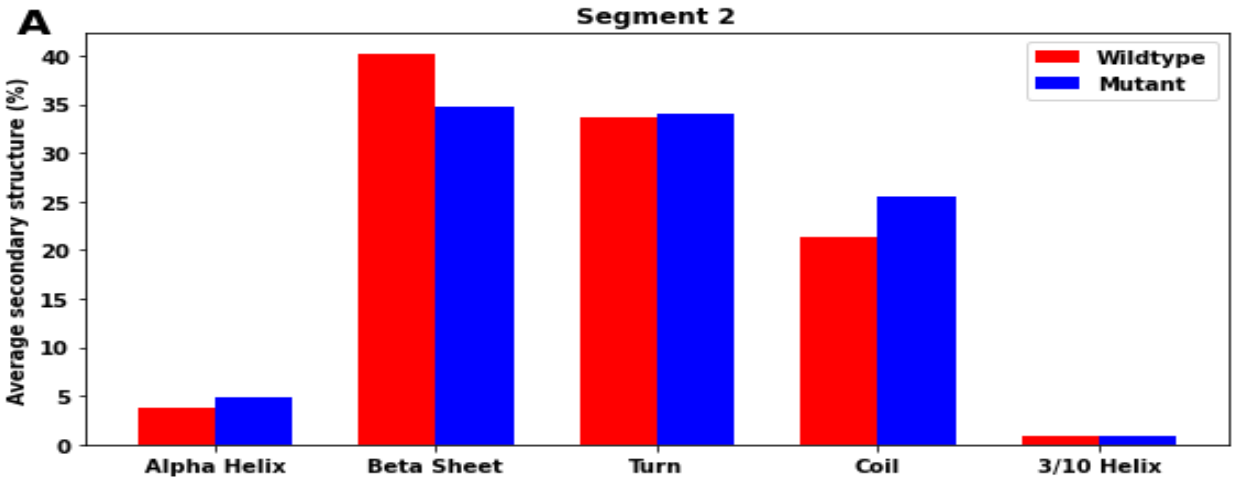


Figure 61 - Distribution of secondary structures in segment 2 after 170 ns of molecular dynamics simulation

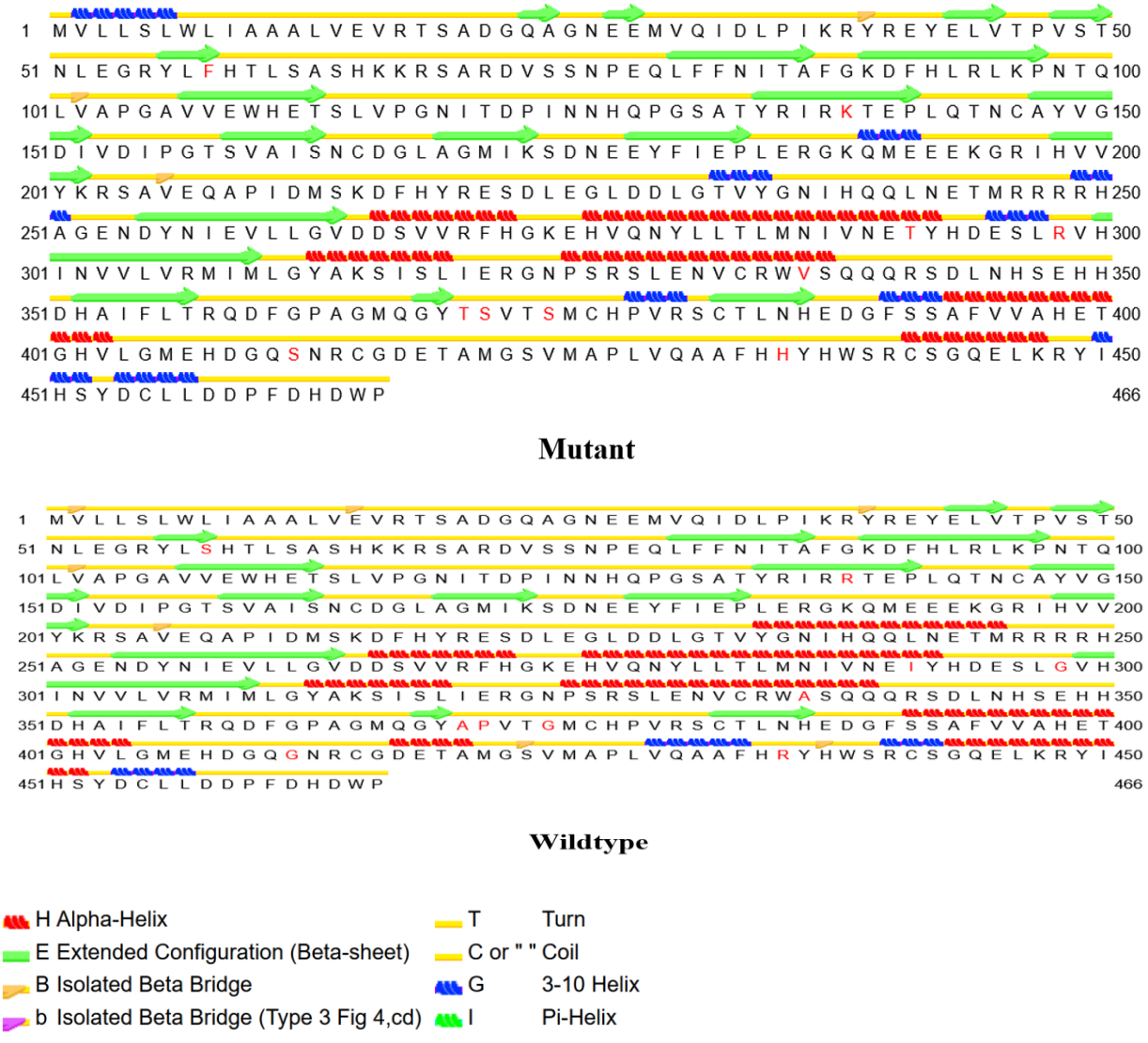


Figure 62 - Analysis of secondary structure in segment 1 after 170 ns of molecular dynamics simulation. Secondary structure elements of wild-type and mutant types of ADAMTS3. Mutated amino acids are marked in red

We also observe that there are no significant changes in the secondary structures near the positions of most of the investigated substitutions. There is no amino acid substitution in the residues of the aforementioned disrupted β-sheets.

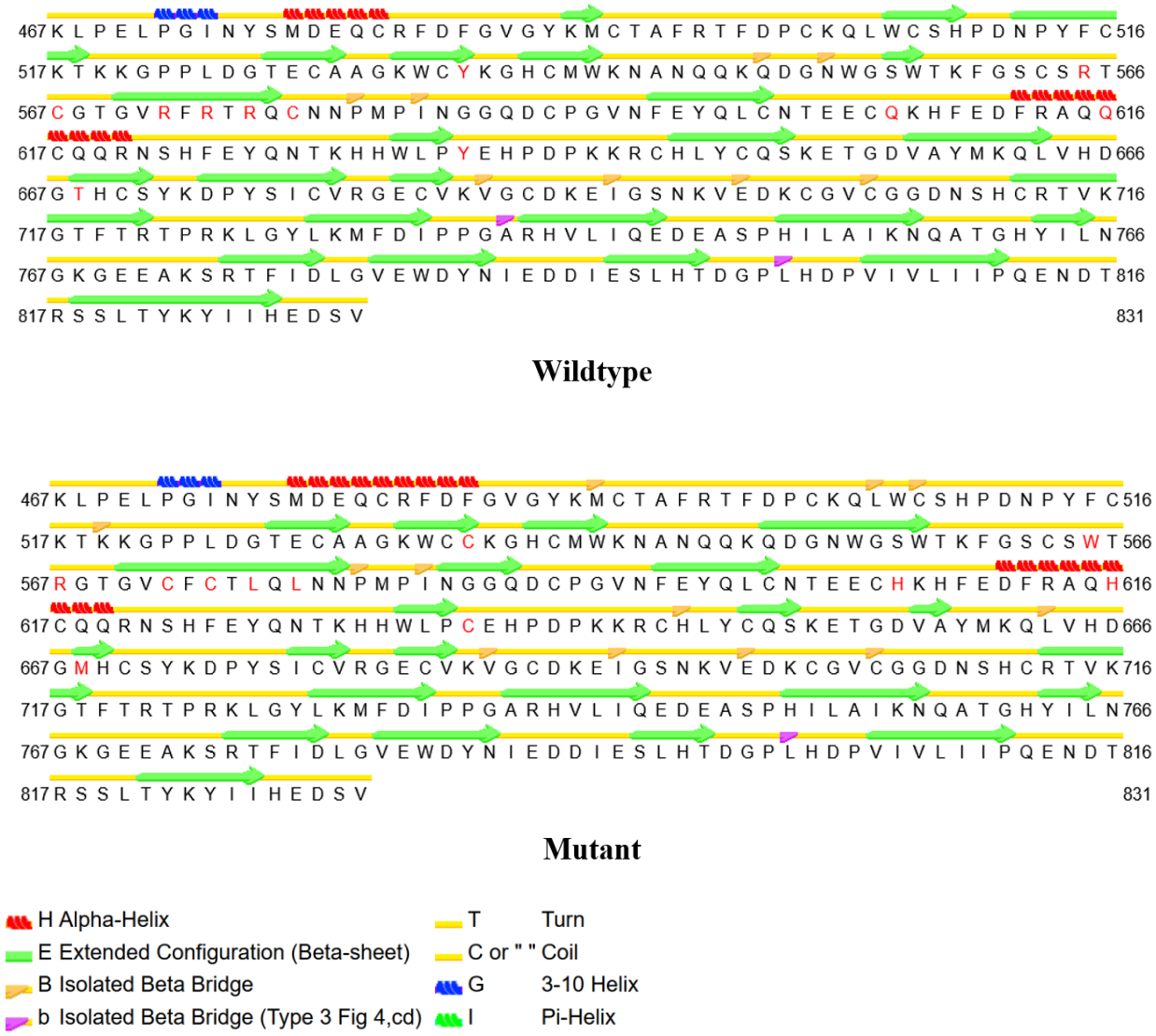


Figure 63 - Analysis of secondary structure in segment 1 after 170 ns of molecular dynamics simulation. Secondary structure elements of wild type and mutant ADAMTS3 are shown. Mutated amino acids are marked in red

Also, the radius of gyration (Rg) analysis is conducted. Two of the most important indicators for determining the structural activity of a macromolecule are Rg determination and calculation of the distance to the center of mass of the molecule. The speed at which the protein folds is proportional to its compactness and can be measured using a complex computer method for calculating the radius of

gyration. From the analysis of the radius of gyration of wild-type and mutant ADAMTS3 structures, it can be observed that the mutant type showed overall higher Rg values throughout the simulation time scale compared to the wild type in segments 1 and 2, but the difference for segment 1 is not as significant (wild type: mean: 23.339 Å, SD: 0.082 Å; mutant: mean: 23.984 Å, SD: 0.139 Å) as for segment 2 (wild type: mean: 27.648 Å, SD: 0.163 Å; mutant: mean: 33.564 Å, SD: 0.402 Å). As a result, the flexibility of the mutant protein is increased (Figure 64).

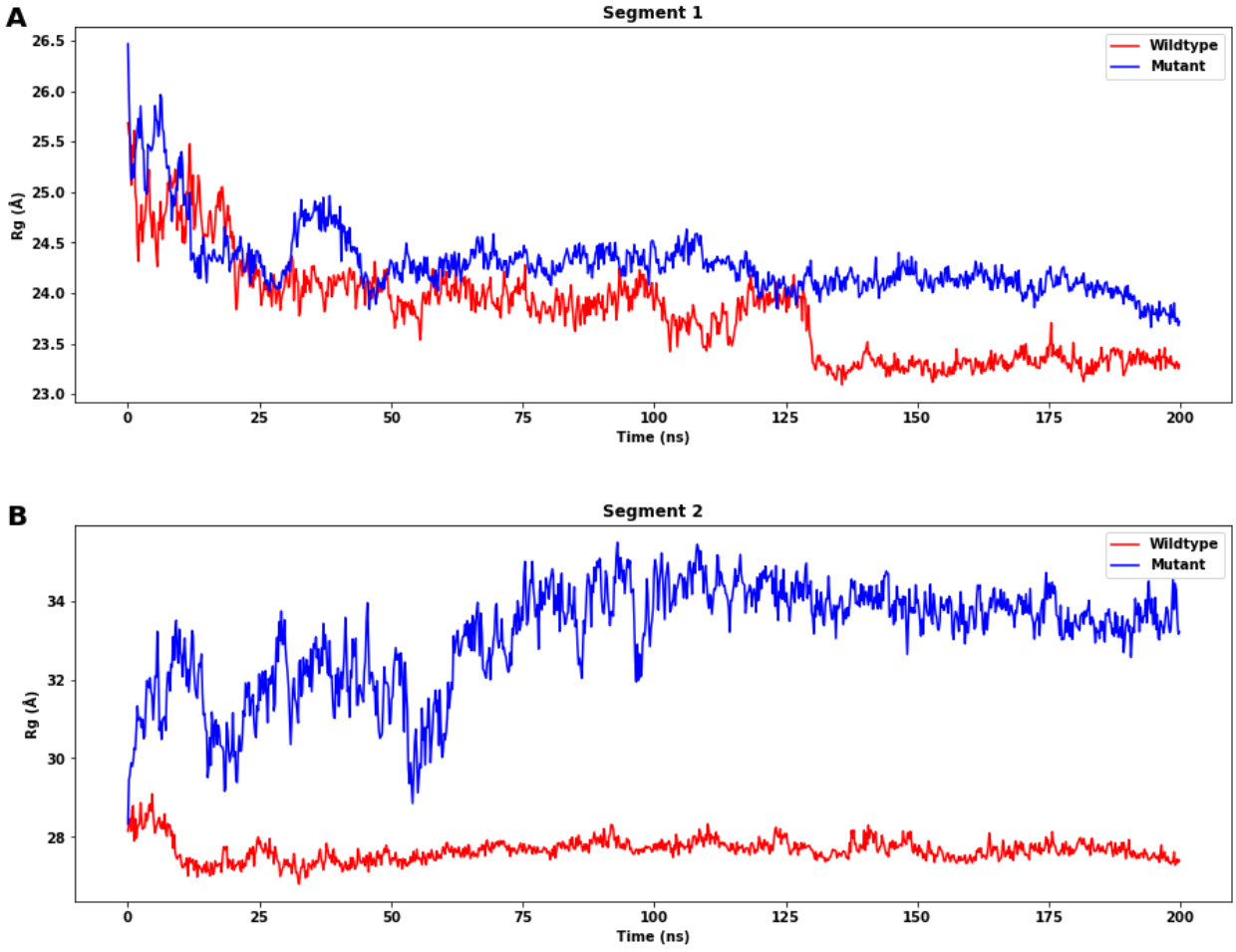


Figure 64 - Radius of gyration of wild-type and mutant protein segments 1 and 2. All mean values and SDs were calculated from values after 170 ns

The analysis of the solvent-accessible surface area (SASA) showed that the mutant structure has a higher SASA value than the wild type for segments 1 and 2 (Figure 65). (Wild type segment 1: mean value: 21906.066 Å2, SD: 282.987 Å2; mutant: mean: 22675.036 Å2, SD: 453.033 Å2; wild type segment 2: mean: 21565.973 Å2, SD: 245.14 Å2; mutant: mean: 22160.942 Å2, SD: 269.095 Å2).

Since a higher SASA value indicates protein expansion, it can be assumed that the wild type is more stable than the mutant protein. The more significant change in SASA value may be due to the amino acid substitution effect, which changes the protein surface size, its hydrophilicity, and other characteristics.

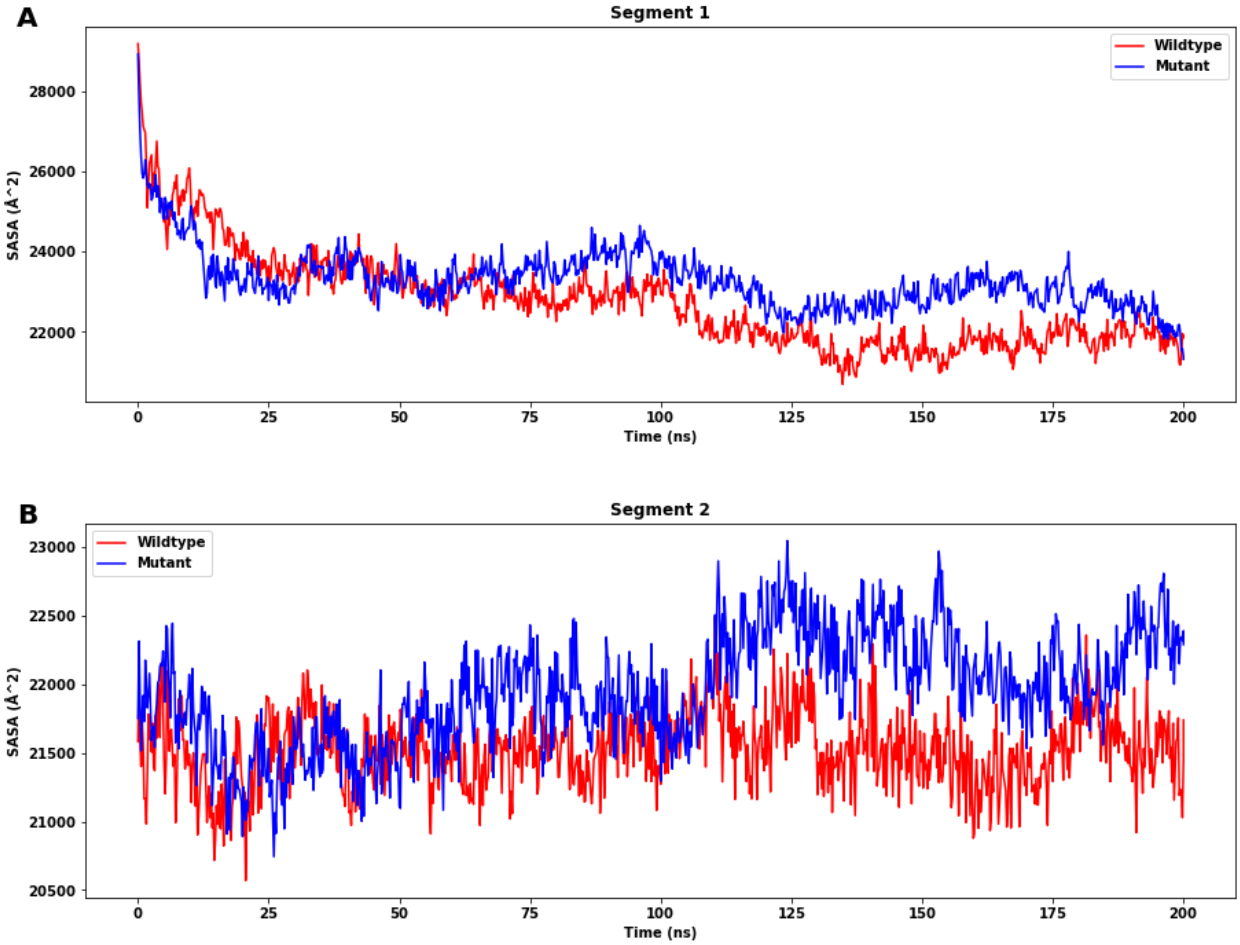


Figure 65 - Solvent accessible surface area (SASA) (in Å2) of wild type and mutant type ADAMTS3 segments 1 (A) and 2 (B)

For both segments, the SASA of the mutant structure is higher than that of the wild type. All mean values and SDs were calculated from values after 170 ns.

During the molecular dynamics simulation, the difference in the number of hydrogen bonds (H-bonds) was also calculated (Figure 66). For segment 1, it is insignificant, which once again indicates that the destabilization effect for substitutions in this segment is small (wild-type: mean: 396.854, SD: 7.427; mutant: mean: 392.351, SD: 9.540). For segment 2, it can be noted that the wild-type structure forms a greater number of H-bonds, while the mutant structure

demonstrates a lower number of H-bonds, which may affect the stability of the mutant protein (wild-type: mean: 263.192, SD: 6.421; mutant: mean: 255.709, SD: 9.473).

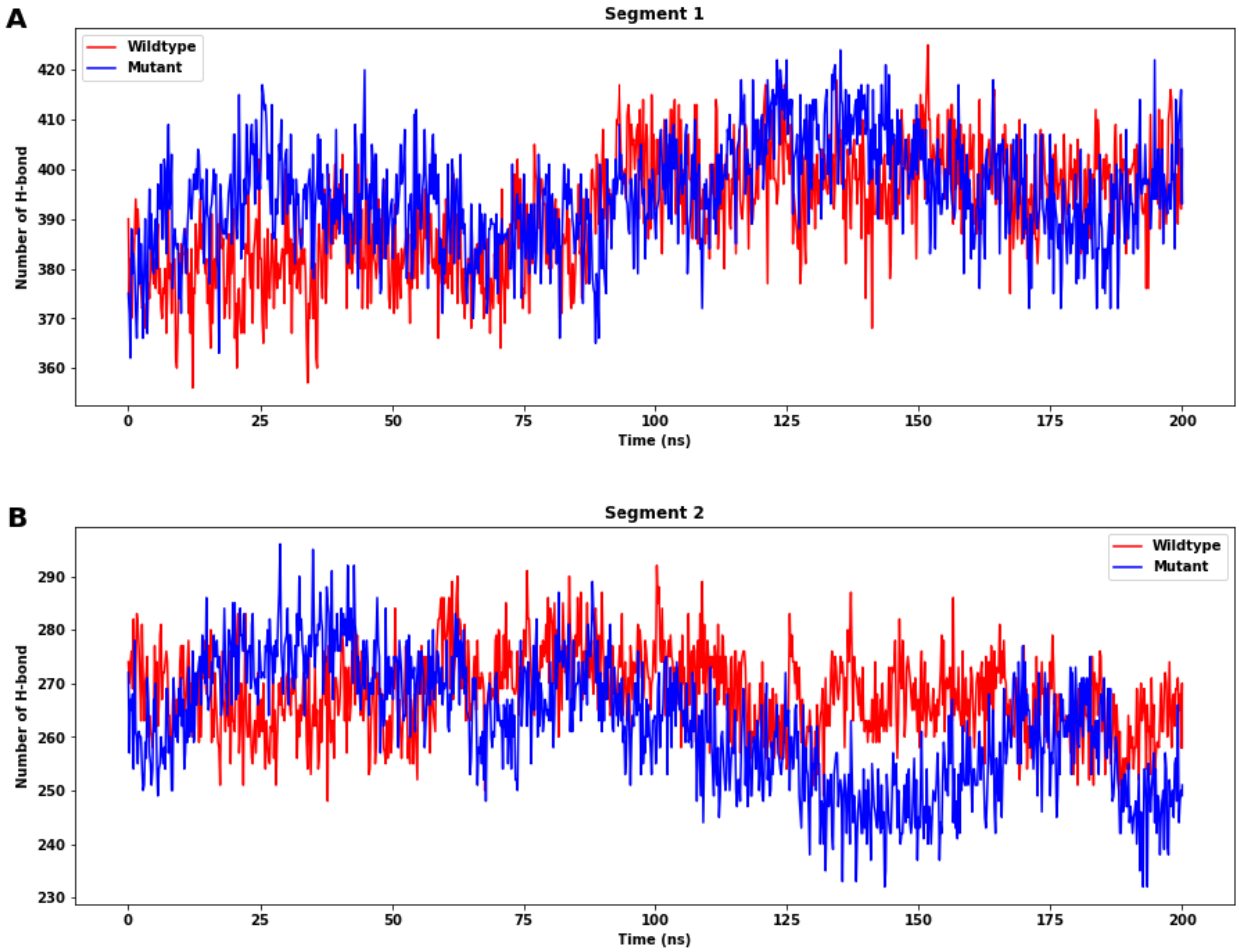


Figure 66 - Total number of hydrogen bonds throughout the simulation of wild-type and mutant ADAMTS3 protein segments 1 and 2. All mean values and SDs were calculated from values after 170 ns

Principal component analysis (PCA) was used in this study to analyze the trajectories and structures of wild-type and mutant ADAMTS3 proteins in segments 1 and 2. The PCA plots show the collective motions of the protein system projected onto the first two principal components.

The plots for segments 1 and 2 (Figure 67) indicate a significant difference in the trajectories and motions of wild-type and mutant protein systems. In segment 1, the plots for wild-type and mutant structures largely overlap, indicating that mutations in this segment have a minor impact on the collective motions of the

protein. However, in segment 2, there is less overlap between the plots for wild-type and mutant structures, indicating that mutations in this segment have a greater impact on the collective motions of the protein.

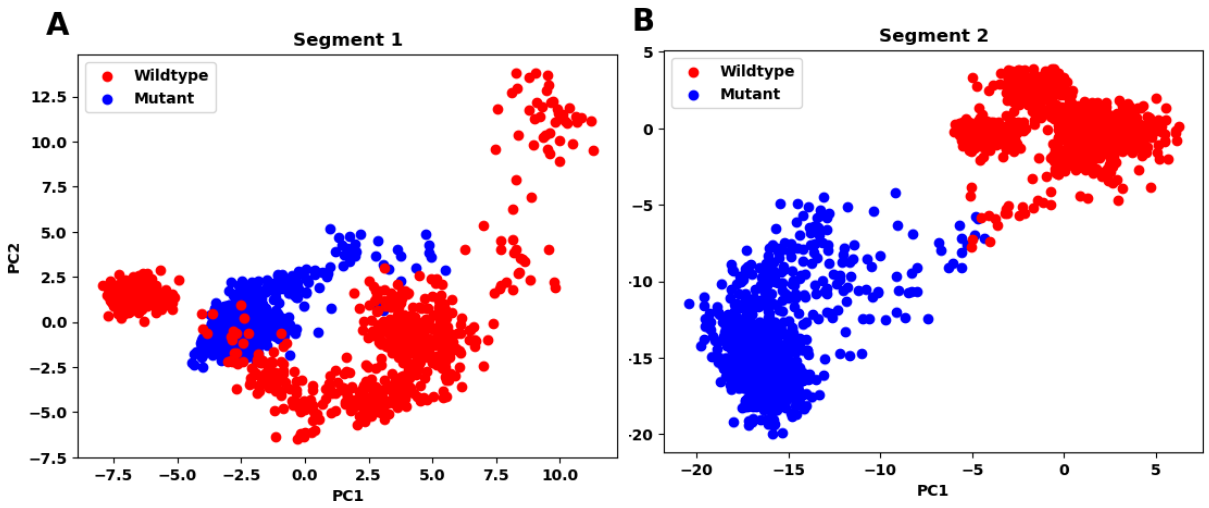


Figure 67 - PCA plots of trajectory analysis for segments 1 (A) and 2 (B) of wild-type and mutant ADAMTS3 proteins. The trajectories of wild-type and mutant structures are represented by blue and red dots, respectively

Free energy landscape (FEL) plots are constructed after PCA analysis using the first two principal components. In the FEL plots (Figure 68), the conformation with the lowest energy is depicted in dark blue. For segment 1, the lowest energy for the wild-type structure is 12.2 kJ/mol, while for the mutant structure, it is 10.9 kJ/mol. For segment 2, the lowest energy for the wild-type structure is 7.80 kJ/mol, while for the mutant structure, it is 9.08 kJ/mol. For both segments, wild-type and mutant structures in ADAMTS3 demonstrate differences in the number and position of stable conformations that correspond to local minima in the FEL plots. This suggests that the mutations have influenced the overall conformational stability of the protein.

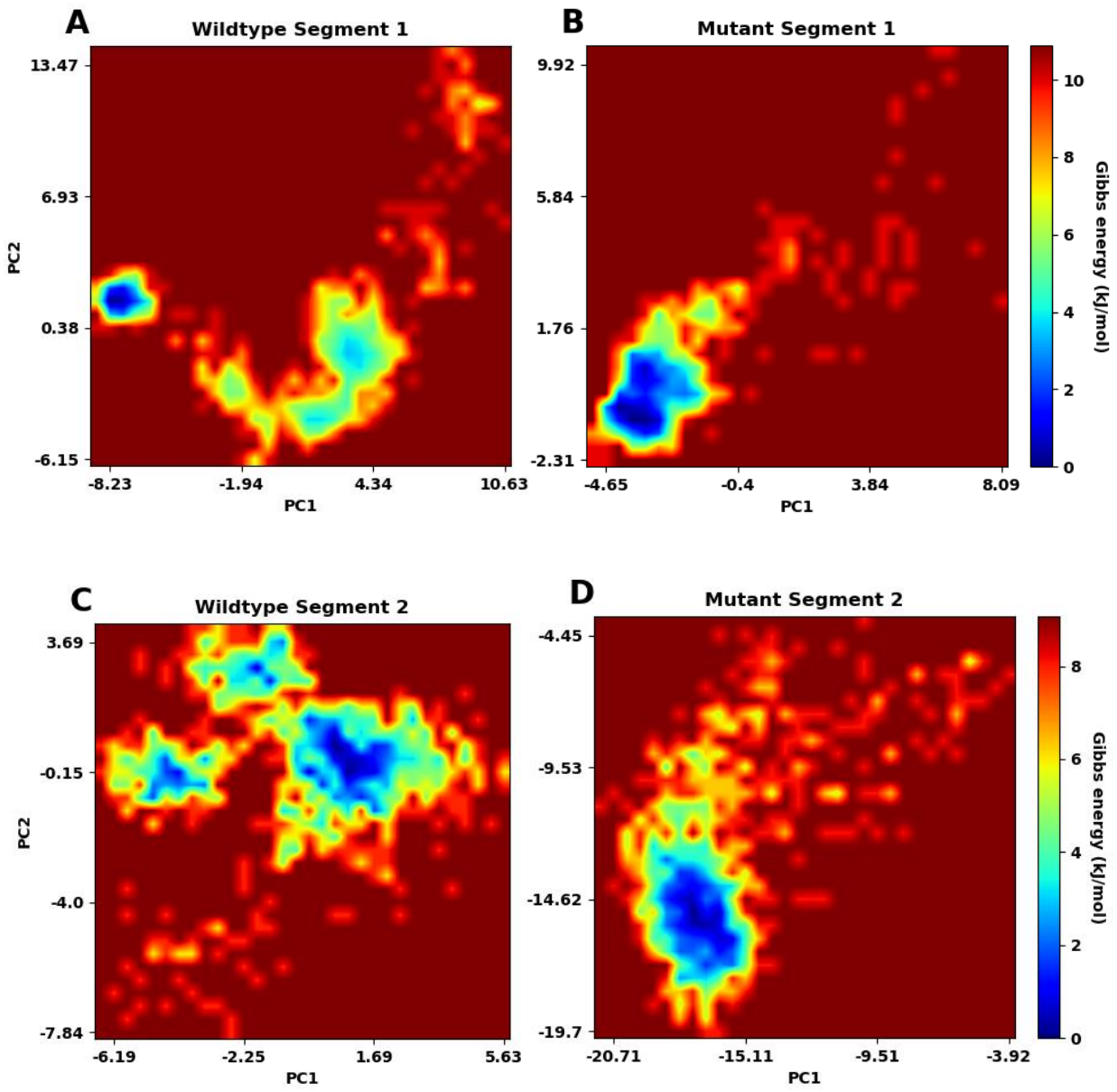


Figure 68 - Free energy landscape (FEL) analysis. The Gibbs energy is plotted as a function of the first two principal components (PC1 and PC2) for segments 1 and 2 of wild-type and mutant ADAMTS3. The conformation with the lowest energy is denoted by dark blue color

5.8 - Molecular dynamics simulation of wild-type and mutant FAT4 protein

The FAT4 protein consists of 4981 amino acid residues, making it a very long protein that needs to be divided into multiple fragments for modeling. In this study, we created five models: model 1 = sequence of 540 amino acid residues; model 2 =

600 residues; model 3 = 660 residues; model 4 = 600 residues; and model 5 = 420 residues. Multiple simulations of these sequences were conducted.

The changes in the root-mean-square deviation (RMSD) values of Cα atoms of wild-type and mutant FAT4 are presented in figures 69-73. Figure 69 shows that the wild-type protein reaches stability almost after 30 ns, and then the system converges and equilibrates after 60 ns. The RMSD values of the mutant protein deviate, and after 60 ns, the RMSD constantly increases until the end of the simulation. Therefore, the model 1 protein deviates more compared to the wild-type throughout the simulation. These results indicate that the wild-type protein is more stable than the mutant protein for model 1.

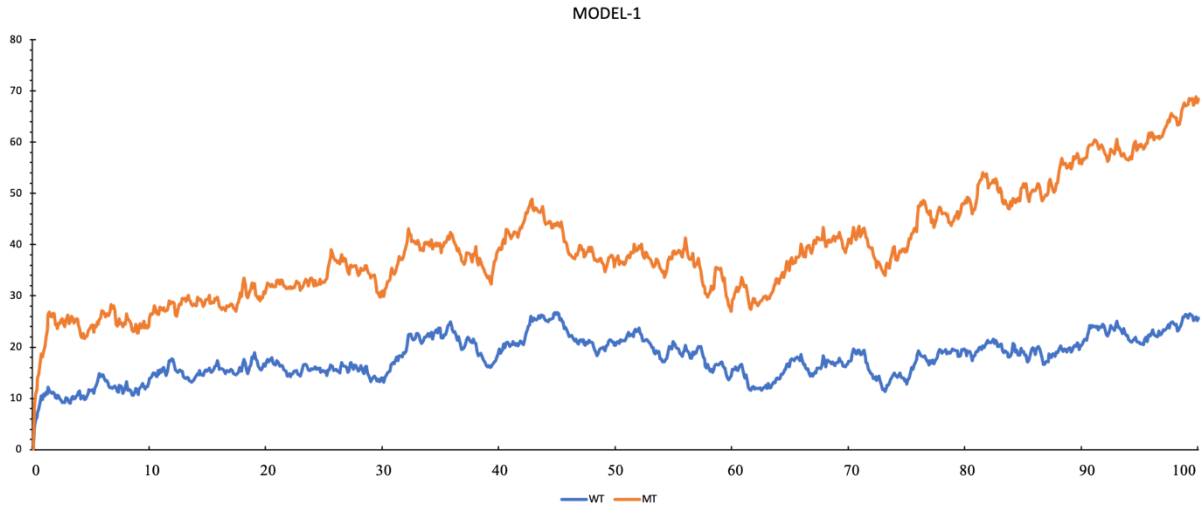


Figure 69 - Root-mean-square deviation (RMSD) of FAT4 protein Cα atoms for wild-type (blue) and mutant model 1 (red) over time

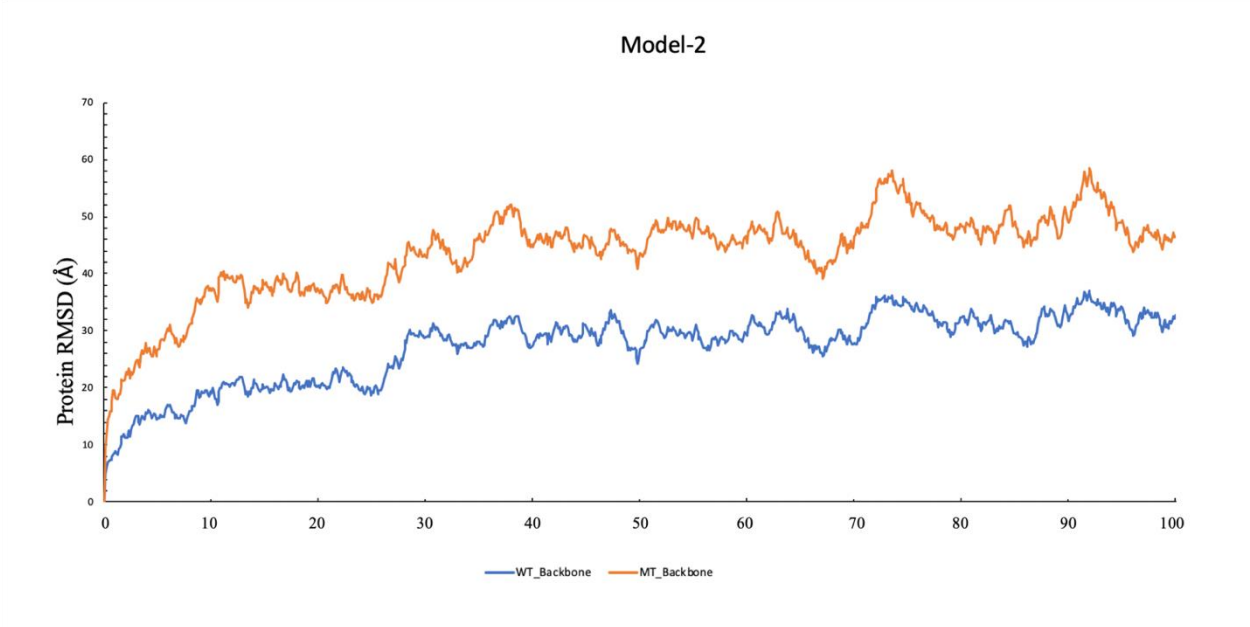


Figure 70 - Root-mean-square deviation (RMSD) of the Cα atoms of wild-type (blue) and mutant model 2 (red) of the FAT4 protein over time

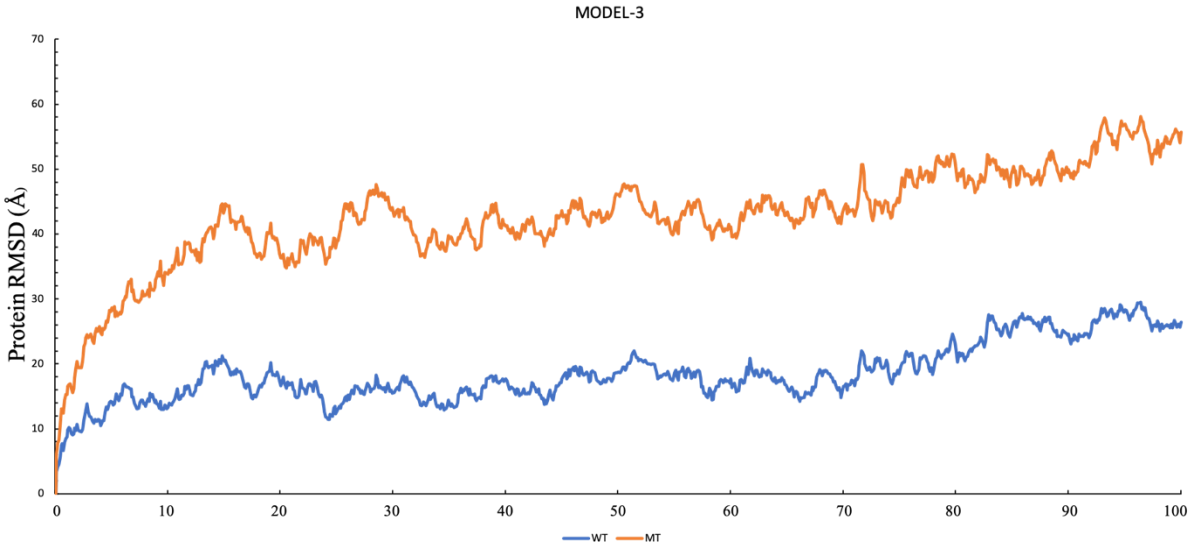


Figure 71 - Root Mean Square Deviation (RMSD) of the Cα atoms of wild-type (blue) and mutant model 3 (red) of FAT4 protein over time

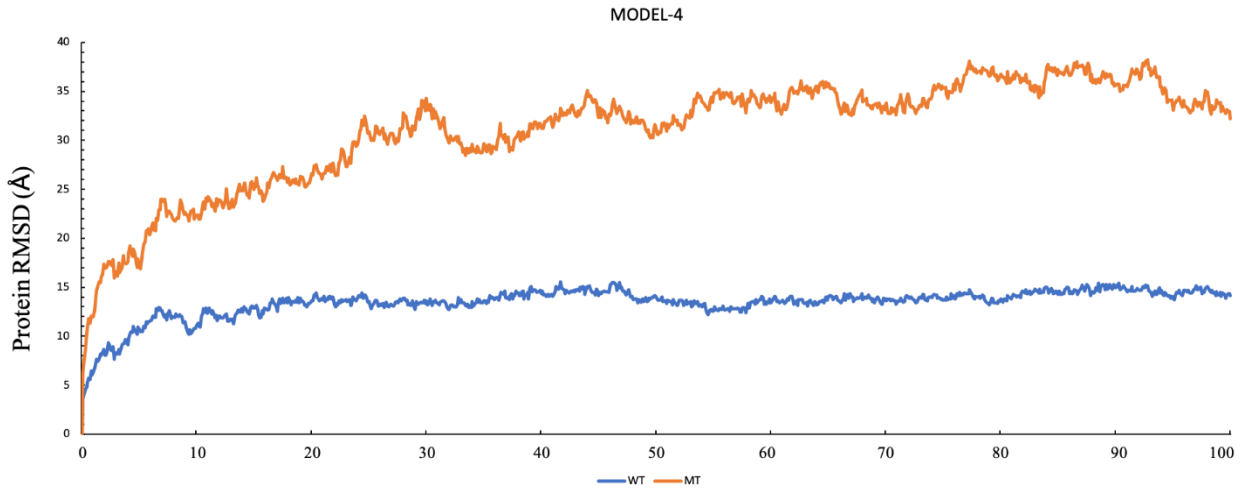


Figure 72 - Root mean square deviation (RMSD) of the Cα atoms of wild-type (blue) and mutant model 4 (red) of FAT4 protein over time

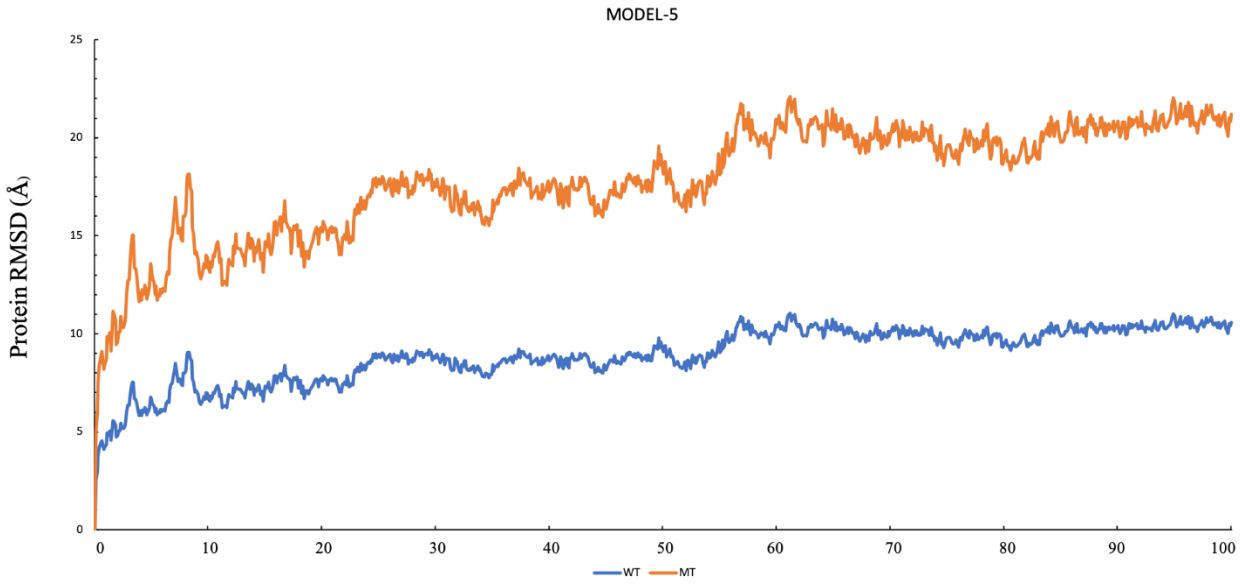


Figure 73 - Root-mean-square deviation (RMSD) of the Cα atoms of wild-type and mutant model 5 of FAT4 protein over time

Similarly, for model-2, after 28 ns, there were fewer deviations in the wild-type protein and the system remained equilibrated throughout the simulation. However, for the mutant variant, an increase in RMSD was observed almost at 30 ns and a larger deviation was observed until 100 ns (Figure 70).

RMSD for model-3 is shown in Figure 71. For the wild-type protein, there was a small deviation at almost 80 ns, after which the simulation converged, while

for the mutant type, RMSD continuously increased after 25 ns. This indicates greater stability of the wild-type protein compared to mutant model 3.

Similarly, for model-4 and model-5, the wild-type FAT4 is more stable than the mutant models. The protein regions that fluctuate the most during simulation are shown as peaks in the RMSF graph (Figures 74-78). Protein tails (both N- and C-termini) undergo changes more often than other regions of the protein. Alpha helices and beta sheets, for example, are usually more rigid and less fluctuating than the disordered parts of the protein. Residues with higher peaks, according to the trajectories, correspond to loop regions or N- and C-terminal zones. RMSF shows that there are more fluctuations in the mutant models compared to the wild-type FAT4 protein.

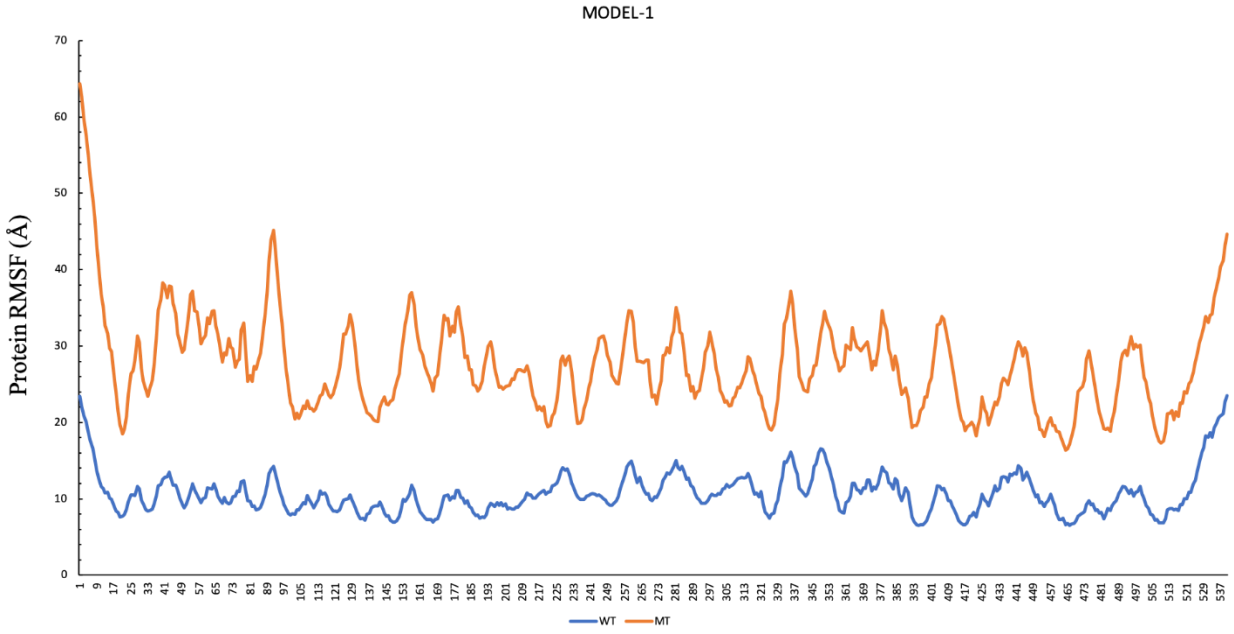


Figure 74 - Root mean square fluctuation (RMSF) of FAT4 wild-type (blue) and mutant model 1 (red) α-carbons over time

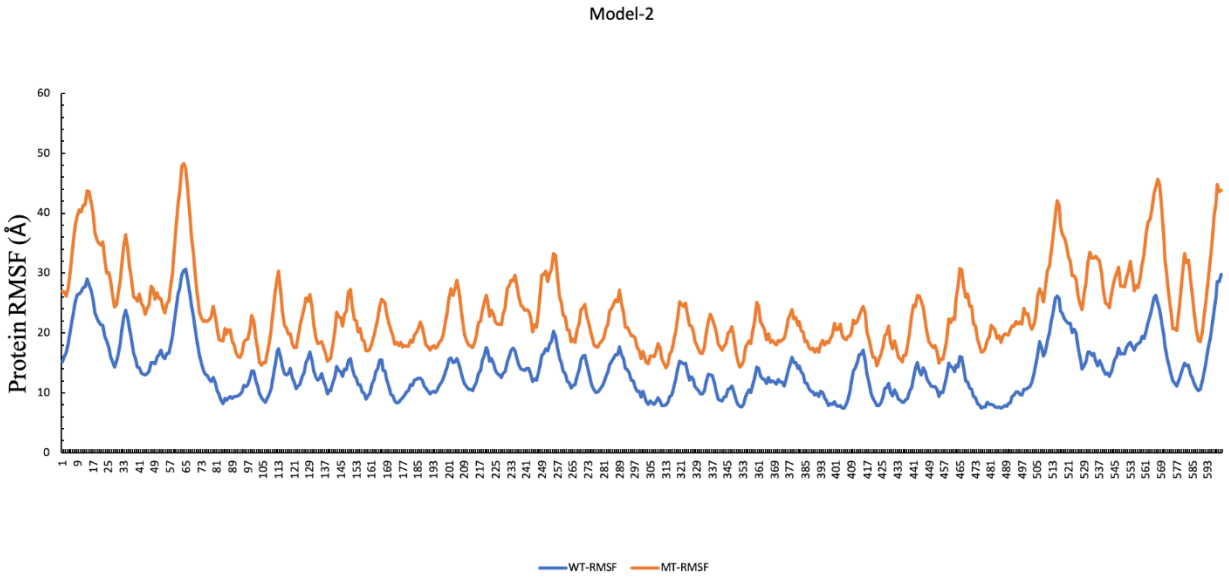


Figure 75 - Root Mean Square Fluctuation (RMSF) of FAT4 protein's α-carbon atoms for the wild-type (blue) and mutant model 2 (red) over time

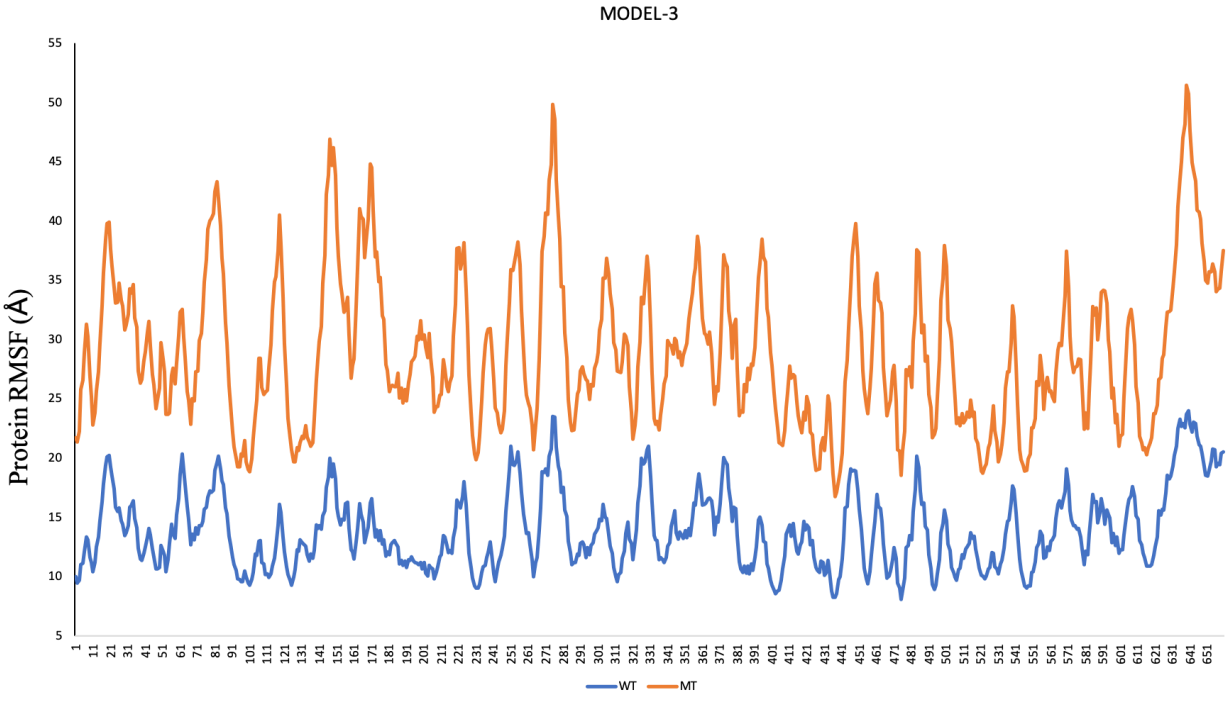


Figure 76 - Root-mean-square fluctuation (RMSF) of FAT4 protein's Cα atoms for the wild-type (blue) and mutant model 3 (red) over time

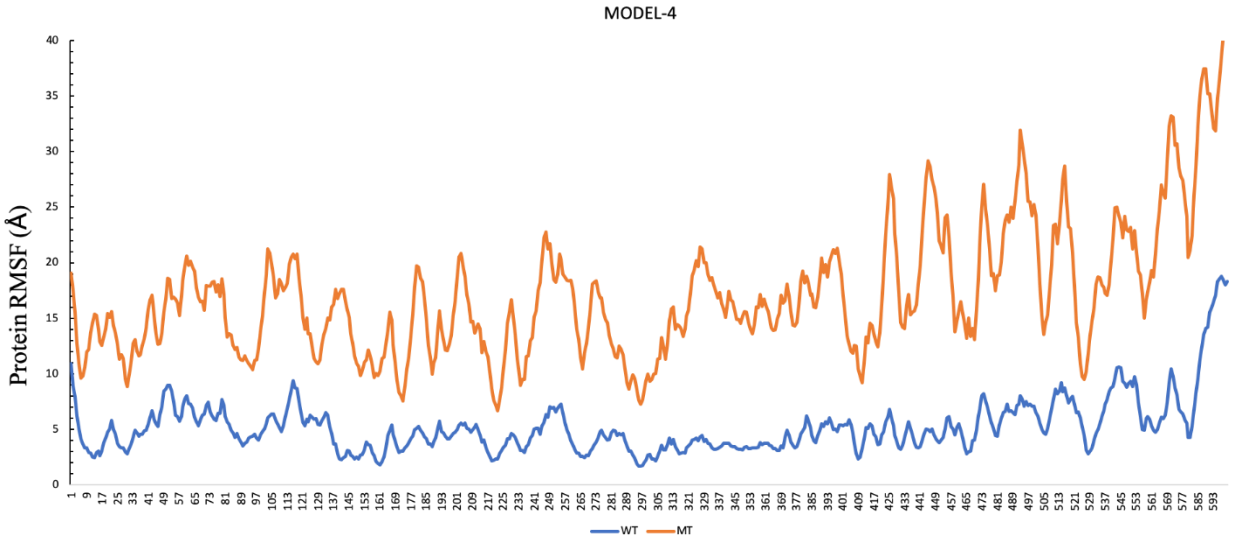


Figure 77 - Root Mean Square Fluctuation (RMSF) of the FAT4 wild-type (blue) and mutant model 4 (red) α-carbon atoms over time

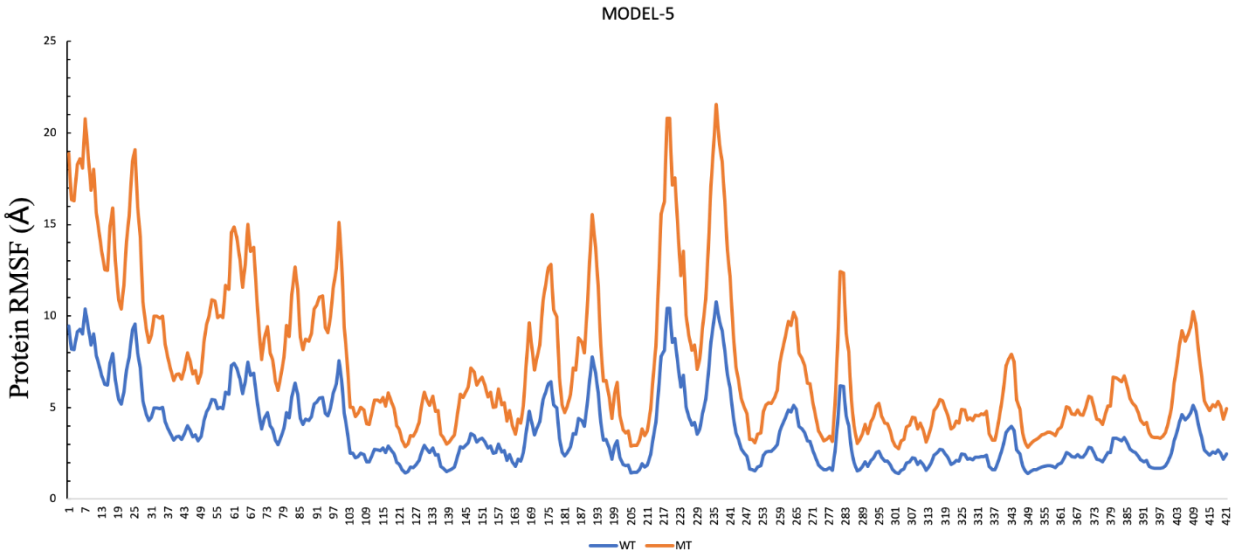


Figure 78 - Root mean square fluctuation (RMSF) of the α-carbon atoms of wild-type FAT4 protein (blue) and mutant model 5 (red) over time

The average distributions of the secondary structure elements (SSE) of the protein were also calculated during the 170 ns simulation. Alpha helices (in orange) and beta sheets (in blue) were tracked as SSEs among other elements of the protein's secondary structure. For wild type protein in model-1, the simulation showed 0.83% helices and 35.49% beta strands, alongside 36.33% SSE, whereas the mutant protein showed 1.33% helices and 36.05% beta strands, alongside 37.38% SSE. For model-

2, the wild type protein showed 0.37% helices and 35.25% beta strands, alongside 35.62% SSE, whereas the mutant protein showed 1.63% helices and 34.43% beta strands, alongside 36.06% SSE. Similar analyses were performed for other models, and these data, along with graphical representations of the results, were published in the corresponding article. These results were not of principal importance for further analysis but were taken into account.

The analysis of the radius of gyration (Rg) of the wild type protein and mutant models showed that the mutants exhibited a higher Rg value on the simulation time scale compared to the wild type. Consequently, the flexibility of the mutants increased (see Figure 79). Similar analyses were performed for other models, and these data, along with graphical representations of the results, were published in the corresponding article.

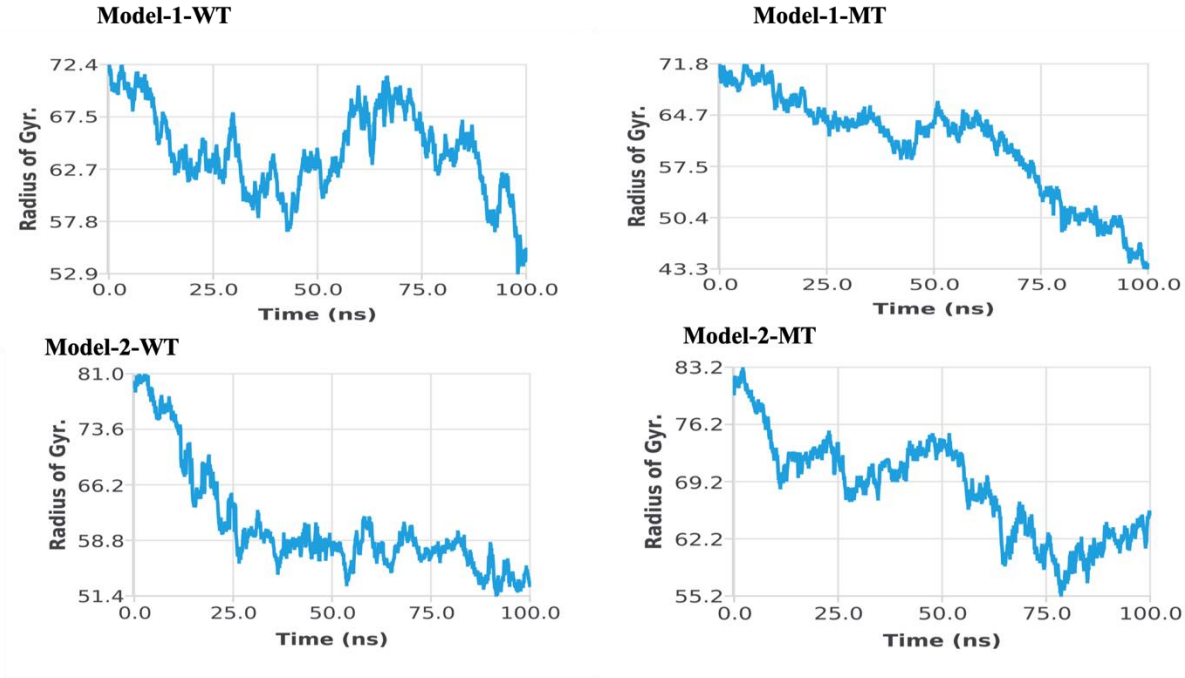


Figure 79 - Analysis of radius of gyration over time during modeling of wild type (left) and mutant (right) variants of FAT4 models 1 (top) and 2 (bottom)

Analysis of the energetic parameters for mutant and wild type models shows that the total energy of the mutant system was increased compared to the wild type,

but not equally for different models. The largest differences were observed for models 1, 2, and 3, while the smallest were observed for models 4 and 5.

5.9 - Assessment of the impact of non-synonymous single nucleotide variants on the structure and function of the FAT4 protein identified in a patient with Hennekam syndrome phenotype

We conducted an identification of the impact of missense mutations in the FAT4 gene, as well as missense mutations p.A807V, p.G3526D, and p.S3875N, which we obtained through VCF from the whole genome data of a patient suspected of having Hennekam syndrome (Table 20). Using multiple algorithms and tools, these substitutions were evaluated as deleterious. By including these amino acid substitutions in the models during molecular dynamics simulation of the FAT4 protein, we tested the hypothesis that these mutations disrupt the structure and function of the protein.

Table 20 - Overview of nsSNPs in the FAT4 gene obtained from the analysis of whole genome sequencing of a patient with Hennekam syndrome phenotype, including the mutation site, pathogenicity predictions, and information on population prevalence

Gene	nsSNP	Genotype	substitution in the protein	Protein mutation	Frequency in the population (1000g/gnomAD/ExAC)	Prediction of pathogenicity (FATHMM/PROVEAN/CADD)
FAT4	rs1039808	Heterozygotic	c.C2420T	p.A807V,	0.424/0.461/0.422	D/N/17.22
FAT4	rs1567047	Heterzygotic	c.G10577A	p.G3526D	0.231/0.225/0.267	D/D/29.7
FAT4	rs12650153	Homozygotic	c.G11624A	p.S3875N	0.009/0.007/0.002	T/N/20.1

Thus, in this chapter, we presented the results of nsSNP studies in the genes CCBE1, ADAMTS3, and FAT4, some of which have already been published by other researchers, while others were newly identified by us or found in the full-genome sequence data of a patient with the phenotype of Hennekam syndrome. We

verified these mutations using various in silico tools to identify harmful nsSNPs among them, confirmed the impact of selected potentially harmful mutations on the structure and function of the investigated proteins, and then performed protein modeling, analysis, and simulation of the molecular dynamics of wild-type and mutant protein models to determine the effects that the identified amino acid substitutions may have on the CCBE1, ADAMTS3, and FAT4 proteins.

List of works published by chapter 5

1. Predicting the Most Deleterious Missense Nonsynonymous Single-Nucleotide Polymorphisms of Hennekam Syndrome-Causing CCBE1 Gene, In Silico Analysis / K. Shinwari, L. Guojun, S.S. Deryabina, M.A. Bolkov, I.A. Tuzankina, V.A. Chereshnev // ScientificWorldJournal. 2021. Jun 10; 2021:6642626. doi: 10.1155/2021/6642626. PMID: 34234628; PMCID: PMC8211529.

2. Novel high-risk missense mutations identification in FAT4 gene causing Hennekam syndrome and Van Maldergem syndrome 2 through molecular dynamics simulation / K. Shinwari, H.M. Rehman, N. Xiao, L. Guojun, M.A. Khan, M.A. Bolkov, I.A. Tuzankina, V.A. Chereshnev // Informatics in Medicine Unlocked. 2023. 37. [101160]. https://doi.org/10.1016/j.imu.2023.101160.

CONCLUSION

With the advancement of technology and accumulation of new knowledge, there is a constant improvement in the understanding of significant developmental anomalies affecting the human population. However, much still remains unknown and awaits discovery. Predicting any deviation in a living system that may lead to disease/syndromes/death depends on multiple variables. Studying the genetic factors opens up significant opportunities for diagnosis, prognosis, and personalized treatment for physicians and biologists, but first, it is necessary to evaluate the impact of each individual genomic variant on the structure and function of the protein, as well as its influence on the phenotype as a whole.

The results presented in this work were achieved through the development and integration of modern technologies for evaluating the impact of non-synonymous single nucleotide polymorphisms (nsSNPs) on the encoded protein into the analysis process. Along with other evaluation methods and the interpretation of large-scale and complex multidimensional data in systems biology, this may become another instrument for regular research on various human pathologies, including immunopathology.

To understand the mechanisms of increased susceptibility of patients with RBCK1 deficiency to infections, as well as their relationship with amylopectinosis present in the disease phenotype, we conducted a bioinformatics study of a previously described case of RBCK1 deficiency accompanied by autoinflammation and an infectious syndrome, which distinguishes this disease from other autoinflammatory syndromes in this group.

Genes with increased and decreased expression in autoinflammatory syndrome RBCK1 deficiency (also known as HOIL1 deficiency), identified during our research, allowed us to identify key signaling pathways involved in the development of this disease (signaling pathways involved in Staphylococcus aureus,

Vibrio cholerae infections, leishmaniasis, intracellular signal transduction, antigen processing and presentation, NK-mediated cytotoxicity, and others).

These signaling pathways and corresponding proteins directly or indirectly reflect the deep involvement of molecular processes related to the immune system in the pathogenesis of the disease. In addition, a general understanding of the immune mechanisms involved has been developed.

In particular, changes in the activity of mTOR, PI3K/AKT, Rho, and Nf-kB signaling pathways have been shown to affect the expression of immune system genes, cell apoptosis, and sensitivity to the key cytokine of the immune response, IL-1β.

Moreover, the gene CSID2 significantly affects cell susceptibility to ER stress, apoptosis, and cell death. Its expression was significantly reduced in RBCK1-deficient cells compared to mononuclear cells from the peripheral blood of healthy children ($p=0.00000000000000007537936$).

We also found that the differences in the expression of genes related to viral infections, including the signaling pathway involved in SARS-CoV-2 infection, are insignificant. This is also confirmed by clinical observations described in publications by other researchers, where particular susceptibility in patients is identified specifically with respect to bacterial agents [129].

Overall, it can be concluded that increased susceptibility to pyogenic infections is complicated by general protein ubiquitination disorders, extensive glycophagy with glycogen depletion and accumulation of polysaccharides, as well as identified differences in gene expression and, most likely, production of various immune response proteins.

In our study, the network density and the so-called biological distance for genes related to primary immunodeficiencies (PIDs) and congenital neutropenia, in particular, were found to be functionally similar to each other and closely interact compared to other PID genes. Using these data, as well as identifying genes whose expression differs significantly from normal in severe congenital neutropenia, and combining this information with data on protein-protein interaction and gene

function characterization data ("biological distance" and "network density"), we were able to predict causally significant genes for the development of congenital neutropenia that were not previously described in the classification of primary immunodeficiencies in this role.

In our study, we identified 15 novel candidate genes for the development of congenital neutropenia that are interdependent with known genes involved in the same biological pathways, demonstrating the high biological significance of their correlation with known congenital neutropenia genes. The confirmation of several predicted genes and their impact on neutrophil functions in recent studies on patients with neutrophil defects convincingly demonstrate the significance of these candidate genes for the development of pathology.

Additionally, in our investigation of congenital neutropenia, we analyzed the pathogenicity of single nucleotide variants in the ELANE and TCIRG1 genes. Using several computational tools, we identified 8 non-synonymous single nucleotide variants (rs28931611, rs57246956, rs137854448, rs193141883, rs201723157, rs201139487, rs137854451, and rs200384291) in the neutrophil elastase (ELANE) gene that are most disruptive to the protein structure and function. Variants with substitutions F218L, R34W, G203S, R193W, and T175M have not yet been detected in patients with severe congenital neutropenia, while variants C71Y, P139R, C151Y, G214R, and G203C, which we reported in our study, are already associated with both disorders. These mutations destabilize the structure, disrupt the activation, splicing, and folding of the ELANE protein and may decrease the efficiency of the trypsin-like serine protease.

The TCIRG1 gene defect has recently been considered by various scientists not only as a cause of hereditary osteopetrosis (aggressive osteoporosis and increased risk of fractures), but also as a cause of congenital neutropenia. The results of whole-genome sequencing of a patient with congenital neutropenia at our disposal allowed us to suspect the TCIRG1 gene variant as the cause, especially since other hereditary causes of congenital neutropenia had not been previously identified by

specialists, and the patient receives specific treatment and is under observation by an immunologist with this diagnosis.

To assess the pathogenicity of the identified non-synonymous single nucleotide variant, molecular dynamics simulation of the TCIRG1 protein was conducted with consideration of the given amino acid substitution (V52L). The tests showed that the resulting modified protein is less stable, indicating a higher probability of inadequate functioning in the patient.

To provide convincing evidence, it is necessary to conduct testing of the identified protein in the patient and their parents, as well as exclude other potential causes that may be discovered in the future. Nonetheless, the in silico investigation method has allowed for the increased significance of such a substitution in the protein and presents it as a substitution requiring special attention.

In addition to the TCIRG1 V52L mutation, we tested other single amino acid changes in highly conservative regions of the TCIRG1 protein, and a total of 15 nsSNPs (rs199902030, rs200149541, rs372499913, rs267605221, rs374941368, rs375717418, rs80008675, rs149792489, rs116675104, rs121908250, rs121908251, rs121908251, rs149792489, and rs116675104) were identified, which are likely pathogenic gene variants since they destabilize the structure and function of the wild-type protein. Some of these variants are located in the conserved domain of V-ATPase I. These variants have not yet been identified in patients with congenital neutropenia and/or osteopetrosis, while the G405R, R444L, and D517N variants that were reported in our study have already been confirmed by other researchers as variants associated with osteopetrosis [26, 34]. The results of the investigation may help in further understanding the broad spectrum of diseases associated with the activation of the TCIRG1 kinase catalytic domain and assist in developing effective treatments for diseases associated with changes in this protein.

Similar methods were applied in assessing the impact of non-synonymous single nucleotide substitutions on the structure and function of proteins responsible for the development of Hennekam syndrome - FAT4 and ADAMTS3. In addition to these two genes, this autosomal recessive disorder, in which lymphangiectasia and

lymphedema play a key role in its pathogenesis, is also associated with defects in the CCBE1 gene. Three corresponding proteins affect the activation of the primary lymphangiogenic growth factor VEGF-C.

Using modern in silico tools, this study investigated the most pathogenic non-synonymous single nucleotide polymorphisms (nsSNPs) in the CCBE1, FAT4, and ADAMTS3 genes. Our results demonstrate that seven nsSNPs in the CCBE1 gene (rs115982879, rs149792489, rs374941368, rs121908254, rs149531418, rs121908251, and rs372499913) are likely to have a pathogenic impact, with four of them (G330E, C102S, C174R, and G107D) having a very high probability of being pathogenic, and two of them (G330E and G107D) never having been reported in the context of Hennekam syndrome. In addition, two important substitutions in the CCBE1 gene (rs374941368 and rs200149541) were evaluated, which may have an impact on post-translational modifications, as they affect a potential phosphorylation site. The web-based ligand-binding analysis service FTSite was used to assess the impact of these substitutions on molecule function, and the two substitutions were found to be potentially highly deleterious and should be taken into account when diagnosing Hennekam syndrome.

When analyzing variants of the ADAMTS3 gene from the dbSNP database, 919 nsSNPs were initially sorted, of which five substitutions (G298R, C567Y, A370T, C567R, and G374S) were predicted to be the most dangerous and potentially associated with disease. Protein modeling showed that the protein can be divided into segments 1, 2, and 3, which are connected by short loops. Using molecular dynamics simulation tools, it was found that some substitutions significantly destabilize the protein structure and disrupt secondary structures, especially in segment 2. The pathogenic effect of mutations in segment 1 may be related not to destabilization, but to other factors, such as changes in phosphorylation, as suggested by post-translational modification studies.

Our work represents the first study of ADAMTS3 gene polymorphisms using multiple tools, including molecular dynamics simulation. Some of the predicted substitutions in the ADAMST3 protein are not yet reported in the PubMed library,

and we hope that the obtained data will be useful for diagnostic tasks and the search for therapy methods.

In analyzing various variants of the FAT4 gene among 3,343 nsSNPs available in the NCBI library using different tools to predict pathogenicity, 11 substitutions in the FAT4 protein (D2978G, V986D, Y1912C, R4799C, D1022G, G4786R, D2439E, E2426Q, R4643C, N1309I, and Y2909H) were predicted as potentially pathogenic. In addition, three substitutions in the FAT4 gene (rs12650153, rs1567047, and rs1039808) were previously detected in a patient with the presumed Henneman syndrome by filtering candidate variants during whole-genome sequencing, and in silico study of these mutations showed that they strongly destabilize the protein structure and function.

In this study, using the molecular dynamics simulation method (MDS), we focused on 19 mutations in the FAT4 gene - 11 predicted in our in silico study, 3 nsSNPs detected in the patient, and 5 nsSNPs already published as likely causes of Henneman and Van Maldergem syndromes, which differ phenotypically from Henneman syndrome.

The results of the applied molecular dynamics simulation method confirmed lower stability of the "mutant" protein compared to the "wild" type. Genetic variants detected in this cohort of studies were not previously registered as causes of Henneman syndrome. It is worth noting that due to the limited resources of the supercomputer and software, such a long protein as FAT4, consisting of 4981 amino acids, could only be simulated fragmentarily, in segments containing the analyzed substitution of less than 1000 amino acids. Nevertheless, we hope that these results can contribute to a better understanding of the predisposition to diseases associated with the activation of the FAT4 protein and may help in the development of effective approaches for the diagnosis and treatment of diseases related to this gene.

In general, it should be noted that the FAT4 molecule has a huge size and is itself a flexible structure, providing the transmission of not fully understood intercellular signals. Perhaps this molecule provides spatial orientation, cell polarization, and signal transmission about intercellular contact, among other things.

Due to its length, it is difficult to predict how much of a serious impact a single amino acid substitution has. Accumulation of differences and especially impairment of functional active binding centers of molecules should be more significant than a single amino acid substitution.

Prospects for further development of the topic

Identification of specific genetic changes and determination of the molecular basis of immunopathology will enable the study of pathogenetic mechanisms, differentiation of nosological forms from a vast heterogeneous group of inborn errors of immunity, and approach the creation of specific targeted therapies, including gene editing and antisense oligonucleotides. This will make it possible to address issues of radical patient cure. In addition, even a simple acceleration of the diagnostic process will help timely diagnosis, prescribe replacement and pathogenetic therapy, improve the prognosis, and quality of life of patients.

The process of verifying genes of primary immunodeficiency can be improved by developing software to predict candidate genes of various immunopathologies, and incorporating methods for predicting the impact of genetic changes on protein in silico provides the possibility of its effective use in clinical research.

In addition, studying rare cases of human pathology allows us to address general pathological issues of disease formation, enriching science with knowledge of the functioning laws of the immune system and the human body as a whole. By delving into the molecular level of pathology, researchers gain objective justifications for the development and application of targeted therapeutic tactics, which opens up the prospect of creating new targeted drugs.

Thus, our research has allowed us to draw the following conclusions.

FINDINGS

1. New genetic findings have been identified in three types of primary immunodeficiency disorders: RBCK1 deficiency, congenital neutropenia, and Hennekam syndrome.

2. Significant differences in gene expression have been found in RBCK1 deficiency compared to healthy children and patients with CINCA/NOMID syndrome, Muckle-Wells syndrome, and mevalonate kinase deficiency.

3. Non-synonymous single nucleotide substitutions in the TCIRG1 gene (rs199902030, rs200149541, rs372499913, rs267605221, rs374941368, rs375717418, rs80008675, rs149792489, rs116675104, rs121908250, rs121908251, rs121908251, rs149792489, rs116675104) and ELANE gene (rs200384291, rs201163886, rs193141883, rs201139487, rs201723157) destabilize the TCIRG1 and ELANE proteins in neutrophils.

4. The genes CDC42, CRKL, FGR, CRC, NYK, PLCG1, ARRB2, PIK3CG, PTK2, STAT1, STAT2, STAT3, STAT5B, VAV1, and ITK are new candidate genes for the development of congenital neutropenia.

5. Non-synonymous single nucleotide substitutions in the CCBE1 (rs115982879, rs149792489, rs374941368, rs121908254, rs149531418, rs121908251, and rs372499913), FAT4 (rs147663284, rs192514171, rs138137489, rs199895179, rs372060616, rs138173652, rs142184187, rs147633644, rs181607904, rs184971791, rs148655455), and ADAMTS3 (rs61757480, rs61741624, rs140806973, rs140595148, rs140914273, rs142268705, rs142781084, rs143059623, rs146979323, rs372067284, rs370857003, rs375983592, rs367831484, rs202031187, and rs150012152) genes lead to destabilization of the CCBE1, FAT4, and ADAMTS3 proteins and may cause Hennekam syndrome.

6. The developed program for sequential use of bioinformatics methods is effective in identifying genes that influence the pathogenesis of diseases associated with primary immunodeficiency disorders.

PRACTICAL RECOMMENDATIONS

1. In diagnosing primary immunodeficiencies (inborn errors of immunity), it is necessary to determine the gene expression profile by analyzing differential gene expression, signaling pathways, and genetic ontologies, as well as identifying biomarkers of pathology, which will reduce the costs of treatment and prevent the development of side effects.

2. When conducting research on predicting new candidate genes for congenital neutropenia, it is necessary to include co-expression factors, protein-protein interactions, and signaling pathways in the analysis.

3. For the differential diagnosis of congenital neutropenia, in addition to the genes listed on the ESID website and in the IUIS classification, additional genes (CDC42, CRKL, FGR, CRC, NYK, PLCG1, ARRB2, PIK3CG, PTK2, STAT1, STAT2, STAT3, STAT5B, VAV1, and ITK), identified in our study as candidate genes, should be included.

4. For the differential diagnosis of Henneman syndrome and congenital neutropenia, in addition to the listed missense mutations in the genes ADAMTS3, FAT4, CCBE1, ELANE, and TCIRG1, it is necessary to assess the presence of nsSNP missense mutations identified in our study for the following genes: CCBE1 (rs115982879, rs149792489, rs374941368, rs121908254, rs149531418, rs121908251, and rs372499913), ELANE (rs200384291, rs201163886, rs193141883, rs201139487, and rs201723157), TCIRG1 (rs199902030, rs200149541, rs372499913, rs267605221, rs374941368, rs375717418, rs80008675, rs149792489, rs116675104, rs121908250, rs121908251, rs121908251, rs149792489, and rs116675104), FAT4 (rs147663284, rs192514171, rs138137489, rs199895179, rs372060616, rs138173652, rs142184187, rs147633644, rs181607904, rs184971791, rs148655455), and ADAMTS3 (rs61757480, rs61741624, rs140806973, rs140595148, rs140914273, rs142268705, rs142781084, rs143059623, rs146979323, rs372067284, rs370857003, rs375983592, rs367831484, rs202031187, and rs150012152).

REFERENCES

1. Анализ уровней TREC и KREC в высушенных пятнах крови здоровых новорожденных с различным сроком беременности и весом. / Д.А. Черемохин, [и др.] // Acta Naturae. 14 (1). С.101-108.

2. Воронина Л.И. Факторы влияния на отношение медицинского сообщества к генетическим исследованиям и оказанию медицинских услуг детям с врожденными ошибками иммунитета / Л.И. Воронина, Е.В. Зайцева, И.А. Тузанкина // Социология медицины. 2019/ 18(2)/ С.92–97.

3. Изменчивость симптомокомплекса CATCH-22 в рамках синдрома делеции 22q11.2. / Д.А. Черемохин, [и др.] // Медицинская иммунология. 2021/ 23 (6). С.1357-1366.

4. Классификация врожденных ошибок иммунитета человека, обновленная экспертами комитета Международного союза иммунологических обществ в 2019 году / М.А. Болков, [и др.] // Российский иммунологический журнал. 2021. Т. 24. №1. С. 7-68.

5. Неонатальный скрининг на тяжелую комбинированную иммунную недостаточность в России: прекрасное далеко или завтрашняя реальность? / С.С. Дерябина, [и др.] // Вопросы современной педиатрии. 2017. 16 (1). С. 59-66.

6. Первичные иммунодефициты (врожденные ошибки иммунитета) в раннем возрасте: монография / И.А.Тузанкина и др.; рец.: А.С.Симбирцев, З.Ж.Рахманкулова; – Ташкент: Изд-во «Adast poligraf», 2022. – 232 с.

7. Приказ Министерства здравоохранения Российской Федерации от 21.04.2022 № 274н. – Москва, 2022.

8. Ретроспективный анализ случаев первичных иммунодефицитов у детей с врожденными пороками сердца / С.С. Дерябина, [и др.] // Российский иммунологический журнал. 2020. Т. 23, № 4. С. 505-514.

9. Ретроспективная диагностика первичных иммунодефицитных состояний у детей в Свердловской области / С.С. Дерябина, [и др.] // Медицинская иммунология. 2016. 18 (6). С. 583-588.

10. Роль врожденных ошибок иммунитета в группе детей с летальными исходами на первом году жизни / Д.А. Черемохин, [и др.] // Российский иммунологический журнал. 2022. Т.25, № 4. С. 555-560.

11. Фенотипические варианты манифестации гомозиготной делеции сегмента хромосомы 1, захватывающей участки гена CFHR3 / И.А. Тузанкина, [и др.] // Медицинская иммунология. 2020, 22(3), С. 569-576

12. Эпидемиология первичных иммунодефицитов в Российской Федерации / А.А. Мухина, [и др.] // Педиатрия им. Г.Н. Сперанского. 2020; 99 (2): С. 16-32.

13. A catalytic-independent role for the LUBAC in NF-κB activation upon antigen receptor engagement and in lymphoma cells / S.M. Dubois, [et al.] // Blood. 2014, № 14 (123). С. 2199–2203.

14. Aksentijevich, I. NF-κB Pathway in Autoinflammatory Diseases: Dysregulation of Protein Modifications by Ubiquitin Defines a New Category of Autoinflammatory Diseases / I. Aksentijevich, Q. Zhou // Frontiers in Immunology. 2017. (8). С. 399.

15. Al-Mousa, H. Primary Immunodeficiency Diseases in Highly Consanguineous Populations from Middle East and North Africa: Epidemiology, Diagnosis, and Care / H. Al-Mousa, B. Al-Saud // Frontiers in Immunology. 2017. (8). С. 678.

16. A Next-Generation Sequencing Test for Severe Congenital Neutropenia: Utility in a Broader Clinicopathologic Spectrum of Disease / S.N. McNulty, [et al.]. // The Journal of molecular diagnostics: JMD. 2021. № 2 (23). С. 200–211.

17. A novel computational and structural analysis of nsSNPs in CFTR gene / C. George Priya Doss, [et al.]. // Genomic Medicine. 2008. № 1–2 (2). С. 23–32.

18. A partial form of recessive STAT1 deficiency in humans / A. Chapgier, [et al.]. // The Journal of Clinical Investigation. 2009. № 6 (119). С. 1502–1514.

19. A proteome-scale map of the human interactome network / T. Rolland, [et al.]. // Cell. 2014. № 5 (159). С. 1212–1226.

20. As Little as Needed: The Extraordinary Case of a Mild Recessive Osteopetrosis Owing to a Novel Splicing Hypomorphic Mutation in the TCIRG1 Gene / C. Sobacchi, [et al.]. // Journal of Bone and Mineral Research. 2014. № 7 (29). С. 1646–1650.

21. A Systematic Review on Predisposition to Lymphoid (B and T cell) Neoplasias in Patients With Primary Immunodeficiencies and Immune Dysregulatory Disorders (Inborn Errors of Immunity) / I.B. Riaz, [et al.]. // Frontiers in Immunology. 2019. Vol. 10. С. 777.

22. An integrative approach to predicting the functional effects of non-coding and coding sequence variation / H.A. Shihab, [et al.] // Bioinformatics. – 2015. – Т. 31. – №. 10. – С. 1536-1543.

23. Atp6i-deficient mice exhibit severe osteopetrosis due to loss of osteoclast-mediated extracellular acidification / Y.P. Li, [et al.]. // Nature Genetics. 1999. № 4 (23). С. 447–451.

24. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. / J. Eberhardt, [et al.] //Journal of chemical information and modeling. – 2021. – Т. 61. – №. 8. – С. 3891-3898

25. Autosomal recessive intestinal lymphangiectasia and lymphedema, with facial anomalies and mental retardation / R.C. Hennekam, [et al.] // American Journal of Medical Genetics. 1989. № 4 (34). С. 593–600.

26. Autosomal recessive osteopetrosis type I: description of pathogenic variant of TCIRG1 gene / L.E. Chávez-Güitrón, [et al.] // Boletín médico del Hospital Infantil de México. 2018. № 4 (75). С. 255–259.

27. Bach, E.A. The IFN gamma receptor: a paradigm for cytokine receptor signaling / E.A. Bach, M. Aguet, R.D. Schreiber // Annual Review of Immunology. 1997. (15). С. 563–591.

28. Ballgown bridges the gap between transcriptome assembly and expression analysis / A.C. Frazee, [et al.] // Nature Biotechnology. 2015. № 3 (33). С. 243–246.

29. BioGRID: a general repository for interaction datasets / C. Stark, [et al.]. // Nucleic Acids Research. 2006. № Database issue (34). С. D535-539.

30. Bioinformatics analysis of key genes and latent pathway interactions based on the anaplastic thyroid carcinoma gene expression profile / Y. Huang, [et al.]. // Oncology Letters. 2017. № 1 (13). С. 167–176.

31. Blom N. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. / N. Blom, S. Gammeltoft, S. Brunak // Journal of Molecular Biology, 1999, № 5 (294), С. 1351–1362,

32. Bondos, S.E. Physical and genetic interactions link hox function with diverse transcription factors and cell signaling proteins / S.E. Bondos, X.-X. Tan, K.S. Matthews // Molecular & cellular proteomics: MCP. 2006. № 5 (5). С. 824–834.

33. Borregaard, N., Granules of the Human Neutrophilic Polymorphonuclear Leukocyte / N. Borregaard, J.B. Cowland // Blood. 1997. № 10 (89). С. 3503–3521.

34. Buried in the Middle but Guilty: Intronic Mutations in the TCIRG1 Gene Cause Human Autosomal Recessive Osteopetrosis / E. Palagano, [et al.]. // Journal of Bone and Mineral Research. 2015. № 10 (30). С. 1814–1821.

35. Capriotti, E. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure / E. Capriotti, P. Fariselli, R. Casadio // Nucleic Acids Research. 2005. № Web Server issue (33). С. W306-310.

36. Casanova, J.-L. Genetic dissection of immunity to mycobacteria: the human model / J.-L. Casanova, L. Abel // Annual Review of Immunology. 2002. Vol. 20. С. 581–620.

37. CCBE1 is required for embryonic lymphangiogenesis and venous sprouting / B.M. Hogan, [et al.]. // Nature Genetics. 2009. № 4 (41). С. 396–398.

38. CDG: An Online Server for Detecting Biologically Closest Disease-Causing Genes and its Application to Primary Immunodeficiency / D. Requena, [et al.]. // Frontiers in Immunology. 2018. Vol. 9. C. 1340.

39. Chen H. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R / Chen H., Boutros P. C. // BMC bioinformatics. 2011. Т. 12, № 1. С. 1-7.

40. Cheng, J. Prediction of protein stability changes for single-site mutations using support vector machines / J. Cheng, A. Randall, P. Baldi // Proteins. 2006. № 4 (62). С. 1125–1132.

41. Chitrala, K.N. Computational screening and molecular dynamic simulation of breast cancer associated deleterious non-synonymous single nucleotide polymorphisms in TP53 gene / K.N. Chitrala, S. Yeguvapalli // PloS One. 2014. № 8 (9). С. e104242.

42. Choi Y. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels / Y. Choi, A.P. Chan // Bioinformatics. 2015. Т. 31, № 16. С. 2745-2747.

43. Clinical-Epidemiological Pattern of Primary Immunodeficiencies in Malaysia 1987-2006: A 20 year experience in Four Malaysian Hospitals / L.M. Noh, [et al.]. // The Medical Journal of Malaysia. 2013. № 1 (68). С. 13–17.

44. CISD2 maintains cellular homeostasis / Z.Q. Shen, [et al.] //Biochimica et Biophysica Acta (BBA)-Molecular Cell Research. 2021. Т. 1868, №. 4. С. 118954.

45. Clinical features of dominant and recessive interferon gamma receptor 1 deficiencies / S.E. Dorman, [et al.]. // Lancet (London, England). 2004. № 9451 (364). С. 2113–2121.

46. Comparison of predicted and actual consequences of missense mutations / L.A. Miosge, [et al.]. // Proceedings of the National Academy of Sciences of the United States of America. 2015. № 37 (112). С. E5189-5198.

47. Computational prediction of methylation types of covalently modified lysine and arginine residues in proteins. / W. Deng et al. // Briefings in Bioinformatics, 2017 №4 (18), С. 647–658.

48. Congenital neutropenia: diagnosis, molecular bases and patient management / J. Donadieu, [et al.]. // Orphanet Journal of Rare Diseases. 2011. № 6. С. 26.

49. Congenital neutropenia in the era of genomics: classification, diagnosis, and natural history / J. Donadieu, [et al.]. // British Journal of Haematology. 2017. № 4 (179). С. 557–574.

50. Congenital neutropenia with variable clinical presentation in novel mutation of the SRP54 gene / L. Goldberg, [et al.]. // Pediatric Blood & Cancer. 2020. № 6 (67). С. e28237.

51. Condino-Neto, A. Changing the Lives of People With Primary Immunodeficiencies (PI) With Early Testing and Diagnosis / A. Condino-Neto, F.J. Espinosa-Rosales // Frontiers in Immunology. 2018. № 9. С. 1439.

52. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies / C. Dong, [et al.] // Human molecular genetics. 2015. Т. 24, № 8. С. 2125-2137.

53. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. / H. Ashkenazy, [et al.] // Nucleic Acids Research. 2016. 8.44(W1): W344-50

54. Contribution of Iran in Elucidating the Genetic Causes of Autosomal Recessive Intellectual Disability / R. Ataei, [et al.]. // Archives of Iranian Medicine. 2019. № 8 (22). С. 461–471.

55. Csardi, G. The Igraph Software Package for Complex Network Research / G. Csardi, T. Nepusz // InterJournal. 2005. (Complex Systems). С. 1695.

56. Current Perspectives and Unmet Needs of Primary Immunodeficiency Care in Asia Pacific / D. Leung, [et al.]. // Frontiers in Immunology. 2020. Vol. 11. С. 1605.

57. Current status and prospects of primary immunodeficiency diseases in Asia / R.K. Pilania, [et al.] // Genes & Diseases. 2019. № 1 (7). С. 3–11.

58. Cytoscape: a software environment for integrated models of biomolecular interaction networks / P. Shannon, [et al.]. // Genome Research. 2003. № 11 (13). С. 2498–2504.

59. dbNSFP v3. 0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs / X. Liu, [et al.] // Human mutation. 2016. Т. 37, №. 3. С. 235-241.

60. Defects in TCIRG1 subunit of the vacuolar proton pump are responsible for a subset of human autosomal recessive osteopetrosis / A. Frattini, [et al.]. // Nature Genetics. 2000. № 3 (25). С. 343–346.

61. Diagnostic Tools for Inborn Errors of Human Immunity (Primary Immunodeficiencies and Immune Dysregulatory Diseases) / A.M. Richardson, [et al.]. // Current Allergy and Asthma Reports. 2018. № 3 (18). С. 19.

62. Dominant-negative mutations in the DNA-binding domain of STAT3 cause hyper-IgE syndrome / Y. Minegishi, [et al.]. // Nature. 2007. № 7157 (448). С. 1058–1062.

63. Editorial: Methods and Applications of Computational Immunology / B. Chain, [et al.]. // Frontiers in Immunology. 2019. № 10. С. 2818.

64. Ekblom, R. A field guide to whole-genome sequencing, assembly and annotation / R. Ekblom, J.B.W. Wolf // Evolutionary Applications. 2014. № 9 (7). С. 1026–1042.

65. Epidemiology of congenital neutropenia / J. Donadieu, [et al.] // Hematology/Oncology Clinics. 2013. Т. 27, № 1. С. 1-17.

66. Epidermodysplasia Verruciformis: Inborn Errors of Immunity to Human Beta-Papillomaviruses / S.J. de Jong, [et al.]. // Frontiers in Microbiology. 2018. (9). С. 1222.

67. Evaluation of Clinical Manifestations in Patients with Severe Lymphedema with and without CCBE1 Mutations / M. Alders, [et al.]. // Molecular Syndromology. 2013. № 3 (4). С. 107–113.

68. Expression of the neutrophil elastase gene during human bone marrow cell differentiation // The Journal of Experimental Medicine. 1989. № 3 (169). С. 833–845.

69. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features / M.F. Rogers, [et al.]. // Bioinformatics (Oxford, England). 2018. № 3 (34). С. 511–513.

70. Flanagan S.E. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations / S.E. Flanagan, A.M. Patch, S. Ellard // Genetic testing and molecular biomarkers. – 2010. Т. 14, №. 4. С. 533-537.

71. Gene Ontology Consortium The Gene Ontology (GO) project in 2006 // Nucleic Acids Research. 2006. № Database issue (34). С. D322-326.

72. Genome bioinformatic analysis of nonsynonymous SNPs / D.F. Burke, [et al.] // BMC bioinformatics. 2007. (8). С. 301.

73. Ghosh, S. NF-kappa B and Rel proteins: evolutionarily conserved mediators of immune responses / S. Ghosh, M.J. May, E.B. Kopp // Annual Review of Immunology. 1998. (16). С. 225–260.

74. Global overview of primary immunodeficiencies: a report from Jeffrey Modell Centers worldwide focused on diagnosis, treatment, and discovery / V. Modell, [et al.] // Immunologic Research. 2014. № 1 (60). С. 132–144.

75. Gomber, S. Vaccine Associated Paralytic Poliomyelitis Unmasking Common Variable Immunodeficiency / S. Gomber, V. Arora, P. Dewan // Indian Pediatrics. 2017. № 3 (54). С. 241–242.

76. GPS: A comprehensive www server for phosphorylation sites prediction. / Y. Xue, [et al.] // Nucleic acids research. 2005. Т. 33, № suppl_2. С. W184-W187.

77. Gassoum, A. Comprehensive analysis of rsSNPs associated with hypertension using in-silico bioinformatics tools / A. Gassoum, N. Abdelraheem, N. Elsadig // Open Access Library Journal. 2016. Vol.3, No.7, July P. 1-24.

78. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers / M.J. Abraham, [et al.] //SoftwareX. – 2015. – Т. 1. – С. 19-25.

79. Growth in diagnosis and treatment of primary immunodeficiency within the global Jeffrey Modell Centers Network / J. Quinn, [et al.]. // Allergy, Asthma, and Clinical Immunology: Official Journal of the Canadian Society of Allergy and Clinical Immunology. 2022. № 1 (18). С. 19.

80. Guidelines for genetic studies in single patients: lessons from primary immunodeficiencies / J.-L. Casanova, [et al.]. // The Journal of Experimental Medicine. 2014. № 11 (211). С. 2137–2149.

81. Hauck, F. Pathogenic mechanisms and clinical implications of congenital neutropenia syndromes / F. Hauck, C. Klein // Current Opinion in Allergy and Clinical Immunology. 2013. № 6 (13). С. 596–606.

82. Hecht M. Better prediction of functional effects for sequence variants. / M. Hecht, Y. Bromberg, B. Rost // BMC Genomics. 2015; 16, Прил. 8. S.1.

83. Hennekam syndrome can be caused by FAT4 mutations and be allelic to Van Maldergem syndrome / M. Alders, [et al.]. // Human Genetics. 2014. № 9 (133). С. 1161–1167.

84. Hershfield, M.S. Genotype is an important determinant of phenotype in adenosine deaminase deficiency // Current Opinion in Immunology. 2003. № 5 (15). С. 571–577.

85. Highly accurate protein structure prediction with AlphaFold / J. Jumper, [et al.]. // Nature. 2021. № 7873 (596). С. 583–589.

86. Hildebrand, P.W. Bringing Molecular Dynamics Simulation Data into View / P.W. Hildebrand, A.S. Rose, J.K.S. Tiemann // Trends in Biochemical Sciences. 2019. № 11 (44). С. 902–913.

87. Hilliard, R.I. Congenital abnormalities of the lymphatic system: a new clinical classification / R.I. Hilliard, J.B. McKendry, M.J. Phillips // Pediatrics. 1990. № 6 (86). С. 988–994.

88. Horvath, S. Geometric interpretation of gene coexpression network analysis / S. Horvath, J. Dong // PLoS computational biology. 2008. № 8 (4). C. e1000117.

89. Human Inborn Errors of Immunity: 2022 Update on the Classification from the International Union of Immunological Societies Expert Committee / S.G. Tangye, [et al.]. // Journal of Clinical Immunology. 2022. № 7 (42). C. 1473–1507.

90. Human Inborn Errors of Immunity: 2019 Update of the IUIS Phenotypical Classification / A. Bousfiha, [et al.]. // Journal of Clinical Immunology. 2020. № 1 (40). C. 66–81.

91. Human Protein Reference Database and Human Proteinpedia as resources for phosphoproteome analysis/ Goel R. et al. //Molecular BioSystems. – 2012. – T. 8. – №. 2. – C. 453-463.

92. Hypermutable Non-Synonymous Sites Are under Stronger Negative Selection / S. Schmidt, [et al.]. // PLOS Genetics. 2008. № 11 (4). C. e1000281.

93. Identification, analysis, and prediction of protein ubiquitination sites / P. Radivojac, et al. // Proteins: Structure, Function, and Bioinformatics. 2010. T. 78. № 2. C. 365-380.

94. Identification of candidate disease genes in patients with common variable immunodeficiency/ G. Liu, [et al.] // Quantitative Biology. 2019. № 3 (7). C. 190–201

95. Immunodeficiency, autoinflammation and amylopectinosis in humans with inherited HOIL-1 and LUBAC deficiency / B. Boisson, [et al.]. // Nature Immunology. 2012. № 12 (13). C. 1178–1186.

96. Inducible expression of a disease-associated ELANE mutation impairs granulocytic differentiation, without eliciting an unfolded protein response / B. Garg [et al.] //Journal of Biological Chemistry. 2020. T. 295, № 21. C. 7492-7500.

97. In Silico Analysis Revealed Five Novel High-Risk Single-Nucleotide Polymorphisms (rs200384291, rs201163886, rs193141883, rs201139487, and rs201723157) in ELANE Gene Causing Autosomal Dominant Severe Congenital

Neutropenia 1 and Cyclic Hematopoiesis / K. Shinwari, [et al.]. // The Scientific World Journal. 2022. № 6. С. 3356835.

98. Increasing incidence of human melioidosis in Northeast Thailand / D. Limmathurotsakul, [et al.]. // The American Journal of Tropical Medicine and Hygiene. 2010. № 6 (82). С. 1113–1117.

99. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2 / V. Pejaver, [et al.]. // Nature Communications. 2020. № 1 (11). С. 5918.

100. InnateDB: systems biology of innate immunity and beyond-recent updates and continuing curation / K. Breuer, [et al.]. // Nucleic Acids Research. 2013. № Database issue (41). С. D1228-1233.

101. INstruct: a database of high-quality 3D structurally resolved protein interactome networks / M.J. Meyer, [et al.]. // Bioinformatics (Oxford, England). 2013. № 12 (29). С. 1577–1579.

102. IntAct: an open source molecular interaction database / H. Hermjakob, [et al.]. // Nucleic Acids Research. 2004. № Database issue (32). С. D452-455.

103. Intrinsic and extrinsic causes of malignancies in patients with primary immunodeficiency disorders / F. Hauck, [et al.]. // The Journal of Allergy and Clinical Immunology. 2018. № 1 (141). С. 59-68. e4.

104. Introduction on primary immunodeficiency diseases / N. Rezaei [et al.]; под ред. N. Rezaei, A. Aghamohammadi, L.D. Notarangelo. Springer Verlag, 2017. С. 1–81.

105. Investigation of deleterious effects of nsSNPs in the POT1 gene: a structural genomics-based approach to understand the mechanism of cancer development / M. Amir, [et al.]. // Journal of Cellular Biochemistry. 2019. № 6 (120). С. 10281–10294.

106. iStable 2.0: Predicting protein thermal stability changes by integrating various characteristic modules / C.-W. Chen. [et al.]. // Computational and Structural Biotechnology Journal. 2020. (18). С. 622–630.

107. Itan, Y. Novel Primary Immunodeficiency Candidate Genes Predicted by the Human Gene Connectome / Y. Itan, J.-L. Casanova // Frontiers in Immunology. 2015. (6). C. 142.

108. Iwasaki, A. Regulation of adaptive immunity by the innate immune system / A. Iwasaki, R. Medzhitov // Science (New York, N.Y.). 2010. № 5963 (327). C. 291–295.

109. Kanehisa, M. KEGG: kyoto encyclopedia of genes and genomes / M. Kanehisa, S. Goto // Nucleic Acids Research. 2000. № 1 (28). C. 27–30.

110. Kohn L.A. Gene Therapies for Primary Immune Deficiencies / L.A. Kohn, D.B. Kohn. // Frontiers in Immunology. 2021. № 12 C.648951

111. Komander, D. The ubiquitin code / D. Komander, M. Rape // Annual Review of Biochemistry. 2012. (81). C. 203–229.

112. Ku, B.C. Neutropenia in the Febrile Child / B.C. Ku, C. Bailey, F. Balamuth // Pediatric Emergency Care. 2016. № 5 (32). C. 329–334.

113. Lee, P.P.W. Cellular and Molecular Defects Underlying Invasive Fungal Infections-Revelations from Endemic Mycoses / P.P. Lee, Y.-L. Lau // Frontiers in Immunology. 2017. (8). C. 735.

114. Lee, P. P.W. Endemic infections in Southeast Asia provide new insights to the phenotypic spectrum of primary immunodeficiency disorders / P.P.W. Lee, Y.-L. Lau // Asian Pacific Journal of Allergy and Immunology. 2013. № 3 (31). C. 217–226.

115. Lee, P.P.W. Improving care, education, and research: the Asian primary immunodeficiency network / P.P.W. Lee, Y.-L. Lau // Annals of the New York Academy of Sciences. 2011. Vol. 1238. C. 33–41.

116. Lee, P.P.W. Primary immunodeficiencies: «new» disease in an old country / P.P.W. Lee, Y.-L. Lau // Cellular & Molecular Immunology. 2009. № 6 (6). C. 397–406.

117. Leiding, J.W. Warts and All: HPV in Primary Immunodeficiencies / J.W. Leiding, S.M. Holland // The Journal of allergy and clinical immunology. 2012. № 5 (130). C. 1030–1048.

118. Lessons learned from the study of human inborn errors of innate immunity / G. Bucciol, [et al.]. // The Journal of Allergy and Clinical Immunology. 2019. № 2 (143). С. 507–527.

119. Li, H. Fast and accurate short read alignment with Burrows-Wheeler transform / Li H., Durbin R. // Bioinformatics (Oxford, England). 2009. № 14 (25). С. 1754–1760.

120. Llimma powers differential expression analyses for RNA-sequencing and microarray studies / M.E. Ritchie, [et al.] // Nucleic acids research. 2015. Т. 43, № 7. С. e47-e47.

121. Linear ubiquitination prevents inflammation and regulates immune signalling / B. Gerlach, [et al.]. // Nature. 2011. № 7340 (471). С. 591–596.

122. Liu B. C. Host-intrinsic interferon status in infection and immunity / B.C. Liu, J. Sarhan, A. Poltorak //Trends in molecular medicine. 2018. Т. 24, № 8. С. 658-668.

123. Loging, W. High-throughput electronic biology: mining information for drug discovery / W. Loging, L. Harland, B. Williams-Jones // Nature Reviews Drug Discovery. 2007. № 3 (6). С. 220–230.

124. Loss of ADAMTS3 activity causes Hennekam lymphangiectasia-lymphedema syndrome 3 / P. Brouillard, [et al.]. // Human Molecular Genetics. 2017. № 21 (26). С. 4095–4104.

125. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity / K.A. Jagadeesh, [et al.] //Nature genetics. 2016. Т. 48. № 12. С. 1581-1586.

126. Maere, S. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks / S. Maere, K. Heymans, M. Kuiper // Bioinformatics (Oxford, England). 2005. № 16 (21). С. 3448–3449.

127. Malaysia’s First Transplanted Case of Chronic Granulomatous Disease: The Journey of Overcoming Obstacles / I.H. Ismail, [et al.]. // Children (Basel, Switzerland). 2016. № 2 (3). С. 9.

128. Malaviya, A.N. Ataxia telangiectasia: immunological abnormalities in probands and first degree relatives in 5 families / A.N. Malaviya, K.K. Sachdeva, N. Singh // The Journal of the Association of Physicians of India. 1973. № 8 (21). С. 701–705.

129. McDermott, D.H. WHIM syndrome: Immunopathogenesis, treatment and cure strategies / D.H. McDermott, P.M. Murphy // Immunological Reviews. 2019. № 1 (287). С. 91–102.

130. Medzhitov, R. Decoding the patterns of self and nonself by the innate immune system / R. Medzhitov, C.A. Janeway // Science (New York, N.Y.). 2002. № 5566 (296). С. 298–300.

131. Mehta, S. R. Agammaglobulinaemia / S.R. Mehta, L.R. Sarin, L.M. Sanghvi // Journal of the Indian Medical Association. 1964. Vol. 42. С. 539–541.

132. Molecular docking and structure-based drug design strategies / L.G. Ferreira, [et al.]. // Molecules (Basel, Switzerland). 2015. № 7 (20). С. 13384–13421.

133. Molecular markers in bladder cancer / F. Soria, [et al.]. // World Journal of Urology. 2019. № 1 (37). С. 31–40.

134. Mooney, M.A. Gene set analysis: A step-by-step guide / M.A. Mooney, B. Wilmot // American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics. 2015. № 7 (168). С. 517–527.

135. MutationTaster evaluates disease-causing potential of sequence alterations/ Schwarz J. M. et al. //Nature methods. – 2010. – Т. 7. – №. 8. – С. 575-576.

136. Mutations in CCBE1 cause generalized lymph vessel dysplasia in humans / M. Alders, [et al.]. // Nature Genetics. 2009. № 12 (41). С. 1272–1274.

137. Mutations in ELA2, encoding neutrophil elastase, define a 21-day biological clock in cyclic haematopoiesis / M. Horwitz, [et al.]. // Nature Genetics. 1999. № 4 (23). С. 433–436.

138. Mutations in the a3 subunit of the vacuolar H(+)-ATPase cause infantile malignant osteopetrosis / U. Kornak, [et al.]. // Human Molecular Genetics. 2000. № 13 (9). С. 2059–2063.

139. Mutations in the ELA2 gene correlate with more severe expression of neutropenia: a study of 81 patients from the French Neutropenia Register / C. Bellanné-Chantelot, [et al.]. // Blood. 2004. № 11 (103). С. 4119–4125.

140. Mutations in the gene encoding neutrophil elastase in congenital and cyclic neutropenia. / Dale D.C. et al. / Blood, The Journal of the American Society of Hematology. – 2000. – Т. 96. – №. 7. – С. 2317-2322

141. MutationTaster evaluates disease-causing potential of sequence alterations / J.M. Schwarz, [et al.]. // Nature Methods. 2010. № 8 (7). С. 575–576.

142. Nationwide survey of patients with primary immunodeficiency diseases in Japan / M. Ishimura, [et al.]. // Journal of Clinical Immunology. 2011. № 6 (31). С. 968–976.

143. Network-based approach to prediction and population-based validation of in silico drug repurposing / F. Cheng, [et al.]. // Nature Communications. 2018. № 1 (9). С. 2691.

144. Neutropenia and primary immunodeficiency diseases / N. Rezaei, [et al.]. // International Reviews of Immunology. 2009. № 5 (28). С. 335–366.

145. Ng, P.C. SIFT: Predicting amino acid changes that affect protein function / P.C. Ng, S. Henikoff // Nucleic Acids Research. 2003. № 13 (31). С. 3812–3814.

146. Notarangelo, L.D. Primary immunodeficiencies: novel genes and unusual presentations / Notarangelo, L.D. G. Uzel, V.K. Rao // Hematology. American Society of Hematology. Education Program. 2019. № 1 (2019). С. 443–448.

147. Novel HAX1 mutations in patients with severe congenital neutropenia reveal isoform-dependent genotype-phenotype associations / M. Germeshausen, [et al.]. // Blood. 2008. № 10 (111). С. 4954–4957.

148. Osteopetrosis: genetics, treatment and new insights into osteoclast function / C. Sobacchi, [et al.]. // Nature Reviews Endocrinology. 2013. № 9 (9). C. 522–536.

149. Padron G.T. Autoimmunity in Primary Immunodeficiencies (PID) / G.T. Padron, V.P. Hernandez-Trujillo// Clinical Reviews in Allergy & Immunology. – 2022. – C. 1-18.

150. PANTHER: a library of protein families and subfamilies indexed by function / P. D. Thomas, [et al.] //Genome research. 2003. Т. 13, № 9. С. 2129-2141.

151. Papadopoulou-Alataki, E. Prevention of infection in children and adolescents with primary immunodeficiency disorders / E. Papadopoulou-Alataki, A. Hassan, E.G. Davies // Asian Pacific Journal of Allergy and Immunology. 2012. № 4 (30). C. 249–258.

152. Paradis, E. APE: Analyses of Phylogenetics and Evolution in R language / E. Paradis, J. Claude, K. Strimmer // Bioinformatics (Oxford, England). 2004. № 2 (20). C. 289–290.

153. Patients with Primary Immunodeficiencies Are a Reservoir of Poliovirus and a Risk to Polio Eradication / A. Aghamohammadi, [et al.]. // Frontiers in Immunology. 2017. Vol. 8. C. 685.

154. PDBe: towards reusable data delivery infrastructure at protein data bank in Europe / S. Mir, [et al.]. // Nucleic Acids Research. 2018. № D1 (46). C. D486–D492.

155. Penicillium marneffei infection and impaired IFN-γ immunity in humans with autosomal-dominant gain-of-phosphorylation STAT1 mutations / P.P.W. Lee, [et al.]. // The Journal of Allergy and Clinical Immunology. 2014. № 3 (133). C. 894-896.e5.

156. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse / P.V. Hornbeck, [et al.]. // Nucleic Acids Research. 2012. № Database issue (40). C. D261-270.

157. Phylogenetic perspectives in innate immunity / J.A. Hoffmann, [et al.]. // Science (New York, N.Y.). 1999. № 5418 (284). С. 1313–1318.

158. Picard, C. Infectious diseases in patients with IRAK-4, MyD88, NEMO, or IκBα deficiency / C. Picard, J.-L. Casanova, A. Puel // Clinical Microbiology Reviews. 2011. № 3 (24). С. 490–497.

159. PIDO: the primary immunodeficiency disease ontology / N. Adams, [et al.]. // Bioinformatics (Oxford, England). 2011. № 22 (27). С. 3193–3199.

160. PINA v2.0: mining interactome modules / M.J. Cowley, [et al.]. // Nucleic Acids Research. 2012. № Database issue (40). С. D862-865.

161. Polyglucosan body myopathy caused by defective ubiquitin ligase RBCK1 / J. Nilsson, [et al.]. // Annals of Neurology. 2013. № 6 (74). С. 914–919.

162. Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology / C. Steentoft, [et al.] //The EMBO journal. 2013. Т. 32, № 10. С. 1478-1488.

163. Prediction of Candidate Primary Immunodeficiency Disease Genes Using a Support Vector Machine Learning Approach / S. Keerthikumar, [et al.]. // DNA Research. 2009. № 6 (16). С. 345–351.

164. Prevalence and Outcomes of Primary Immunodeficiency in Hospitalized Children in the United States / Z. Rubin, [et al.]. // The Journal of Allergy and Clinical Immunology. In Practice. 2018. № 5 (6). С. 1705-1710.e1.

165. Primary immunodeficiencies in India: a perspective / S. Gupta, [et al.]. // Annals of the New York Academy of Sciences. 2012. Vol. 1250. С. 73–79.

166. Primary Immunodeficiencies in Russia: Data From the National Registry / A.A. Mukhina, [et al.]. // Frontiers in Immunology. 2020. Vol. 11. С. 1491.

167. Primary Immunodeficiency Diseases; A 20 Years Experience in a Tertiary University Hospital at Ramathibodi / O. Luecha, [et al.]. // Journal of Allergy and Clinical Immunology. 2012. № 2 (129). С. AB158.

168. Primary immunodeficiency diseases in Singapore--the last 11 years / D.L. Lim, [et al.]. // Singapore Medical Journal. 2003. № 11 (44). С. 579–586.

169. Primary Immunodeficiency Disorders in India-A Situational Review / A.K. Jindal, [et al.]. // Frontiers in Immunology. 2017. Vol. 8. C. 714.

170. PROCHECK: a program to check the stereochemical quality of protein structures / R.A. Laskowski, [et al.] // Journal of applied crystallography. 1993. Т. 26, № 2. С. 283-291.

171. Protein structure prediction / S. Agnihotry, [et al.] // Bioinformatics. Academic Press, 2022. С. 177-188.

172. Quantitative network mapping of the human kinome interactome reveals new clues for rational kinase inhibitor discovery and individualized cancer therapy / F. Cheng, [et al.]. // Oncotarget. 2014. № 11 (5). С. 3697–3710.

173. Ramachandran GN. Stereochemistry of polypeptide chain configurations. / Ramachandran GN, Ramakrishnan C, Sasisekharan V. // J. mol. Biol. – 1963. – Т. 7. – С. 95-99.

174. RBCK1 negatively regulates tumor necrosis factor- and interleukin-1-triggered NF-kappaB activation by targeting TAB2/3 for degradation / Y. Tian, [et al.]. // The Journal of Biological Chemistry. 2007. № 23 (282). С. 16776–16782.

175. Recurrent staphylococcal cellulitis and subcutaneous abscesses in a child with autoantibodies against IL-6 / A. Puel, [et al.]. // Journal of Immunology (Baltimore, Md.: 1950). 2008. № 1 (180). С. 647–654.

176. Reva B. Predicting the functional impact of protein mutations: application to cancer genomics/ Reva B., Antipin Y., Sander C. //Nucleic acids research. – 2011. – Т. 39. – №. 17. – С. e118-e118.

177. Revisiting human IL-12Rβ1 deficiency: a survey of 141 patients from 30 countries / L. de Beaucoudrey, [et al.]. // Medicine. 2010. № 6 (89). С. 381–402.

178. Revisiting human primary immunodeficiencies / J.-L. Casanova, [et al.]. // Journal of Internal Medicine. 2008. № 2 (264). С. 115–127.

179. Robert, F. Exploring the Impact of Single-Nucleotide Polymorphisms on Translation / F. Robert, J.Pelletier // Frontiers in Genetics. 2018. Vol. 9. C. 507.

180. Rodriguez-Esteban, R. Differential gene expression in disease: a comparison between high-throughput studies and the literature / R. Rodriguez-Esteban, X. Jiang // BMC medical genomics. 2017. № 1 (10). С. 59.

181. Rosenzweig, S.D. Defects in the interferon-gamma and interleukin-12 pathways / S.D. Rosenzweig, S.M. Holland // Immunological Reviews. 2005. (203). С. 38–47.

182. Roy, A. I-TASSER: a unified platform for automated protein structure and function prediction / A. Roy, A. Kucukural, Y. Zhang // Nature Protocols. 2010. № 4 (5). С. 725–738.

183. Samarghitean, C. Bioinformatics services related to diagnosis of primary immunodeficiencies / C. Samarghitean, M. Vihinen // Current Opinion in Allergy and Clinical Immunology. 2009. № 6 (9). С. 531–536.

184. Sasaki, Y. Crucial Role of Linear Ubiquitin Chain Assembly Complex-Mediated Inhibition of Programmed Cell Death in TLR4-Mediated B Cell Responses and B1b Cell Development / Y. Sasaki, K. Iwai // Journal of Immunology (Baltimore, Md.: 1950). 2018. № 10 (200). С. 3438–3449.

185. Scalable algorithms for molecular dynamics simulations on commodity clusters / K.J. Bowers, [et al.] // Proceedings of the 2006 ACM/IEEE Conference on Supercomputing. 2006. С. 84-es.

186. Second Case of HOIP Deficiency Expands Clinical Features and Defines Inflammatory Transcriptome Regulated by LUBAC / H. Oda, [et al.]. // Frontiers in Immunology. 2019. (10). С. 479.

187. Severe congenital neutropenia: inheritance and pathophysiology / J. Skokowa, [et al.]. // Current Opinion in Hematology. 2007. № 1 (14). С. 22–28.

188. Severe congenital neutropenias / J. Skokowa [et al.]. // Nature Reviews. Disease Primers. 2017. (3). С. 17032.

189. SHARPIN is a component of the NF-κB-activating linear ubiquitin chain assembly complex / Tokunaga F. [et al.]. // Nature. 2011. № 7340 (471). С. 633–636.

190. Shore, G.C. Signaling cell death from the endoplasmic reticulum stress response / G.C. Shore, F.R. Papa, S.A. Oakes // Current Opinion in Cell Biology. 2011. № 2 (23). С. 143–149.

191. SignaLink 2 - a signaling pathway resource with multi-layered regulatory networks / D. Fazekas, [et al.]. // BMC systems biology. 2013. № 7. С. 7.

192. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets / D. Szklarczyk, [et al.]. // Nucleic Acids Research. 2019. № D1 (47). С. D607–D613.

193. Survival of mature T cells depends on signaling through HOIP / K. Okamura, [et al.]. // Scientific Reports. 2016. Vol. 6. С. 36135.

194. Susceptibility to mycobacterial infections in children with X-linked chronic granulomatous disease: a review of 17 patients living in a region endemic for tuberculosis / P.P.W. Lee, [et al.]. // The Pediatric Infectious Disease Journal. 2008. № 3 (27). С. 224–230.

195. Systems Bioinformatics: increasing precision of computational diagnostics and therapeutics through network-based approaches / A. Oulas, [et al.]. // Briefings in Bioinformatics. 2017. № 3 (20). С. 806–824.

196. Systems Biology as a Comparative Approach to Understand Complex Gene Expression in Neurological Diseases / L. Diaz-Beltran, [et al.]. // Behavioral Sciences. 2013. № 2 (3). С. 253–272.

197. Takada, H. Primary immunodeficiency in Japan; epidemiology, diagnosis, and pathogenesis // Pediatrics International: Official Journal of the Japan Pediatric Society. 2013. № 6 (55). С. 671–674.

198. TCIRG1-dependent recessive osteopetrosis: mutation analysis, functional identification of the splicing defects, and in vitro rescue by U1 snRNA / L. Susani, [et al.]. // Human Mutation. 2004. № 3 (24). С. 225–235.

199. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data / A. McKenna, [et al.]. // Genome Research. 2010. № 9 (20). С. 1297–1303.

200. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists / D.W. Huang, [et al.]. // Genome Biology. 2007. № 9 (8). C. R183.

201. The human gene connectome as a map of short cuts for morbid allele discovery / Y. Itan, [et al.]. // Proceedings of the National Academy of Sciences of the United States of America. 2013. № 14 (110). C. 5558–5563.

202. The mutational spectrum of human malignant autosomal recessive osteopetrosis / C. Sobacchi, [et al.]. // Human Molecular Genetics. 2001. № 17 (10). C. 1767–143.

203. The Sequence Alignment/Map format and SAMtools / H. Li, [et al.] // Bioinformatics (Oxford, England). 2009. № 16 (25). C. 2078–2079.

204. The variant call format and VCFtools / P. Danecek, [et al.]. // Bioinformatics (Oxford, England). 2011. № 15 (27). C. 2156–2158.

205. Using the MINT database to search protein interactions / Calderone A. et al. //Current Protocols in Bioinformatics. – 2020. – Т. 69. – №. 1. – С. e93.

206. Wang, K. ANNOVAR functional annotation of genetic variants from high-throughput sequencing data / K. Wang, M. Li, H.K. Hakonarson // ANNOVAR // Nucleic Acids Research. 2010. № 16 (38). C. e164.

207. The Phyre2 web portal for protein modeling, prediction and analysis / L.A. Kelley, [et al.]. // Nature Protocols. 2015. № 6 (10). C. 845–858.

208. The prevalences [correction] and patient characteristics of primary immunodeficiency diseases in Turkey - two centers study / S.S. Kilic, [et al.]. // Journal of Clinical Immunology. 2013. № 1 (33). C. 74–83.

209. The Primary Immunodeficiency Database in Japan / K. Mitsui-Sekinaka, [et al.]. // Frontiers in Immunology. 2022. Vol. 12.

210. UCSF Chimera--a visualization system for exploratory research and analysis / E.F. Pettersen, [et al.]. // Journal of computational chemistry. 2004. № 13 (25). C. 1605–1612.

211. Uses of Next-Generation Sequencing Technologies for the Diagnosis of Primary Immunodeficiencies / M. Seleman, [et al.]. // Frontiers in Immunology. 2017. (8). C. 847.

212. VarCards: an integrated genetic and clinical database for coding variants in the human genome / J. Li, [et al.] // Nucleic Acids Research. 2018. № D1 (46). C. D1039–D1048.

213. Visualizing the drug target landscape / S.J. Campbell, [et al.]. // Drug Discovery Today. 2010. № 1–2 (15). C. 3–15.

214. Wiskott-Aldrich syndrome. A case report / M.C. Gupta, [et al.]. // The Journal of the Association of Physicians of India. 1964. Vol. 12. C. 531–533.

215. Wu J. Prediction of deleterious nonsynonymous single-nucleotide polymorphism for human diseases / Wu J., Jiang R. //The Scientific World Journal. – 2013. – T. 2013.

216. X-linked agammaglobulinemia: Twenty years of single-center experience from North West India / S. Singh, [et al.]. // Annals of Allergy, Asthma & Immunology: Official Publication of the American College of Allergy, Asthma, & Immunology. 2016. № 4 (117). C. 405–411.

217. clusterProfiler: an R package for comparing biological themes among gene clusters / G. Yu, [et al.]. // Omics: A Journal of Integrative Biology. 2012. № 5 (16). C. 284–287.

218. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data / T. Wang, [et al.]. // BMC bioinformatics. 2019. № 1 (20). C. 40.

219. Distribution, clinical features and molecular analysis of primary immunodeficiency diseases in Chinese children: a single-center study from 2005 to 2011 / Z.-Y. Zhang, [et al.]. // The Pediatric Infectious Disease Journal. 2013. № 10 (32). C. 1127–1134.

220. Inborn errors of interferon (IFN)-mediated immunity in humans: insights into the respective roles of IFN-alpha/beta, IFN-gamma, and IFN-lambda in

host defense / S.-Y. Zhang, [et al.]. // Immunological Reviews. 2008. (226). С. 29–40.

221. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets / Y., Zhou [et al.]. // Nature Communications. 2019. № 1 (10). С. 1523.

222. Wang, K. A brief procedure for big data analysis of gene expression / K. Wang, W. Wang, M. Li // Animal Models and Experimental Medicine. 2018. № 3 (1). С. 189–193.

223. Welte, K. Severe congenital neutropenia / K. Welte, C. Zeidler, D.C. Dale // Seminars in Hematology. 2006. № 3 (43). С. 189–195.

224. Yang, C.-H. Chromobacterium violaceum infection: a clinical review of an important but neglected infection / C.-H. Yang, Y.-H. Li // Journal of the Chinese Medical Association: JCMA. 2011. № 10 (74). С. 435–441.

225. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles | PNAS [Электр. ресурс]. URL: https://www.pnas.org/doi/10.1073/pnas.0506580102 (дата обращения: 06.01.2023).

226. Lymphedema-lymphangiectasia-mental retardation (Hennekam) syndrome: a review. - PubMed [Электр. ресурс]. URL: https://pubmed.ncbi.nlm.nih.gov/12376947/ (дата обращения: 29.12.2022).

227. Primary Immunodeficiency Diseases: an Update on the Classification from the International Union of Immunological Societies Expert Committee for Primary Immunodeficiency / C. Picard, [et al.]. // Journal of Clinical Immunology. 2015. № 8 (35). С. 696–726.

228. Primary Immune Deficiency Diseases in America: The Third National Survey of Patients (2007) | SCID Compass [Электр. ресурс]. URL: https://scidcompass.org/publication/surveys/primary-immune-deficiency-diseases-america-third-national-survey-patients-2007 (дата обращения: 19.12.2022).

ABBREVIATIONS

1000g	1000 Genomes
2D	2 dimentiaonal
3D	3 dimentional
ANC	Absolute neutrophil counts
AR	Autosomal recicive
BCG	Bacillus Calmette Guerin
BWA	Burrows-Wheeler Aligner
CCBE1	Collagen and calcium-binding protein 1 containing an epidermal growth factor-like domain
CGD	CGD: chronic granulomatous disease
CINCA	Chronic Infantile Neurological Cutaneous and Articular
CISD2	CDGSH iron sulfur domain 2
CN	Congenital neutropenia
CXCR4	Chemokine receptor 4
CADD	Combined Annotation Dependent Depletion
DANN	Annotating genetic variants, especially non-coding variants
dbSNP	Single Nucleotide Polymorphism Database
DEG	Differentially expressed gene
ELANE	Elastase, Neutrophil Expressed
ExAC	Exome Aggregation Consortium
FAT4	FAT atypical cadherin 4
FATHMM	Functional Analysis through Hidden Markov Models

FATHMM-MKL	Functional Analysis through Hidden Markov Models - Multiple Kernel Learning
FDR	False Discovery Frequency
FGA	Functional genomic alignment
FEL	Free energy landscape
GAD	Genetic Association Database
GOF	Gain of Function
GATK	The Genome Analysis Toolkit
GEO	Gene Expression Omnibus
gnomAD	The Genome Aggregation Database
GO	Gene Ontology
GSEA	Gene set enrichment analysis
HapMap	Haplotype Mapping Project
HGMP	Human Genome Mapping Project
HGC	Human Genome Connectome
HH-Pred	Homology detection and structure prediction by HMM-HMM comparison
HLH	Hemophagocytic lymphohistiocytosis
HOLP1	HOIL-1-interacting protein
HOL1	Heme-oxidized IRP2 ubiquitin ligase-1
HS	Hennkam syndrom
HSCT	Hematopoietic stem cell transplantation
IEI	Inborn errors of Immunity
IFN-γ	Interferon Gamma
IFNGIR1	Interferon-gamma receptor 1
IgA	Immunoglobuline A
IgM	Immunoglobuline M
IGV	Integrative Genome Viewer
IL-1β	Interleukin-1 beta

I-mutant	In Silico MUTation Analysis
I-Tasser	Iterative Threading ASSEmbly Refinement
IUIS	International Union of Immunological Societies
IVIG	Intravenous immunoglobulin therapy
iVDPVs	immunodeficiency-associated polioviruses obtained through vaccination
JMF	Jeffrey Modell Foundation
KEGG	Kyoto Encyclopedia of Genes and Genomes
KREC	Kappa-deleting recombination excision circle
LOF	Lost of Function
LRT	Likelihood Ratio Test
LUBAC	Linear Ubiquitin Chain Assembly Complex
MAF	Minor allele frequency
M-CAP	Mendelian Clinically Applicable Pathogenicity
MDS	Molecular Dynamics Simulation
MuPred	Mutation Prediction
MSMD	Mendelian susceptibility to mycobacterial diseases
MTOR	Mechanistic Target of Rapamycin
MU-Pro	"Mutation Protein
MYD88	Myeloid Differentiation primary response gene 88
NAEPID	National Association of Experts on Primary Immunodeficiencies
NCBI	National Center for Biotechnology Information
NEMO	NF-κB essential modulator
NES	Normalized enrichment score
NF-Kb	Nuclear factor-kB
NGS	Next-generation sequencing

NOMID	Neonatal-Onset Multisystem Inflammatory Disease
nsSNPs	non-synonymous single nucleotide polymorphisms
OMIM	Online Mendelian Inheritance in Man
PANTHER	Protein ANalysis THrough Evolutionary Relationships
PCA	Principal Component Analysis
PCC	Pearson's correlation analysis
PDB	Protien data bank
PHD-SNP	Predictor of Human Deleterious Single Nucleotide Polymorphisms
PIK3/AKT	Phosphoinositide 3-kinase/Protein Kinase B
PID	Primary immunodeficiency
PolyPhen2	Polymorphism Phenotyping v2
PPI	Protein-protein interaction
PROVEAN	Protein Variation Effect Analyzer
PBMCs	Peripheral blood mononuclear cells
PTM	Post translation modification
RBCK1	RanBP-type and C3HC4-type zinc finger-containing protein 1
Rg	Radius of gyaration
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
SASA	Solvent Accessible Surface Area
SCN	Severe congenital neutropenia
SCID	Severe combined immune deficiency
SHARPIN	SHANK-associated RH domain interactor
SIFT	Sorting Intolerant From Tolerant
SNAP2	Scaled Neural Network Aligning Profile 2

SNPs	Single nucleotide polymorphisms
SNVs	Single nucleotide variants
SNP-GO	Single Nucleotide Polymorphism Gene Ontology
SOPMA	Self-Optimized Prediction Method with Alignment
SSE	Secondary Structure Element
STAT1	Signal Transducer and Activator of Transcription 1
STRINGS	Search Tool for the Retrieval of Interacting Genes/Proteins
TCIRG1	T-cell immune regulator 1
TLR	Toll-like receptor
TMS	Template Modeling score
TNF	Tumor necrosis factor
TREC	T-cell receptor excision circles
TYK2	Tyrosine Kinase 2
Uniprot	Universal Protein Resource
USA	United states of America
VARCARDS	Variant Interpretation Cards
VAK	Higher Attestation Commission (of the Russian Academy of Sciences)
VCF	Variant Call Format
VEGF-C	Vascular Endothelial Growth Factor C
VEST	Variant Effect Scoring Tool"
WAS	Wiskott-Aldrich syndrome
WGS	Whole genome sequencing
WHIM	Warts, Hypogammaglobulinemia, Infections, and Myelokathexis
XLA	X-linked agammaglobulinemia