ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

FAKULTA BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

Katedra informačních a komunikačních technologií v lékařství

Bioinformatics analysis of genomic data of patients with atypical Wilson disease

Bachelor's Thesis

Study Programme:	Biomedicínská a klinická technika
Field of Study:	Informační a komunikační technologie v lékařství
Author:	Petr Štěpánek
Supervisor:	Mgr. Veronika Vymětalová, Ph.D.
Consultant:	doc. Vladimír Rogalewicz, CSc.

Kladno 2022

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení:	**Štěpánek**	Jméno:	**Petr**	Osobní číslo:	**491816**
Fakulta:	**Fakulta biomedicínského inženýrství**				
Garantující katedra:	**Katedra informačních a komunikačních technologií v lékařství**				
Studijní program:	**Biomedicínská a klinická technika**				
Studijní obor:	**Informační a komunikační technologie v lékařství**				

II. ÚDAJE K BAKALÁŘSKÉ PRÁCI

Název bakalářské práce:

Bioinformatická analýza genomových dat pacientů s atypickou Wilsonovou chorobou

Název bakalářské práce anglicky:

Bioinformatics analysis of genomic data of patients with atypical Wilson disease

Pokyny pro vypracování:

Perform a literature review focused on Wilson disease, autosomal recessive inherited metabolic disease. Describe the genetic factors influencing the origin and development of this disease, clinical manifestations and spread in the population in the Czech Republic and in the world. Process a comprehensive analysis of genomes in a family with atypical Wilson disease from the obtained data. Develop and describe a bioinformatics procedure for DNA sequencing of patients with atypical disease. Analyze rare variants in the genome of members of the family under study. Using appropriate methods, identify causal changes leading to the manifestation of the phenotype associated with Wilson disease. Design and select genomic and bioinformatics procedures and methods applicable to the screening and prevention of this disease. Analyze your results and discuss them.

Seznam doporučené literatury:

[1] Nussbaum R. at al Thompson & Thompson , Genetics in Medicine, Elsevier, 2015, ISBN 1437706967
[2] Hoffmann G. F., Dědičné metabolické poruchy, Grada Publishing a.s., 2006, ISBN 80-247-0831-0
[3] Hepsyba S.H., Basic Bioinformatics, Harrow, Middlesex, U.K. , Alpha Science International, 2013, ISBN 1842658042
[4] Hepsyba S.H., Basic Bioinformatics, ed. MJP, 2019, Přístupné z: https://play.google.com/store/books/details/Basic_Bioinformatics?id=GYOcDwAAQBAJ&gl=US

Jméno a příjmení vedoucí(ho) bakalářské práce:

Mgr. Veronika Vymětalová, Ph.D.

Jméno a příjmení konzultanta(ky) bakalářské práce:

doc. Vladimír Rogalewicz, CSc.

Datum zadání bakalářské práce: **14.02.2022**
Platnost zadání bakalářské práce: **18.09.2023**

doc. Ing. Karel Hána, Ph.D. vedoucí katedry	prof. MUDr. Jozef Rosina, Ph.D., MBA děkan

Declaration

I declare that I have written the bachelor thesis entitled **Bioinformatics analysis of genomic data of patients with atypical Wilson disease** independently and used a complete list of citations of the sources used, which I provide in the attached reference list.

I have no valid reason against the use of this school work within the meaning of Section 60 of the Act No. 121/2000 Coll., on Copyright, on Rights Related to Copyright and on Amendments to Certain Acts (Copyright Act), as amended.

Kladno, 9.8.2022 …..……..……...

Petr Štěpánek

Acknowledgements

I would like to thank Mrs. Mgr. Veronika Vymětalová, Ph.D. for supervising my bachelor thesis and for her trust in my work. I would also like to thank prof. Ing. Stanislav Kmoch, CSc. for providing data for the bioinformatics analysis.

Abstract

Bioinformatics analysis of genomic data of patients with atypical Wilson disease

Wilson's disease is an autosomal recessive disorder that results from mutations in the ATP7B gene and is characterized by the accumulation of copper in the body of affected homozygotes. The manifestations of the disease are most often hepatic and neurological. The aim of this study was to analyze whole-genome sequencing data from three families that developed Wilson's disease with an atypical form of inheritance. With no known family history, probands developed symptoms in the form of copper accumulation in the liver and subsequently neurological problems. The paper includes a description of the bioinformatics procedure for processing next-generation sequencing data into a form for clinical processing and evaluation. The Wilson's disease is discussed in the research chapter. In conclusion, no direct causal variants in the ATP7B genes of patients and their family members included in this study were detected. However, a group of genes related to other diseases occurring in the families was also analyzed. There may be epigenetic factors in effect, that could explain the atypical cause of conditions of these families.

Keywords

Wilson's disease, ATP7B gene, whole genome sequencing, bioinformatics, genomics, family history analysis

Abstrakt

Bioinformatická analýza genomických dat pacientů s atypickou Wilsonovou chorobou

Wilsonova nemoc je autosomálně recesivní onemocnění, které vzniká v důsledku mutací v genu ATP7B a projevuje se hromaděním mědi v těle postižených homozygotů. Projevy onemocnění jsou nejčastěji hepatické a neurologické. Cílem práce bylo analyzovat data z celogenomového sekvenování tří rodin, u kterých se objevila Wilsonova nemoc s atypickou formou dědičnosti. Bez známé rodinné anamnézy se u probandů rozvinuly symptomy v podobě hromadění mědi v játrech a následně také neurologické potíže. Součástí práce je popis bioinformatického postupu zpracování dat z next-generation sequencingu do podoby pro klinické zpracování a vyhodnocení. Ve zpracované rešerši je rozebrána Wilsonova nemoc. V případě sledovaných rodin se nepodařilo odhalit přímé kauzální varianty v genech ATP7B pacientů a jejich rodinných příslušníků. Nicméně byla analyzována také skupina genů souvisejících s dalšími onemocněními vyskytujícími se v rodinách. Mohou zde působit epigenetické faktory, které by mohly vysvětlit atypickou příčinu onemocnění těchto rodin.

Klíčová slova

Wilsonova nemoc, gen ATP7B, celogenomové sekvenování, bioinformatika, genomika, analýza rodinné anamnézy

Table of Contents

List of Abbreviations

Abbreviation	Meaning
bp	Base pair
MBD	Metal-binding domain
WD	Wilson's disease
MRI	Magnetic Resonance Imaging
CBC	Complete Blood Count
NGS	Next Generation Sequencing
NMD	Nonsense mediated decay
CTBP2	C-terminal binding protein 2

1. Introduction of Wilson's Disease

Wilson's disease is an autosomal-recessive disease caused by pathogenic variants in the ATP7B gene. Early intervention and treatment are necessary to prevent frequent neurological, hepatological, and multisystem complications. The clinical manifestations and the age at which the disease develops can be wide-ranging. To date, more than 600 variants associated with this disease have been identified. These are predominantly single nucleotide missense mutations but also include indels and rarely splice-site gene mutations. The primary manifestation of Wilson's disease is impairment of copper excretion. Mutations in the ATP7B gene result in a disorder of biliary excretion of copper from the organism. The clinical manifestation of the disease results from the accumulation of copper in various tissues. Usually, the first signs are hepatic, but the copper overload affects nerve tissue, including the brain, resulting in multiple neurological symptoms. (1)

Diagnosis and early treatment of Wilson's disease are crucial and can help the affected reach a normal lifespan without many limitations. If left untreated, Wilson's disease most often results in the patient's death. There are a few unique cases, such as that of a woman who was first diagnosed at the age of 54 during a routine screening. She was asymptomatic at the time, only with Kayser-Fleischer Rings. She refused treatment and remained mostly symptom-free for another 20 years when she started to demonstrate slight evidence of liver dysfunction. For more than 30 years after diagnosis, at the age of 84, she remains symptom-free. (2).

This case represents a few outliers showing that Wilson's disease diagnosis can be complicated and that there are possibly many more unknown factors in play that lead to the development of symptoms. There is a significant discrepancy between the number of diagnosed cases (approx. 1 in 30 000) and the number of genetic cases (approx. 1 in 7026). (3)

Managing the disease, which mainly consists of lifelong treatment with chelating agents and a low-copper diet, can be complicated. Another outlier case described by Sohtaoglu et al. (2007) was of a woman who developed tremors, slurred speech, and postural instability at the age of 75. Mild symptoms had begun in her case 8 years earlier. The treatment with penicillamine, a standard-issued drug for Wilson's disease patient, had significantly deteriorated her overall health. This case has led the authors to question the safety of penicillamine in older adults. (2)

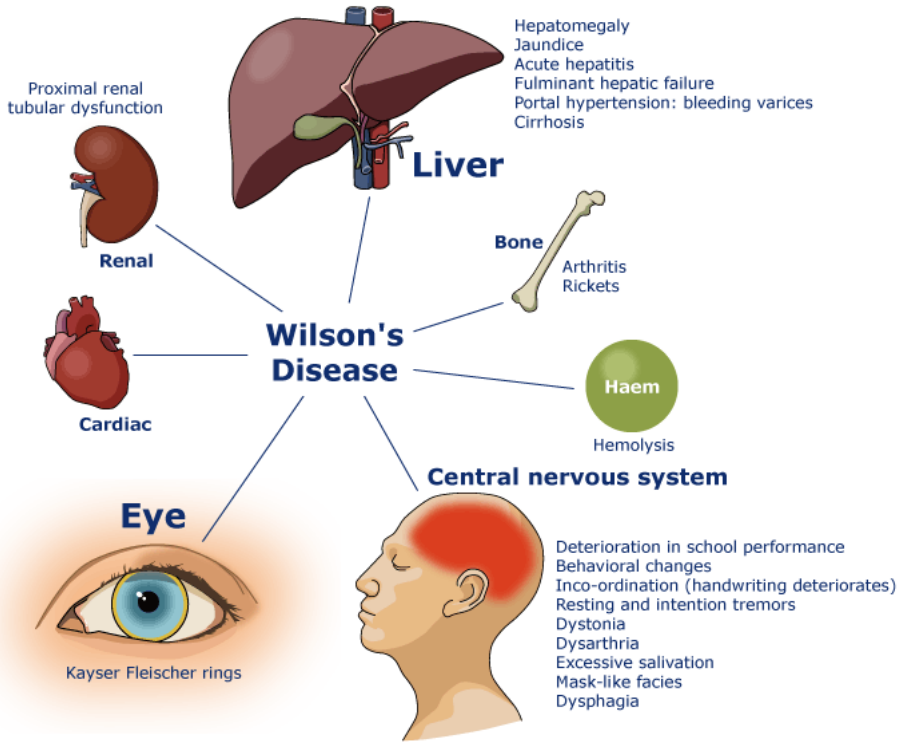

Figure 1 - Wilson's disease and its range of clinical manifestations. As a disorder of copper metabolism, it can affect multiple systems across the whole body. (4)

1.1 Prevalence and epidemiology

The incidence of Wilson's disease is thought to be around 1 in 30 000 births worldwide. There is a high degree of genetic heterogeneity among patients with Wilson disease. It has been historically complicated to evaluate the actual prevalence of Wilson's disease mutations in the general population. Even the rate of affected individuals is hard to specify because of the codominant nature of many variants that must be considered for the diagnosis. The mutation rate varies widely across populations, ranging from 0.3% in Japan to 2.5% in Italy. The highest rates occur in Ashkenazi Jews and individuals of Eastern European descent. (2)

First prevalence estimates were minuscule compared to our modern understanding of the disease, which is based chiefly on population-wide genomic studies. Sternlieb and Scheinberg have estimated, in 1968, the prevalence to be 5 per million individuals (1 in 200 000). Later, Saito reported a higher prevalence of 33 per million (1 in 30 303) in 1981. These studies were based on a limited understanding of the disease. Many biochemical studies, mainly on newborns, were trying to screen for Wilson's disease. Unfortunately, these didn't yield any results other than that the changes associated with Wilson's disease appear with increasing age. A screening done by Kroll et al., 2006 of 1045 newborns utilized a technique to measure holoceruloplasmin from dry blood spot tests that used an enzyme-linked immunosorbent assay with specific monoclonal antibodies. However, it has failed to reveal even a single case of the disease. Previously, a screening done by Yamaguchi et al. 1999, of a total of 126 810 newborns has led to the same results. Much better success was with a group of 24 165 children aged six months to nine years old. Out of that group, only three patients were diagnosed with Wilson's disease, the youngest being eight months old. This equates to a much higher prevalence of 124 per million (1 in 8065). (2)

Modern genomic studies have opened a new way to analyze the prevalence of Wilson's disease and its associated mutations in the population. There is a very significant difference between the number of individuals that present with symptoms and the individuals that are predicted to be affected by the disease based on genetic studies. Most affected individuals have different mutations in the two ATP7B alleles. Genetic studies have led to a prevalence estimate of 142 per million (1 in 7042). Many isolated

populations, namely in Romania and Sardinia, have been found to have a much greater prevalence, 885 per million and 370 per million, respectively. Furthermore, complete exon sequencing of the ATP7B gene in 1000 DNA and exons 8, 14, and 18 in 5000 samples in the United Kingdom has led to an estimated frequency of heterozygotes carrying the mutations to be about 4%. (2)

1.1.1 Prevalence in the Czech Republic

Wilson's disease has a prevalence in the Czech Republic that is similar to the global average. There are around 333 cases of Wilson's disease in the Czech population. The prevalence equates to approximately 1 in 30 000. Around 1 in 90 people are heterozygous carriers of some mutated ATP7B allele. The following table describes the most common mutations in patients with Wilson's disease in the Czech Republic. (5)

Table 1 - Frequency of Wilson's Disease recessive mutation in the Czech Population. (5)

Mutation	Frequency
H1069Q	57%
340delC	3.5%
W779X	2.75%
R778G	2.0%
1340del14	1.25%

1.2 Genetics of Wilson's Disease

Most commonly used genetic testing for Wilson's disease is impractical, given that multiple known mutations are causal to Wilson's disease with varying levels of severity. In the past, diagnosis has to rely mainly on a wide range of laboratory and other diagnostic tests. The associated mutations are mostly missense, but deletions, insertions, nonsense, and splice site mutations all occur. Affected individuals are usually compound heterozygotes. Such people have inherited different mutations from each parent that result in the manifestation of clinical symptoms. This combination of various mutations leads to enormous combinations and permutations that make developing a certain phenotype-genotype correlation hard. (1)

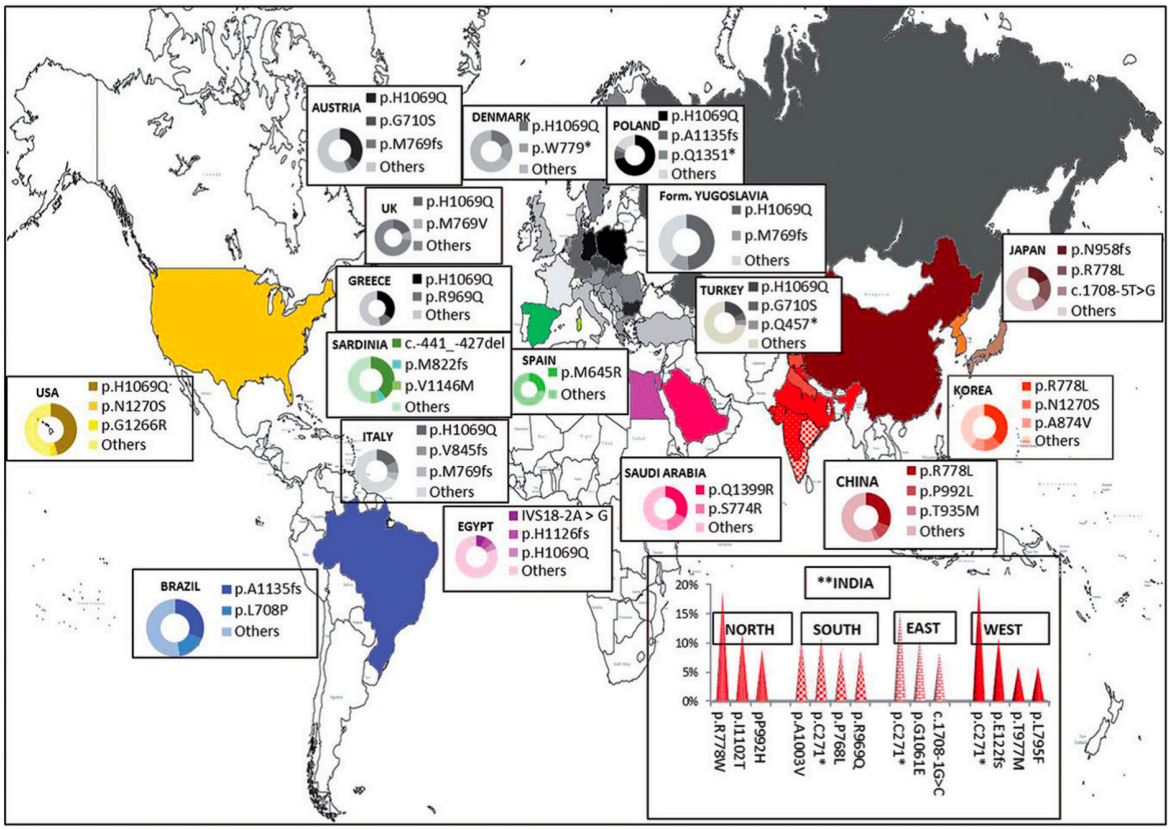

Figure 2 - Map showing a prevalence of ATP7B by geographic region. Higher allelic frequency is shown as a darker gradient on the map. (6)

One of the major breakthroughs was in the year 1993, when it was discovered that the gene ATP7B is a P-type ATPase, a 140 kDa copper-transporting enzyme that is located on chromosome 13q14. (1)

According to the Clinvar database, there are now 315 pathogenic variants that might cause Wilson's disease, and more with either uncertain significance or being likely pathogenic. (7)

Table 2 - Table of the most common mutations associated with Wilson's disease according to the Clinvar database (7)

Clinical significance	**Count**
Uncertain Significance	400
Likely Pathogenic	264
Pathogenic	315
Molecular significance	**Count**
Frameshift	139
Missense	432
Nonsense	82
Splice site	63
UTR	56
Variation Type	**Count**
Deletion	136
Duplication	47
Indel	9
Insertion	64
Single Nucleotide	659

Various mutations affect the function of ATP7B, depending on which coding region they affect the most.

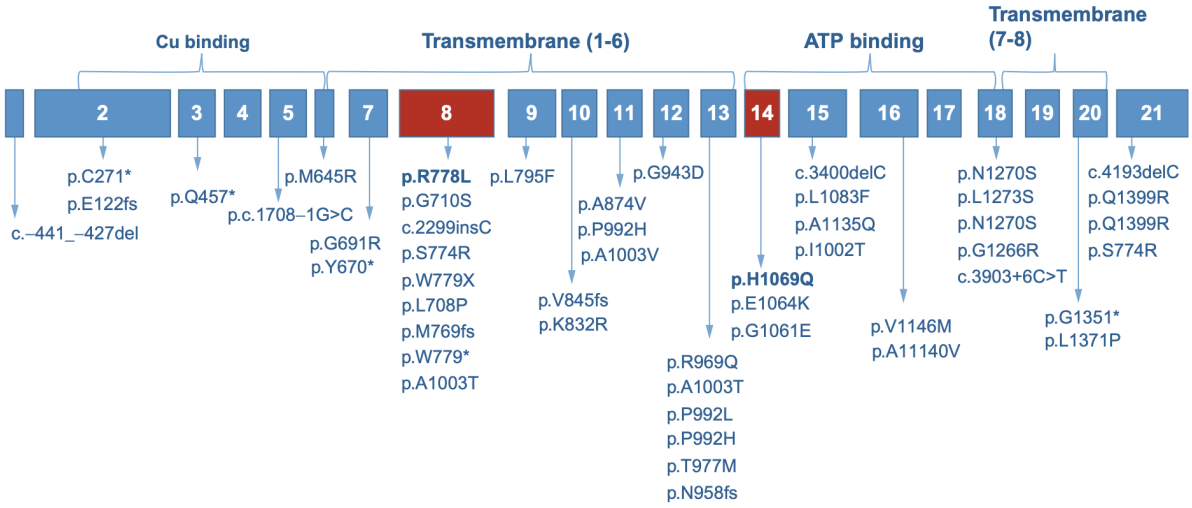

Figure 3 - Illustration of the most common mutations that result in coding changes with significant molecular consequences. The most common allele in the Czech population, H1096Q, affects exon number 14 in the ATP binding region of the protein. (2)

1.2.1 ATP7B Protein Structure and Function

There are two copper transport ATPases in the human body. ATP7A and ATP7B both play a crucial role in copper metabolism and homeostasis. Defects in the APT7A gene lead to Menkes disease, a rare disease that has many characteristic symptoms such as kinky hair, failure to grow, and nervous system deterioration. Menkes disease onsets during infancy, and the affected children (approximately 1 in 100 000 to 1 in 250 000) don't live past the age of three. (8)

Mutations in the gene that codes for the second ATPase, the ATP7B protein, are causal to Wilson's disease. ATP7B transporter has two main functions. It transports copper into the protein called ceruloplasmin found in plasma. The second is the elimination of copper in bile. (9) It uses the energy of ATP hydrolysis to transport copper ions across cellular membranes. It is localized mainly in the Golgi apparatus. (10) ATP7B is the only identified causal gene in Wilson's disease. However, there have been cases that lacked mutations and still developed symptoms characteristic of Wilson's disease, which raises the question of whether a second causal gene affects copper homeostasis. (2)

ATP7B is located on chromosome 13, most specifically on 13q14.3. It contains 20 introns and 21 exons. Its total genomic length is 80 kb. ATP7B is highly conserved and belongs to the P-type ATPase family of proteins responsible for transporting metals across cellular membranes. The gene is primarily expressed in the liver but is also present in kidneys, mammary glands, placenta, lungs, and brain. (11)

The ATP7B protein contains 1465 amino acids and many different domains. (11)

- Phosphatase domain (A-domain)
- Phosphorylation domain (P-domain, amino acid 971-1035)
- Nucleotide binding domain (N-domain, amino acids 1240-1291)
- M domain that consists of eight transmembrane ion channels

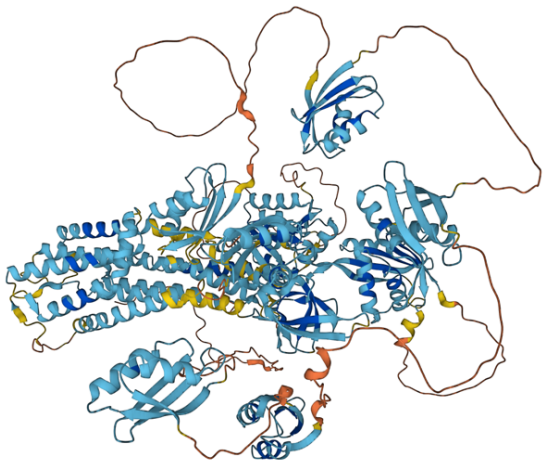

Figure 4 - A prediction of the structure of the ATP7B protein as created by AlphaFold using an AI system developed by DeepMind. It can regularly reach accuracy comparable with an experimental approach. (12)

The complex process of copper transport across cell membranes begins with ATP7B binding the copper at the N-domain. A metal binding domain (MBD) at the N-terminal is composed of six sites that can bind copper, each with its conserved sequence. The metal binding domains have a crucial role in accepting copper from the copper chaperone ATOX1. Afterward, it is transported across the cellular membrane using ATP as a source of energy. Free copper binds intracellularly to GG motifs in the metal binding domain, and finally, dephosphorylation of acyl-phosphate at the A-domain discharges the copper

across the cell membrane. Any of these steps can be impaired by a mutation that can cause copper accumulation. (13)

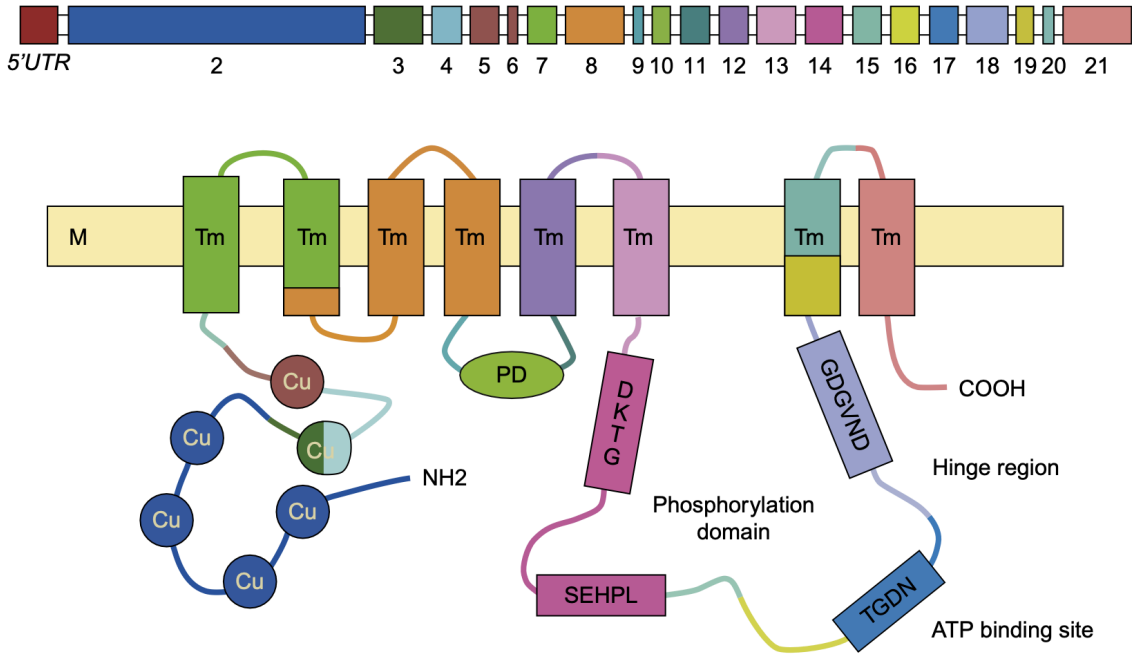

Figure 5 - A schematic representation of the ATP7B gene and protein. Introns and exons are shown on the top diagram, including the 5'UTR region. Domains of the protein are shown on the bottom diagram. (14)

1.3 Clinical Features

The range of manifestations of Wilson's disease is broad, and when left untreated, it results in multisystem complications that can affect various organs and tissues all over the body. The most significant symptoms can be divided into five general categories: hepatic, neurological, psychiatric, ophthalmological, and musculoskeletal. (1)

There is a slight bias caused by specialists when describing the first symptoms that appear in WD cases. The patients that first present themselves with health issues to neurologists have a higher prevalence of neurologic symptoms (69.1%) compared to hepatic

symptoms (14.9%). The hepatic symptoms dominate in patients that present to gastroenterologists first (68.1%). (15)

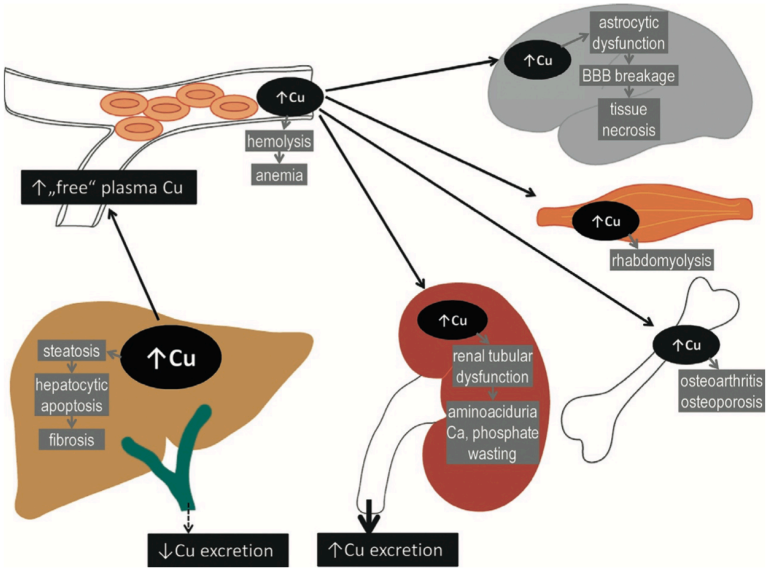

Figure 6 - Illustration of copper toxicity and mechanism in the human body in Wilson's Disease. Impairment of copper excretion in the liver leads to copper overload, affecting multiple systems across the body. The copper overload leads to liver toxicity and subsequent steatosis, hepatocytic apoptosis, and fibrosis. When the liver cannot accumulate any more copper, it is released into the bloodstream, where it can affect red blood cells, causing anemia and hemolysis. From the bloodstream, copper spreads to other organs, damaging them and eventually causing the typical symptoms of advanced Wilson's Disease.

1.3.1 Hepatic Manifestations

In approximately 40 to 50% of individuals with Wilson’s disease, hepatic manifestation in the form of liver dysfunction is the first to appear from the range of possible symptoms. Most cases present with first hepatic symptoms at an average age of 11.4 years. The symptoms can arise at any time but only very rarely appear after the age of 40. In rare cases, Wilson’s disease has been diagnosed in a patient as young as two years. (1)

The liver plays a crucial role in copper homeostasis. A higher concentration of copper in the liver is seen in most patients with liver disease. Still, even WD patients only with neurological symptoms have higher concentrations of copper in their livers. More than half of them also presents with liver cirrhosis. (16)

1.3.2 Neurological Manifestations

Neurological symptoms are initial for around 40-60% of patients that are eventually diagnosed with Wilson's disease, and the manifestations first appear at an average age of 18.9 years. Most frequent symptoms include resting, kinetic or postural tremors. Proximal upper limb tremor in the form of "wing-beating" can also appear. Spotting tremors in Wilson's disease patients can sometimes be challenging, affecting only distal extremities and being minuscule in amplitude. Other manifestations include dysarthria, dystonia, and gait abnormalities of extrapyramidal and cerebellar patterns. Unusual symptoms include chorea, tics, and myoclonus. (1)

1.3.3 Psychiatric Manifestations

The frequency of psychiatric manifestations in people diagnosed with Wilson's disease is mostly unknown. It can range from 20% to 65%, according to multiple studies. (1)

The most frequent psychiatric symptoms are personality changes and mood changes. The mood changes present especially as depression, sometimes aggression, and increased irritability. Depression can range from mild to severe. Almost 16% of patients had a history of suicide attempts. Some less reported symptoms are psychosis, anti-social or criminal behavior, sex preoccupation, and disinhibition. (2)

1.3.4 Musculoskeletal Manifestations

Wilson's disease also affects bones and joints. These effects are primarily under-recognized. Based on radiographic evidence, osteoporosis is present in up to 88% of affected patients. Osteoporosis can result in spontaneous fractures and joint issues, especially at the knees. The vertebral columns are abnormal in approximately 20 to 33% of individuals with Wilson's disease. (1)

1.3.5 Ophthalmological Manifestation

Kayser-Fleischer rings are dark rings that appear in the iris of the eye as a sign of Wilson's disease. They directly result from copper deposition caused by various liver diseases but are most often connected with Wilson's disease. Due to high variance in the clinical presentation of this disease, Kayser-Fleischer rings cannot be the sole manifestation that would be sufficient for diagnosis. Still, if they appear, the chance that the person is affected by Wilson's disease is high. They usually appear clearly visible only in light-eyed people and are most difficult to find in brown eyes. If anticipated in the patient, the rings can be detected with a slit lamp examination much earlier than they would appear to the naked eye. (1)

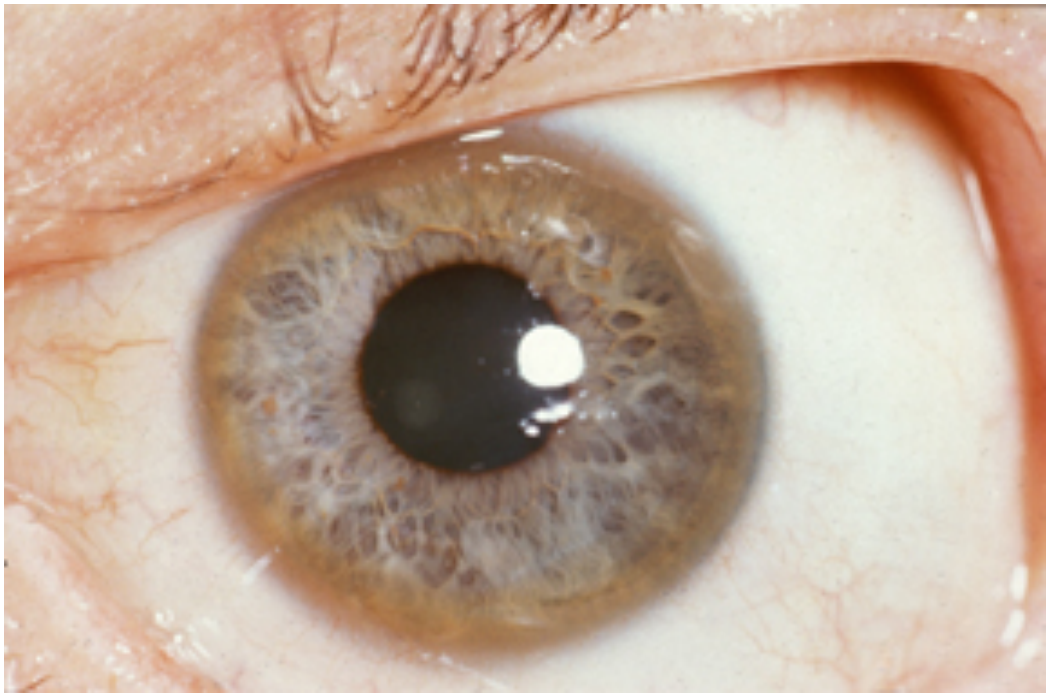

Figure 2 - Photo of the Kayser-Fleischer ring. A characteristic symptom of Wilson's disease. The photo shows a 32-year-old patient with speech difficulties and tremors in his limbs. (17)

A modern machine-learning-based approach to the detection of Kayser-Fleischer rings has been proposed and developed by many teams worldwide. Researchers and developers face the challenge of the lack of a high-quality training set of photos of eyes with Kayser-Fleischer rings.

One notable project is called the Killer Eye. It analyzes each frame of webcam input using TensorFlow, an open-source library for machine learning and artificial intelligence. Therefore, it uses a neural network to recognize Kayser-Fleischer rings. The recognition algorithm was trained on a set of 35 images of Kayser-Fleischer rings from blue-eyed patients. The tool is available free to use. Unfortunately, it doesn't have enough awareness; therefore, it isn't widely used and even developed, with the last update on it being from 2020. (18)

Manufacturers could integrate similar lightweight software into a wide range of ophthalmic equipment, readily taking digital photographs of the patient's eyes during routine eye exams. Building this recognition capabilities into commonly used equipment could be one of the approaches that could lead to earlier diagnoses of Wilson's disease in mostly asymptomatic patients or those that first present with neurological or psychiatric symptoms.

1.4 Diagnosis

The diagnostic methods used depend on whether there are symptoms present, which would indicate that copper has already spread past the patient's liver. Given the high variance in symptoms presentation, diagnosis is complicated, especially in unaffected patients, due to mostly unknown reasons for this large variance. In patients with family history, a routine liver examination repeated every two years is usually recommended, along with genetic tests for Wilson's disease-causing mutations. (2) (1)

1.4.1 24-Hour copper measurement

Evaluating the 24-Hour copper measurement test is especially useful in patients with neurological or psychiatric manifestations. The level of copper is measured in the urine of patients over 24-hour period. It is important to collect it into copper-free containers because outside factors and influences could easily contaminate the amounts that are measured, given that they are minuscule. Urine copper levels in affected, symptomatic patients usually exceed 100 µg/dL. However, urine copper levels can be elevated for different reasons unconnected to Wilson’s disease. Heterozygous carriers of Wilson’s disease mutations also have slightly higher urine copper levels but usually don’t rise above 100 µg/dL. (1)

1.4.2 Serum Ceruloplasmin

The measurement of serum ceruloplasmin may fall slightly below normal range in approximately 5 to 15% of patients affected by Wilson's disease. About 10 to 20% of unaffected carriers have lowered levels. Its measurement is simple but not practical as a standalone screening test, given its low specificity. There are many conditions that can result in reduced ceruloplasmin levels, such as nephritic syndrome, sprue, or chronic liver disease of any cause. It may even be intermittently elevated into normal ranges, even in patients affected by Wilson's disease, because of birth control pills, infections, or steroid ingestion. (1), (2)

1.4.3 Serum Copper Measurement

Most of the copper that is present in the bloodstream is bound to ceruloplasmin. Therefore, measuring free non-ceruloplasmin copper in serum can reveal the amount of possibly toxic copper in the patient's bloodstream. The complication is that laboratories usually cannot directly measure free copper in the blood, so it must be derived by calculating it. The amount can be approximated by multiplying the amount of serum ceruloplasmin by three (in mg/dL) and then subtracting that from the level of total serum copper (in μg/dL). The normal range for free non-ceruloplasmin copper should be 10 to 15 μg/dL for healthy patients. (1)

1.4.4 Hepatic Copper Content Determination

Determining the copper content of a liver biopsy is one of the most accurate methods for diagnosing Wilson's disease. Given its invasive nature, it is usually reserved for patients that present with hepatic manifestations and none other. Patients who already have neurological or psychiatric symptoms can generally be diagnosed by serum and urine tests and by looking for Kayser-Fleischer rings. The amount of copper in the dry mass of a liver sample is significantly increased in Wilson's disease-affected patients, with levels greater than 250 μg/g seen in approximately 83.3% of cases. The normal range is usually specified as 15 to 55 μg/g of copper per dry liver tissue. However, increased copper content in the liver can be caused by a multitude of other liver diseases such as primary biliary cirrhosis, biliary atresia, biliary obstruction, hepatitis, or some autoimmune diseases. Also, in around 3.5% of affected patients, the levels were below 50 μg/g, which

signals that there can be exceptions; the combination of multiple diagnostic methods must cover that to reach a conclusion. (1)

1.4.5 Kayser-Fleischer Rings Examination

Using a slit-lamp, an ophthalmologist can look for early signs of developing Kayser-Fleischer rings, which are deposits of copper that form brown-colored rings that encircle the cornea of the eye. These early signs can be seen in a slit lamp microscope before appearing to the naked eye. Even then, it isn't very easy to use this diagnostic method, given that in specific geographic regions, dark-colored eyes are prevalent, which makes Kayser-Fleischer rings-based diagnosis unviable. They are also often absent in affected patients without neurological or psychiatric symptoms. Those who present only with hepatic symptoms usually don't develop Kayser-Fleischer rings. (1)

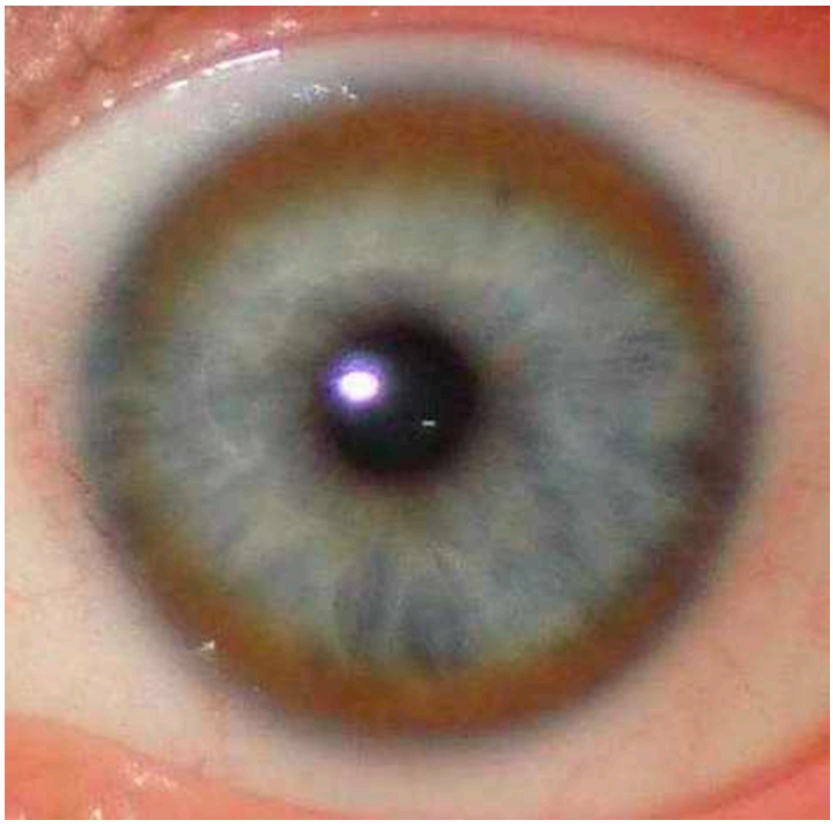

Figure 7 - A Kayser-Fleischer ring in light-eyed patient with Wilson's disease (2)

1.4.6 Other Diagnostic Methos

Neuroimaging studies have shown to be highly informative in patients who present with neurological or psychiatric symptoms. According to one study, some visible changes were present on MRI scans of all tested patients affected by Wilson's disease. A multitude of abnormalities was described based on those studies. Other tests that have also been

described are focused on the use of radioactive copper for more accurate localization of its movement in the body and also the measurement of copper in the cerebrospinal fluid. (1) (2)

Table 3 - Summary of Diagnostic Criteria of Wilson's disease. References are stated in each criterion.

Criteria	Normal Range/State	Wilson Disease Diagnosis
Serum ceruloplasmin is decreased in affected individuals	20-35 mg/dL	< 20-35 mg/dL
24-Hour copper measurement above normal range	20-50 µg/g	> 100 µg
Hepatic copper content in dry tissue	15-55 µg/g	> 250 µg/g
Complete Blood Count (CBC)	Affected by multiple conditions	May show signs of anemia
Non-ceruloplasmin bound free copper	10-15 µg/dL	>15 µg/dL
Hepatic blood tests	Affected by multiple conditions	Mild elevation of liver enzymes
Hepatic manifestations	Affected by multiple conditions	Steatosis, cirrhosis, possibly acute liver failure with hemolysis and hepatic necrosis
Kayser-Fleischer rings	Can be caused by corneal staining unrelated to WD	Appear in almost all cases that present with psychiatric or neurological symptoms
Neurological changes	Affected by multiple conditions	slurred speech, abnormal gait, balance disturbances, drooling, involuntary movements, dystonia, Parkinsonism
Psychiatric changes	Affected by multiple conditions	Signs of depression, irritability, aggression, psychosis, suicidal thoughts
Neuroimaging (MRI)	Affected by multiple conditions	T2-MRI, atrophy of caudate and putamen, increased degenerative changes in the brain
Radioactive Copper Ratio – intravenous infusion of ^{64}Cu, measured within the liver and serum after 2, 24, and 48 hours	24 hours/2 hours ratio > 0.3 48 hours/2 hours ratio > 0.395	24 hours/2 hours ratio < 0.3 48 hours/2 hours ratio < 0.395
Genetic Analysis	No pathogenic variants in the ATP7B gene, or heterozygous	Disease-causing mutations are present in ATP7B

1.4.7 Future of Wilson's disease diagnostics

It has long been impractical to do a genome-wide screening of individuals for the genotypes that are associated with Wilson's disease. Pfeiffer, in 2007, stated that the multitude of documented mutations identified in Wilson's disease makes commercial genetic testing impractical and that advances in technology may make it possible in the future, but that the diagnosis of Wilson's disease still must be made based on a combination of diagnostic tests. (1).

Nevertheless, the technology has advanced significantly over the past years, and as of the year 2022, the total cost of sequencing a human genome has dropped below $1000. In the year 2006, completing a human genome sequence would cost around $20-25 million. Creating a high-quality 'draft' whole human genome sequence in mid-2015 was about $4,000, but by late 2015 the price fell below $1,500. The innovations that led to this dramatic price decrease are mostly credited to the Illumina company, which has developed their own proprietary sequencing by synthesis technology that they utilize in their line of high-throughput sequencers. (19)

Nowadays, more companies are entering the market, such as BGI, with their daughter company MGI. They use technologies similar to Illumina, and with their equipment, a genome can be sequenced with thorough coverage for around $700. (20)

As stated previously, there is a discrepancy between the known, diagnosed prevalence of Wilson's disease, which is 1 in 30 000 people, and the genetic prevalence, which is measured to be as high as 1 in 7026 individuals according to whole-genome sequencing population studies. (21)

This significant difference may be caused by yet unknown genes that affect or modify copper metabolism or possibly by the lack of good, accessible diagnostic standards that could be utilized routinely. (22) Even though genome-wide genetic testing is accessible and affordable today, it is unpractical in the case of Wilson's disease because many studies have shown no clear connection between genotype and phenotype. There have been multiple cases of people who should have been affected but never presented with any symptoms. (23)

Diagnostic methods utilized for Wilson's disease are heading toward optimization. The growing opportunities offered by population wide NGS studies have led to the creation of large datasets which can be utilized to study the genetics behind the disease. Sometimes, clinicians and researchers view Wilson's disease as either too rare or too benign to be worth studying, but many challenges remain to be solved. (22)

1.5 Management

The outcome can be very promising when Wilson's disease is treated early, although long-term survival is not yet well documented. (24)

Most available treatment is only palliative, the only notable exception being a liver transplant. Other forms of treatment cannot treat the underlying cause of copper disbalance. Lifelong pharmacological treatment is necessary because limiting the intake of copper from food is often ineffective. Still, a low-copper diet is usually advised to patients. Furthermore, people need to be aware that copper utensils and cookware can also release a small but not negligible amount of copper into food. A lifelong course of disease management with chelating agents is eventually inevitable for most affected patients. In case of liver failure, a liver transplant is necessary, curing the underlying problem of copper accumulation, but it comes at the cost of life-long immunosuppression. Routine monitoring of ceruloplasmin, serum copper, and liver enzymes should also be considered in affected patients and their relatives. (25)

1.5.1 Zinc

First proposed in 1961, zinc can serve as a management drug because, when ingested, it reduces copper absorption in the intestine. (1) It can be administered in the form of acetate, sulfate, or gluconate. Zinc causes the formation of metallothionein in intestinal enterocytes which then can bind both zinc and copper, which leads to their excretion in feces. Zinc is generally well tolerated; the usual dosage consists of 50 mg of elemental zinc three times a day. (26)

1.5.2 Penicillamine

A widely used drug that soon became the standard of therapy for Wilson's disease after its introduction in the year 1956 (1). D-Penicillamine is a metabolic product of penicillin, and it has the ability to chelate copper with a relatively quick onset of improvement after the beginning of treatment. Most patients improve in 2 weeks, although it may take longer. During treatment copper from various tissues is avidly excreted in urine. Penicillamine is usually taken in four doses, 250-500 mg each, on an empty stomach. There has been a number of documented side effects of taking penicillamine. In nearly a third of all patients an acute case of skin rash, fever, eosinophilia, thrombocytopenia, leukopenia, and lymphadenopathy can develop, which leads to abandonment of the treatment. (1) Also, there have been concerns about neurological decline that might be connected to the use of penicillamine in patients. In 50% of patients, these neurological symptoms don't improve after cancellation of the treatment. Long-term administration of penicillamine often leads to a wide range of complications, which can include anemia, bone marrow suppression, lupus-like syndrome and other. (27)

1.5.3 Trientine

Trientine is a drug with a similar mechanism to penicillamine. It was approved for use in the United States in the year 1969. It is an orally available copper chelating agent. Trientine that is complexed to copper is excreted in the urine. The medication is taken on an empty stomach in a dosage ranging from 750 to 1250 mg for adults and 500 to 750 mg for children, initially given in 2 to 4 divided doses daily. The dosage can be raised to a maximum of 2000 mg daily for adults and 1500 mg for children (28). Common side effects can include headache, arthralgias, myalgias, nausea, anorexia, diarrhea, rash, and renal dysfunction. Other works stated side effect such as: proteinuria, bone marrow suppression, autoimmune reactions, worsening of neurologic symptoms. (27) Side effects are mostly mild but aren't as well studied as those of penicillamine. The frequency of side effects of penicillamine has led to greater attention being given to trientine as the leading drug for treating Wilson's disease. (1). However, trientine is less effective in attaining negative copper balance compared to D-penicillamine.

1.5.4 Tetrathiomolybdate

Also called Tiomolibdic acid is a medication used for the treatment of Wilson's disease in the form of the salt bis-choline tetrathiomolybdate. (29) It was first tested in the year 1984 as a potential treatment for Wilson's disease. It is still subject to clinical studies and is therefore used only as an experimental medication. (1)

1.5.5 Liver Transplantation

A liver transplant is the only known definite cure for Wilson's disease, but it comes with a life-long treatment with immunosuppressives. Patients who develop fulminant liver failure have a very high mortality rate despite medical treatment. A transplant is the only option in those cases. (30)

1.5.6 Proposed and Future Treatments

The current treatment consists mostly of symptom management. To maintain copper homeostasis, either chelating agents, which directly chelate copper from the blood, are used, or zinc salts which decrease copper absorption in the intestine. It is crucial to achieve a correct balance, because during long-term disease management, even copper deficiency can occur.

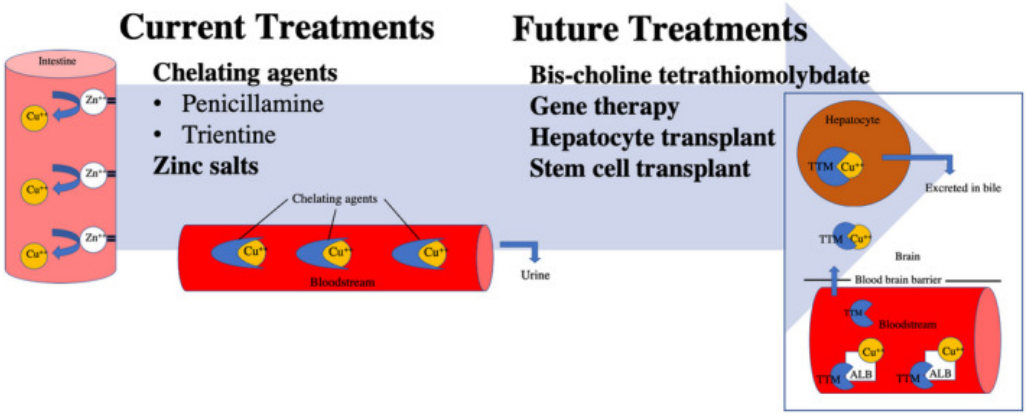

Figure 8 - The current treatment consists mostly of the management of copper levels. Future treatments are focused more on the root of the cause. Promising gene therapies based on viral 8 vector expressing ATP7B transgene have been shown to be helpful in the more long-term treatment of copper metabolism issues. However, it yet has to be tested in humans. (Leung, 2021) (22)

Future possibilities include gene therapies, which could use viral vectors to express ATP7B transgene in the liver of affected patients. A company called Ultragenyx from the United States has announced, that the first patient has been dosed in their CYPRUS2+ clinical study, which aims to find a treatment of Wilson's disease. It consists of a one-time intravenous infusion of UX701, an investigational adeno-associated virus gene therapy which has the potential to restore ATP7B function and improve copper homeostasis. (31)

Also, transplant of healthy hepatocytes has led to decreased inflammation and prevention of fulminant liver failure. Bis-choline tetrathiomolybdate is still in medical trials, but it could offer a more easily tolerated alternative to penicillamine that yields better results in treatment of neurological symptoms. (32)

2. Genomic Data Analysis

The genome represents the collection of all genetic information of an organism. Genomics is an interdisciplinary field that focuses on the study of the structure, function, and evolution of the genome. It is also concerned with mapping the occurrence of variation in genomes and also with editing them. The basis of genomics is the analysis of DNA, which is a macromolecule composed of two complementary polynucleotide chains. (33) Bioinformatics is a discipline combining biology and computer science. It processes medical data and, most often, sequencing data in the field of genomics. It uses many analytical tools to process the data into a form that a physician can interpret. This is most often done using statistics, but increasingly also with the use of machine learning methods. (34)

2.1 Genomic Data Formats

Genomic data is vast compared to most data used in medical practice, maybe with the only exception being medical imaging data. A human genome is approximately 6.4 gigabases in length. It needs to be read multiple times to avoid errors. Therefore, special data formats had to be developed and defined to help facilitate data storage, exchange, and processing.

2.1.1 FASTQ Files – Genomic Raw Reads Data

FASTQ is a text-based format that stores raw sequence reads from NGS sequencers. FASTQ files use four lines for one short read section

The 1st line starts with the '@' character followed by the identifier of the section being read
The 2nd line contains the short nucleotide sequence (approximately 100 to 200 bases)
The 3rd line starts with a '+' character and may optionally contain an identifier or additional description
The 4th line contains characters that symbolize the read quality of the bases from the second line.

For example, a sequence record from a FASTQ file might look like this. (35)

@SEQ_ID
TCGCGCGCATAGACCTTTATCTCCGGTTCAAGTTAGGCATGAGGCTGCATGC
TACGTTG
+
!"*((((***+))%%%++)(%%%%).1***-+*"))**55CCF>>>>>>CCCCCCC65

The symbol representing the quality is in hexadecimal form from 0x21 to 0x7e, converted using the ASCII table into characters starting from the lowest quality '!' to the highest quality '~' as listed below. (36)

!"#$%&'()*+,-
./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmn
opqrstuvwxyz{|}~

FASTQ files must go through a process called sequence alignment before they are anything more than a set of a few randomly placed short stretches of DNA. Sequence alignment is the process by which short sections are compared to the reference sequence of the organism from which the sample most likely came. Often this situation can be complicated by high contamination, e.g. from a buccal swab of a person after a meal, where the sample may then contain DNA from animals or plants in the diet. The mere presence of the oral microbiome often accounts for up to 10% of all short reads when using saliva as a source of genetic information (34). In addition to matching the reference

sequence, it is also important that there is sufficient overlap between the segments so that a complete sequence can be constructed based on their alignment. Duplicate segments are usually aligned to a single genetic locus, corresponding to the reference sequence and the relationship to the surrounding segments. The alignment process results in a SAM file in text form, a BAM file in binary form, or a CRAM file in binary compressed form. Together, the number of segments that were successfully aligned to a sequence at a particular locus determines read coverage. This means the number of times a given section of the genome has been uniquely read. Duplicate sections are not used for diagnostic purposes as they could bias the resulting measurements of, for example, the zygosity of a given variant (34). Freely available software such as bwa (Burrows Wheeler Aligner) can be used for alignment. (37)

2.1.2 Whole-Genome Sequencing data in SAM/BAM/CRAM format

The most commonly used format for whole-genome sequencing data is called SAM/BAM/CRAM. All these formats serve the same purpose. They store the reconstructed sequence of the monitored sample aligned to the reference genome. The SAM file is created as a result of the alignment of the FASTQ files, which store the raw reads from next-generation sequencers. This data can then be further converted into a binary BAM form, which takes up significantly less disk space, e.g., using samtools (36). Alternatively, the file can also be further reduced using compression to CRAM format, which may or may not be lossy. In the case of an emphasis on size reduction, base quality scores are removed from the file, leaving only some which warn about very low quality of reads. (34) All BAM files are always accompanied by a BAI index file. The SAM and CRAM formats also have their indexes in the form of the SAI and CRAI files. These index files do not contain any sequencing data. They serve only for quickly searching the large files they accompany. (34)

2.1.3 VCF – Variant Call File

The VCF files contain a textual listing of changes from the reference genome. They are created in a process called variant calling. The primary use of this file is diagnostics and data analysis. It is a small file that can be easily transferred and processed. In a smaller form, e.g., when dealing with filtered results of whole-exome sequencing or genotyping using microarrays, VCF files can be conveniently browsed using Microsoft Excel or any viewer that can open tab-separated text. Larger files resulting from whole genome

sequencing mostly contain many intron variants, making them hard to analyze without proper filtering. (36)

Below is a snippet of a VCF version 4.2 file describing several example variants. Only some columns are shown in this excerpt. In addition to those listed, there are usually optional columns with descriptions and filter parameters used for variant calling. It may also include information about the variants in the form of their location (in coding or non-coding regions) and their score. (36)

Table 4 - Columns in VCF file version 4.2

CHR	POS	rs ID	REF	ALT	QUAL	FILTER	SAMPLE ID
chr1	4581682	rs241225	G	A	2040.77	PASS	0|1:48:1:51,51
chr1	4581739	rs241224	G	T	1960.77	PASS	0|1:49:3:58,50
chr1	4582097	rs241223	C	T	1796.77	PASS	0|1:54:7:56,60
chr1	4582417	rs184781	C	T	1876.77	PASS	1|0:44:3:58,50
chr1	4583042	rs147765774	C	CCT	678.73	PASS	1|1:49:3:58,40
chr1	4583412	rs725563	G	T	1826.77	PASS	1|0:42:3:52,40

The first column of the CHR indicates the name of the chromosome, taken from the reference genome, in which the variant occurs. The second contains a position denoted by a numeral describing the position of the nucleotide within the chromosome. The rs ID is used to accurately identify the variant and is used by a wide range of databases such as dbSNP (38). The following columns contain the reference and alternative alleles of the variant. The example file shows that these are primarily single-base substitutions. In the fifth row, we can see an example of an indel. The QUAL and FILTER columns indicate whether the variant has passed quality control. Since the usual procedure usually involves calling all variants and then marking whether they passed the filter or not, the VCF file contains both PASS and FAIL variants. Sometimes it is intentional to remove FAIL variants from the file to save space or for use in programs that cannot recognize this parameter. (34)

2.2 WGS Data Processing Pipeline

First, any analysis whole-genome sequencing data must begin by data quality control. Then the data that passed QC has to be aligned to a reference genome (in this case GRCh38) (34). For alignment, I used the Burrows-Wheeler Aligner (BWA) (37). For variant calling, I subsequently used the Genome Analysis Toolkit (GATK) (34). To maintain the quality of the generated SNPs, I followed the practices as recommended by the Broad Institute, GATK (34). The steps are described below in the pipeline chart for WGS data processing. The input of this pipeline are raw reads from NGS sequencers and output is a filtered and recalibrated file that includes all the selected genetic variants.

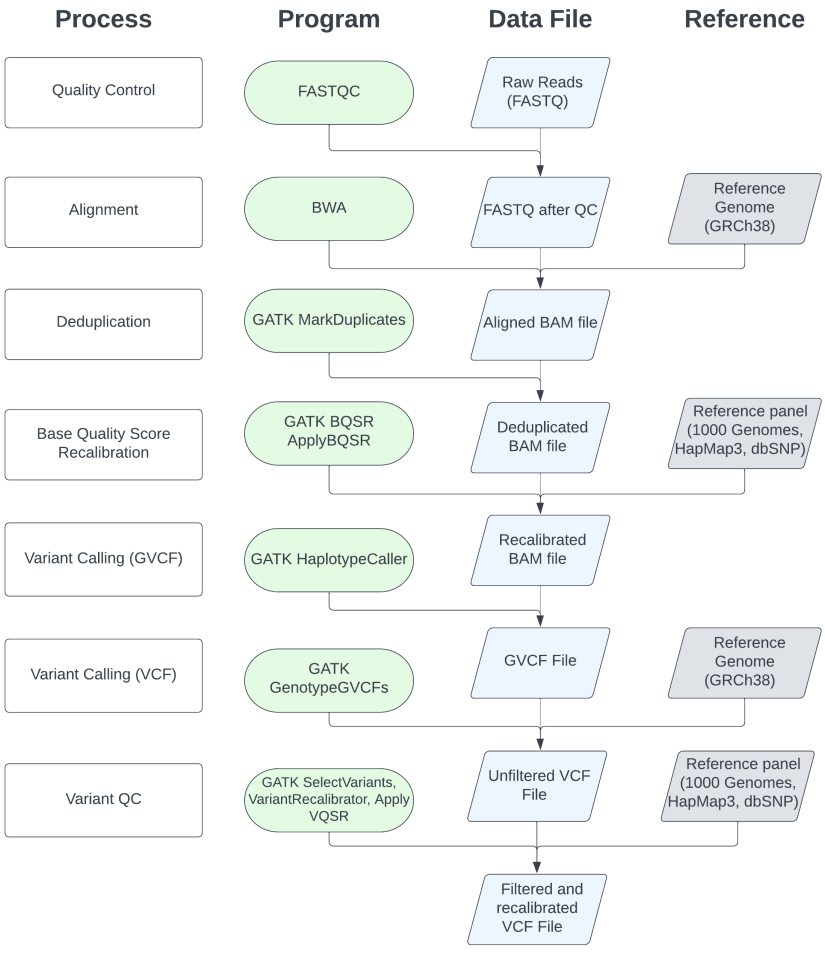

Figure 9 - WGS data processing pipeline. The first step is quality control, and afterward comes the sequence alignment. Refinements such as deduplication of reads and base quality score recalibration have to be made after alignment to ensure high-quality variant calls. After this comes the final step of variant calling and then quality control of the called variants. Some variants don't pass the filters that are set, depending on the use case, and some are selected intentionally. The result is a filtered and recalibrated VCF file that can be used for other diagnostic or research purposes.

2.2.1 Input data QC

A complete quality control of the sequencing results is required before each analysis, especially for diagnostic purposes. In most cases, this process is already handled when the raw sequencing data is converted to FASTQ format within the laboratory. However, to be on the safe side, this check also needs to be performed on our end, assuming we are not fully familiar with the reliability of the input sequences.

A possible option for quality control is the FASTQC program (39). It allows a thorough check of FASTQ files created using Illumina technology, but also MGI. It checks the lengths of the read segments and displays their distribution across the file. It also shows the number of duplicates and the distribution of quality scores.

The biological samples collected without care may be contaminated with foreign DNA, which will also negatively affect the accuracy of the alignment. If separated DNA from a blood sample is used, the risk of contamination is minimal. On the other hand, if buccal swabs or saliva is used as a source of genomic DNA, we have to take into account that the resulting genome coverage will be affected by the presence of microbial DNA or food remnants. Contamination with foreign animal or plant DNA can be avoided by following a regimen of at least half an hour without food before sampling. Microbial contamination can also have advantages, as every sample taken from a person's mouth also contains their oral microbiome, which can then be separated and analyzed from the data. However, if the origin of the sample is uncertain, it is worth checking. For example, the FastQ Screen program can be used for this purpose (40). Simply upload the reference sequences of the organisms suspected of contamination. The program will then reveal which sections our sample are shared with the organisms of interest.

2.2.2 Sequence Alignment and BQSR

After data quality analysis, the FASTQ files need to be further processed. Sequence alignment is the next step. It can be done using several tools. According to the recommendations of the GATK Best Practices manual [2], the current preferred method is to use Burrows-Wheeler Aligner (BWA) [14]. The alignment parameters strongly influence the following interpretation of data and variant calls.

```
bwa mem -t 24 -M -Y -v 3 hg38.fa G000539600_1.fq G000539600_2.fq > G000539600.sam
```

The finished SAM files need to be subsequently converted to BAM binary form. To do this, we use samtools with the view command with the -Sb parameter specifying the output format as BAM. The -@ parameter sets the number of threads available in the process. It will vary based on the hardware used. [19]

```
samtools view -@ 24 -Sb G000539600.sam > G000539600.bam
```

The resulting BAM file needs to be sorted according to the coordinates in the genome. To do this, we will also use samtools with the sort command. We then index the BAM file using samtools index, which allows us to search faster in a large file (approximately 150 GB for a 30x read human genome). [19]

```
samtools sort G000539600.bam -o G000539600_sorted.bam
samtools index G000539600_sorted.bam
```

The BAI index to the corresponding BAM file is created. In this state, the BAM file can be treated as further processable. However, it is highly likely that the data in it is still imperfectly cleaned; therefore, further processing is needed to get the most accurate results. (34) The Genome Analysis Toolkit (GATK) developed at the Broad Institute of MIT and Harvard can be used for this purpose. GATK is the current standard for identifying SNPs and indels in DNA and RNA sequencing data. The tool was originally developed for use with the human genome only. Now it can be used to process data from whole genomes or exomes of any organism. In addition to variant searching, GATK includes tools for deeper processing and quality control of NGS data. (34)

Duplicates must first be flagged in the data cleaning and quality control preparation procedure. Some segments may be read several times resulting in duplicate information that should not be considered in the variant calling process. GATK can be used to mark duplicates with the MarkDuplicates command. This command only marks the duplicates but does not remove them from the BAM file. Therefore, other tools will already treat them as duplicates and, therefore, will not affect the accuracy of the results. The output

is a file with marked duplicates and a text file with metrics about the duplicates in the file. (34) (41)

```
gatk MarkDuplicates -I G000539600_sorted.bam -O G000539600_sorted.dedup.bam -M
G000539600_metrics.txt
```

The variant calling process relies heavily on the base quality indicators included in the file. The problem is that Illumina sequencers tend to overestimate this quality. This could lead to false positive variants. There is a Base Quality Score Recalibration (BQSR) procedure in GATK that can address this problem by using a machine learning process that analyzes the context in the sequence to normalize the quality score. Thus, [2] we will use BaseRecalibrator to correct and normalize quality scores across our data. We also need to supply data on common variants in the population in addition to the reference genome (here we used hg38.fa). Below, data from HapMap, 1000 Genomes, dbSNP and directly from GATK were used (34).

The GATK BaseRecalibrator program requires Read Group Tags to be present in the processed file. If the input file does not contain these tags the program cannot process it. The tags can be added using Picard with the AddOrReplaceReadGroups command. (41)

```
java -jar picard.jar AddOrReplaceReadGroups \
I= MRM002_sorted.dedup.bam \
O= G000539600_sorted_rg.dedup.bam \
RGID=4 \
RGLB=lib1 \
RGPL=ILLUMINA \
RGPU=unit1 \
RGSM=20
```

```
gatk BaseRecalibrator -I MRM002_sorted.dedup.bam \
 -R hg38.fa \
--known-sites 1000G.phase3.integrated.sites_only.no_MATCHED_REV.hg38.vcf \
--known-sites 1000G_omni2.5.hg38.vcf.gz \
--known-sites Mills_and_1000G_gold_standard.indels.hg38.vcf.gz \
--known-sites hapmap_3.3.hg38.vcf.gz \
--known-sites Homo_sapiens_assembly38.known_indels.vcf.gz \
--known-sites GCF_000001405.39.gz \
-O recal_data.table
```

The parameters are:

-I - input BAM file (MRM002_sorted.dedup.bam)

-R - reference genome (hg38.fa)

--known-sites - files with known variants from large genome-wide studies. Based on these and the input data, GATK builds a covariance model.

-O - table with data needed for recalibration of the BAM file

The entire process of recalibrating the quality score can take up to 8 hours. The time required can be estimated if we know the total number of all short genomic segments in the BAM file, which can be determined using the **samtools view** with the **-c parameter**. (36) (34)

```
samtools view -@ 24 -c G000539600_sorted_rg.dedup.bam
```

The resulting value was 1 323 411 520 short reads in this example BAM file. At a processing speed of approximately 3 million short reads per minute, this would mean that the process would take roughly another 8 hours. The problem is that GATK can only use a limited number of processor cores, unlike samtools, in which this number can be specified using the -@ parameter. (34)

But there is a solution in the form of Spark for GATK. Spark is software that enables multithreading for selected supported GATK functions. The BaseRecalibrator function has a counterpart for use with Spark called BaseRecalibratorSpark. So I used the same command as above, but with the parameter that I wanted to use all 24 CPU cores of my

computer. The command itself doesn't change in any way other than changing the name and adding the spark-master local[*] parameter, which means that multithreading will be used on the local device on all available cores. (34)

```
gatk BaseRecalibratorSpark
-I G000539600_sorted_rg.dedup.bam
-R Homo_sapiens_assembly38.fasta
--known-sites 1000G_omni2.5.hg38.vcf.gz
--known-sites Mills_and_1000G_gold_standard.indels.hg38.vcf.gz
--known-sites hapmap_3.3.hg38.vcf.gz
--known-sites Homo_sapiens_assembly38.known_indels.vcf.gz
-O recal_data.table --spark-master local[*]
```

The resulting recalibration table is used in the ApplyBQSR command to generate the recalibrated BAM file. (34)

```
gatk ApplyBQSR -R hg38.fa \
-I G000539600_sorted_rg.dedup.bam \
--bqsr-recal-file recal.table \
-O MRM002_sorted.dedup.recal.bam
```

2.2.3 Variant Calling

After the last step, the genomic data is ready for variant calling. This can be performed using the GATK tools HaplotypeCaller and GenotypeGVCFs. First, the g.vcf.gz file needs to be generated. This is not the final product yet, just an intermediate step leading to the final VCF file. The difference between gVCF and VCF file is that the gVCF file contains all genotype information, including reference variants. The process of generating the g.vcf file using HaplotypeCaller took me a total of 1075 minutes with my hardware in this example pipeline. (34)

```
gatk --java-options "-Xmx4g" HaplotypeCaller -R Homo_sapiens_assembly38.fasta
-I G000539600_sorted_rg_bqsr_dedup.bam
-O G000539600.g.vcf.gz
-ERC GVCF
```

Using the following command, I received a VCF file with all potential variants. This was a quick process taking a total of 26 minutes.

```
gatk --java-options "-Xmx32g" GenotypeGVCFs \
-R Homo_sapiens_assembly38.fasta \
-V G000539600.g.vcf.gz \
-O G000539600.vcf.gz
```

The resulting set of variants is not yet ready for diagnosis. Filtering is required. Any error or omission in this step as well as in previous steps can have wide-ranging consequences. This can be done using the GATK Variant Quality Score Recalibration (VQSR) tool, which uses machine learning methods to filter variants. It is based on the use of high quality reference data on which it builds and trains a predictive model to filter out faulty variants. (34)

First, we create a file containing only SNPs without indels or other variants. We will use the SelectVariants tool from GATK to do this. (34)

```
./gatk SelectVariants \
-R Homo_sapiens_assembly38.fasta \
-V G000539600.vcf.gz \
--select-type-to-include SNP \
--exclude-non-variants true \
-O G000539600_snp_raw.vcf.gz
```

```
./gatk VariantRecalibrator \
-R Homo_sapiens_assembly38.fasta \
-V G000539600_snp_raw.vcf.gz \
--resource:hapmap,known=false,training=true,truth=true,prior=15.0
hapmap_3.3.hg38.vcf.gz \
--resource:omni,known=false,training=true,truth=false,prior=12.0
1000G_omni2.5.hg38.vcf.gz \
--resource:dbsnp,known=true,training=false,truth=false,prior=2.0
Homo_sapiens_assembly38.dbsnp138.vcf \
--resource:1000G,known=false,training=true,truth=false,prior=10.0
1000G_phase1.snps.high_confidence.hg38.vcf.gz \
--an DP -an QD -an FS -an SOR -an ReadPosRankSum \
-mode SNP \
-O G000539600.recal \
--tranches-file G000539600.tranches \
--rscript-file G000539600.plots.R
```

The product of this operation is the important file G000539600.recal, which contains modifications that can be applied to filter out variants from the VCF file. The optional output also includes the graphs below with statistics on the filtered variants.

```
./gatk ApplyVQSR \
  -R Homo_sapiens_assembly38.fasta \
  -V G000539600_snp_raw.vcf.gz \
  -O G000539600_snp_filtered.vcf.gz \
  --truth-sensitivity-filter-level 99.0 \
  --tranches-file G000539600.tranches \
  --recal-file G000539600.recal \
  -mode SNP
```

Then we select from the file only the variants that were not filtered using the SelectVariants command with the --exclude-filtered true parameter.

```
./gatk SelectVariants \
-R Homo_sapiens_assembly38.fasta \
-V G000539600_snp_filtered.vcf.gz \
--exclude-filtered true \
-O G000539600_snps_final.vcf.gz
```

The result is a VCF file usable for diagnostics and further analysis. The variants contained in it can be considered in both polygenic score calculations and searching for monogenic variants.

3. Methods of analysis of genomic data

There are various methods of analysis that can be used to process genomic data. The approach differs based on whether we're looking at polygenic causes or monogenic.

3.1 Monogenic Variant Annotation and Search

Genotype information derived from whole-genome sequencing in the form of variant call files is too large to be browser through manually. Therefore, programs for annotation and search are used with these files, which can easily contain many millions of variants.

3.1.1 IGV

Integrative Genomics Viewer is a freely-available manual genomic data imaging tool developed at the Broad Institute of MIT and Harvard. It is optimized to work with large files, including several whole genome sequences at once. It can visualize BAM/CRAM or VCF files. The selected reference genome is automatically loaded from the Internet. It is an ideal tool when it is necessary to verify the exact conditions under which a variant was called in a VCF file. It allows us to analyze the precise number of short reads that contributed to the variant call. (42)

3.1.2 Gene.iobio.io

Gene.iobio is a tool that simplifies the transition from single-gene and panel-based genetic testing to the analysis of whole-exome and whole-genome sequencing data. It is a variant interrogation and prioritization web-based application that uses a clinically-driven approach to select genes of importance. Genes are selected using its sister tool, genepanel.iobio.io, which outputs genes of importance based on described phenotypes. The variants for the selected genes are displayed graphically in the web application, which allows the tool to be used without extensive technical training, as opposed to many tools which work similarly but require more profound technical knowledge for interaction with their console-based interfaces. (43)

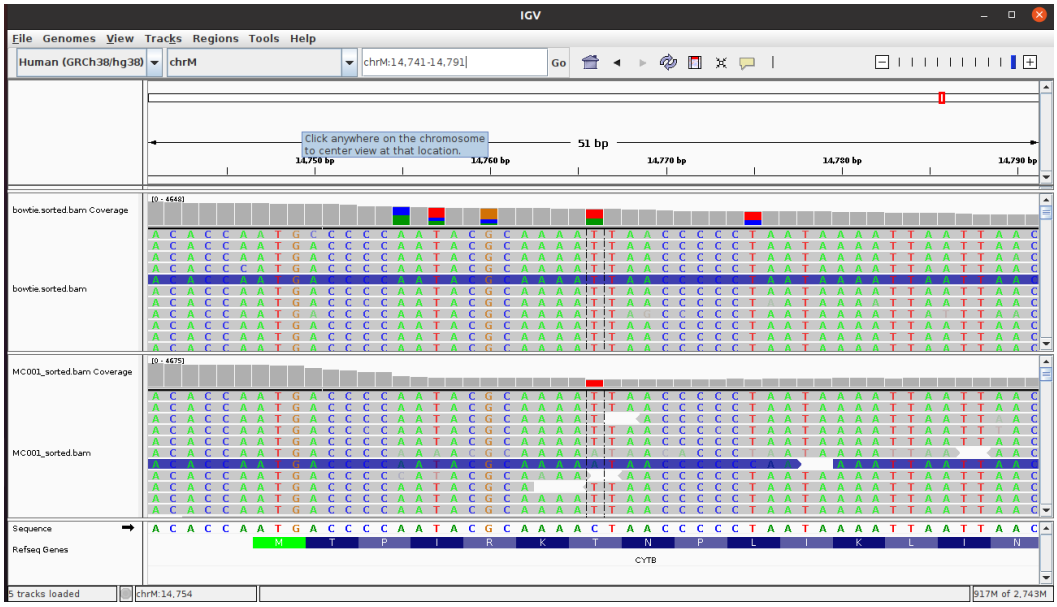

Figure 10 - Screenshot from IGV. This screen shows two whole-genomes loaded in the program. The differences in coverage are caused by different aligners used with various parameters.

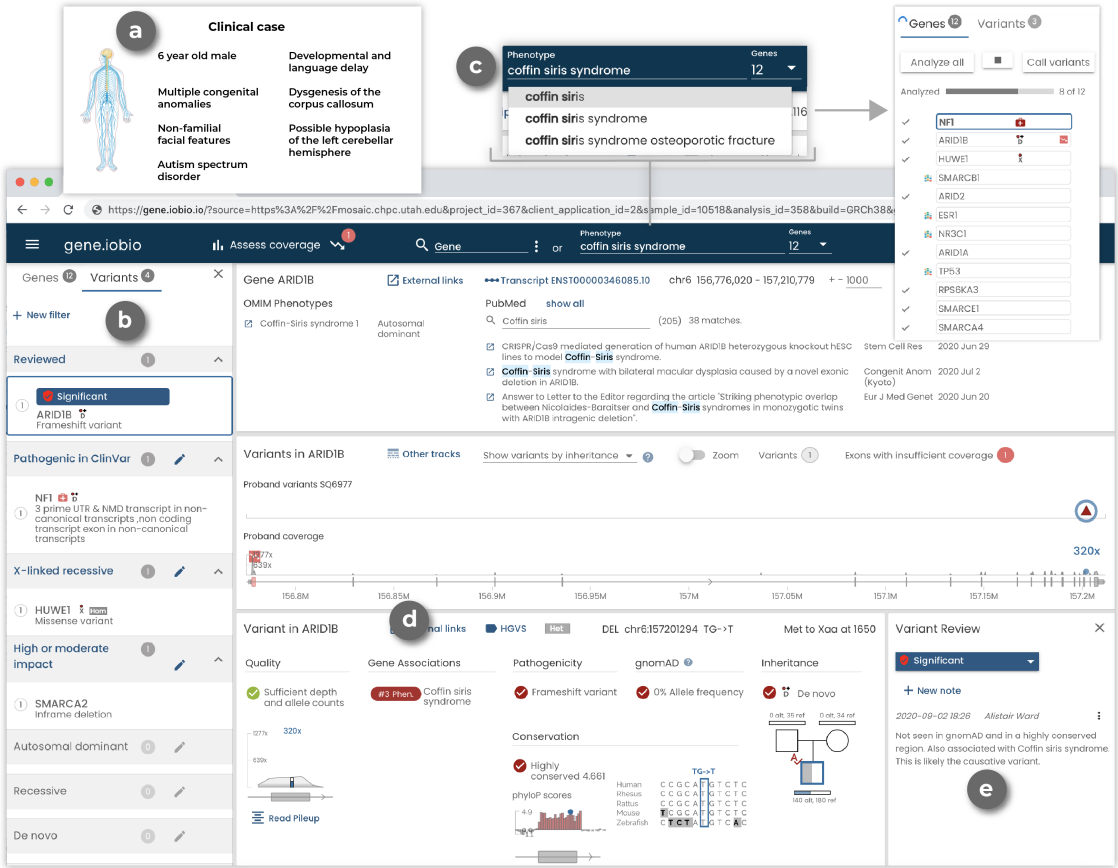

Figure 11 - A representative clinical case as viewed in gene.iobio. (a) Clinical case information and phenotype description. (b) Prioritized variants are shown in the left panel. A list of all loaded genes is available in the Genes tab of the left panel. (c) Phenotype input components. (d) Variant details. (e) Variant review. Caption and picture taken from the original study. (43)

3.2 Polygenic Risk Score Calculation

Polygenic risk is evaluated in diseases that are caused by a range of genetic mutations across multiple genes. These mutations are each assigned an effect size according to genome-wide association studies that compare genotypes between cohorts of affected patients and healthy controls. As a result of those studies, some variants are shown to be statistically more frequent in the affected patients. Those variants are then considered risk factors, if they pass a certain threshold.

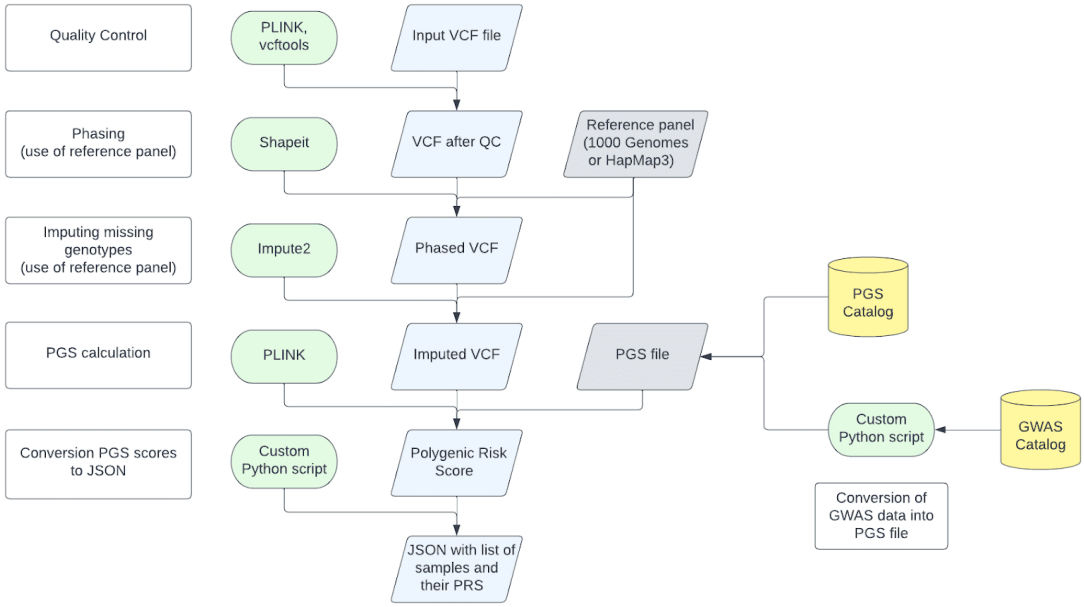

Figure 12 – Polygenic Risk Score Pipeline. The process begins with quality control of the input VCF file. Afterward, the VCF is phased using an available reference panel. If the VCF file is incomplete, the missing genotypes can be imputed based on the reference panel closest to the patient. The imputed VCF is then used for the calculation of polygenic risk scores. One approach is to use PLINK and PGS files, which consist of genotype-phenotype associations that are freely available in the PGS Catalog. (44) (45) Data directly from the GWAS catalog can be used too, but it needs to be processed into PGS files before using PLINK. The final output of the calculation can be stored in a sheet or converted into a JSON file for display in a web or mobile application.

As of current knowledge, Wilson's disease is a monogenic disease, that is caused by specific mutations in the ATP7B. Therefore, the process of calculating polygenic risk scores isn't directly applicable to its diagnosis. Nevertheless, it might be useful for assessing genetic risk scores for some associated risk factors.

4. Analysis of genomic data of the affected patients

Three families have agreed to join a study which could help expand the knowledge of how Wilson's disease can develop in cases with inconclusive family history. All the families are affected by atypical form of disease that manifests as Wilson's disease, but only in the probands. The most notable phenomena that is shared by all these cases is that the disease has appeared de novo in probands, not adhering to the usual pattern of its inheritance.

4.1 Family Number 1

Proband from the family number 1 is a male born in the year 1972. He's married to a spouse with whom he has two children which do not show any Wilson's disease related complications so far.

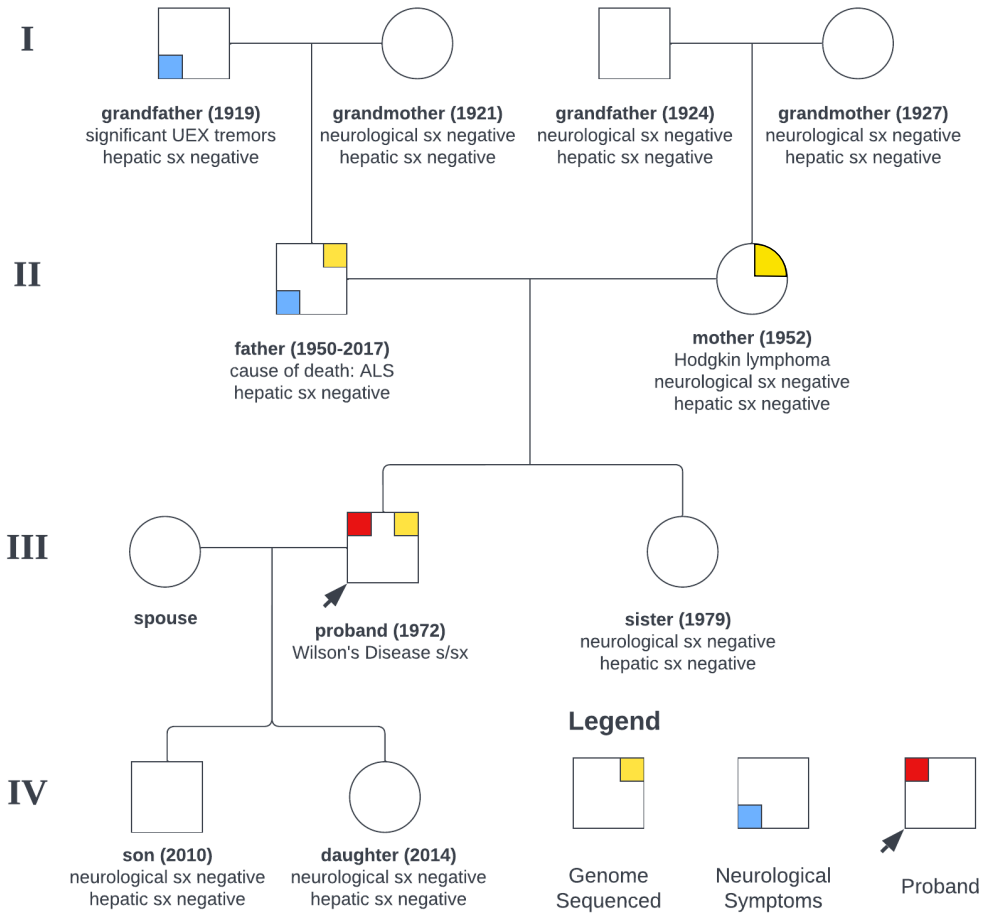

Figure 13 - Pedigree chart of the family no. 1. Whole genomes of three members of the family were sequenced and analyzed.

The proband's first symptoms appeared approximately at age 15 and manifested as progressive upper limb tremors (1987). The diagnosis was made in the year 1999 based on increased hepatic copper content in dry tissue of 324 µg/g. Ceruloplasmin in serum is normal and there are no Kayser-Fleischer rings present. In 28 years, the patient was prescribed penicillamine (brand name Metalcaptase), which initially worsened his condition and later improved with clonazepam. In 2002, the patient discontinued the use of penicillamine for three years, and tremors returned in 2006, which prompted the patient to resume the medication course. The patient negates any psychiatric symptoms. His Father died in the year 2017 from motor neuron disease. Grandfather had upper limb tremors that have been worsening with age.

4.2 Family Number 2

Proband from this family is a woman born in 1973. Her family has no prior WD history.

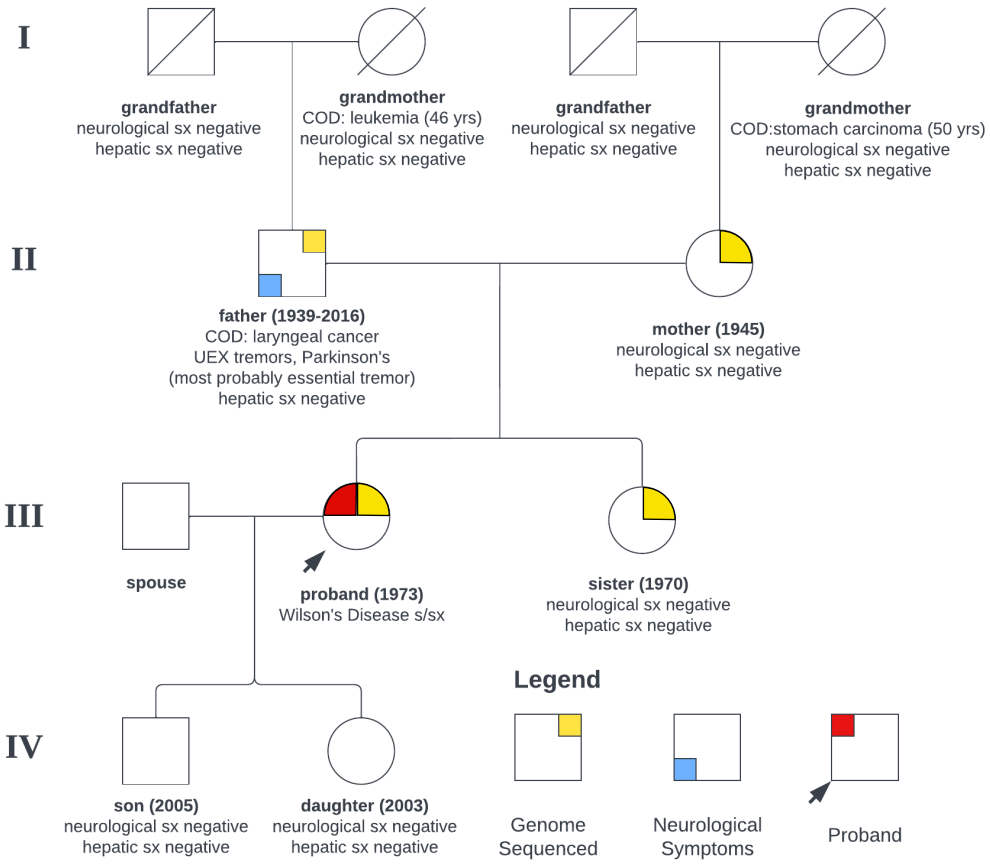

Figure 14 - Pedigree chart of the family no. 2. Whole genomes of four members of the family were sequenced and analyzed.

She has first presented with psychiatric manifestations in the form of depression approximately in 1990. Diagnosis was made based on increased hepatic dry tissue copper content, and low serum ceruloplasmin levels. Management of the disease was started with zinc, which has led to complete stabilization of her condition. Neurological symptoms have not appeared, but psychiatric problems have worsened slightly. The patient was prescribed lithium and an antidepressant (escitalopram). She has two children, a son and a daughter, born in 2005 and 2003 respectively. The children have no manifestation of Wilson's disease so far.

4.3 Family Number 3

The proband is a woman born in 1964, who has been diagnosed in the year 2000.

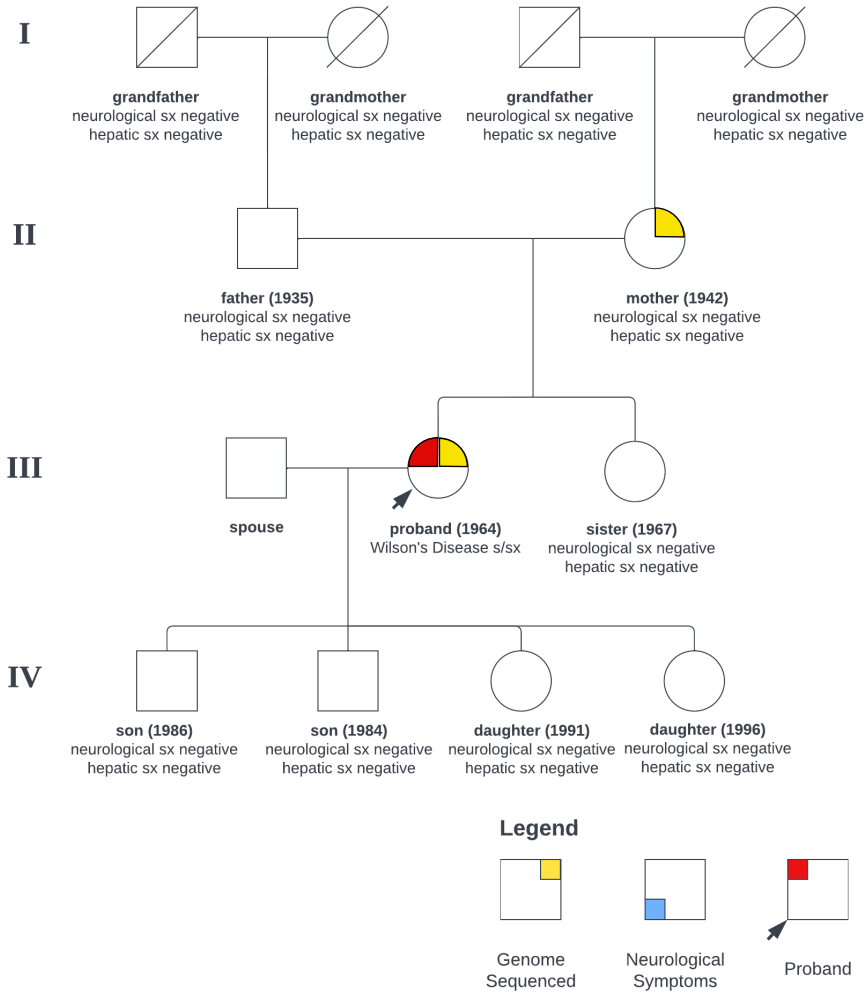

Figure 15 - Pedigree chart of the family no. 3. Whole genomes of two members of the family were sequenced and analyzed.

The diagnosis was based on increased hepatic dry tissue copper content. Since the diagnosis, she is treated with 300mg of penicillamine three times a day. The first manifestation that led to diagnosis were tremors of the whole body and troubles with pronunciation and speaking. The symptoms of the disease have so far remained minimal, following successful treatment with penicillamine. Her family has no prior history of any neurological or hepatic diseases that might be connected to her current diagnosis.

4.4 Methods

I've received both raw data and variant files of probands from the three families. The goal is to generate their variant call files and analyze them for the presence of variants that are either pathogenic and known or other variants that have a significant coding consequence and thus may be candidates for novel WD-causal mutations.

First, I've processed the raw data according to the GATK Best Practices to produce a set of variant call files aligned to the GRCh38 reference genome. The pipeline I used consists of the steps described in **the 2.2 WGS Data Processing Pipeline**.

I've analyzed the affected families' variants in the ATP7B gene.iobio real-time variant analysis tool. I've set up a local instance of gene.iobio according to the instructions listed on its GitHub repository. After analyzing the variants in the ATP7B, I've moved on to searching for variants in other genes that might be causal to the phenotype of the patients. For that, I've used genepanel.iobio.io tool, which generates a list of relevant genes based on suspected phenotypes and conditions. (43)

Afterward, I again used gene.iobio to search for variants of interest in the selected pool of relevant genes. I filtered variants with moderate to high coding consequences and then further analyzed them using Ensembl Variant Effect Predictor (VEP). Followingly, I selected high-effect variants from the VEP report and then did literature research for them in the VarSome database. (46) (47)

4.5 Results

Analysis in the genome of the proband from family number 1 has not revealed any known pathogenic mutations in the ATP7B gene. Some variants have had a moderate impact. Neither the parents are carriers of any ATP7B pathogenic mutation. I have analyzed the proband and both of his parents for mutations in the ATP7B region. The following mutation is present in the coding regions of the proband gene and in their father, who had a medical history of neurological manifestations. None of these mutations are present in the healthy, asymptomatic mother of the proband.

Table 5 - Variants in the genome of the proband form the family no. 1. All variants listed here are also present in the father of the subject, but none of them are present in the mother's genome. The mutations result in various missense mutations.

rs1801249	heterozygous	NC_000013.11:g.51941218A>G	Val1140Ala
rs732774	heterozygous	NC_000013.11:g.51949672C>T	Arg952Lys
rs1061472	heterozygous	NC_000013.11:g.51950352T>C	Lys832Arg
rs1801244	heterozygous	NC_000013.11:g.51970669C>G	Val456Leu
rs1801243	heterozygous	NC_000013.11:g.51974004A>C	Ser406Ala

I have then used the Variant Effect Predictor by Ensembl to further analyze these variants' consequences. The variant rs1061472 has the lowest PolyPhen score of 0.619, indicating that it is probably damaging. Also, the transcript that is affected by this variant is targeted by nonsense-mediated decay. This mutation probably is not the causal variant, given that its frequency in the general population is approximately 0.53 (based on gnomAD). (46)

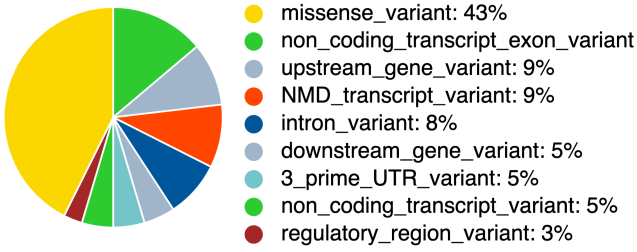

Figure 16 - Pie chart showing the coding consequences of the ATP7B variants in the proband from the family no. 1. Generated by Ensembl. (46)

I used the same process to analyze the proband, her sister, mother, and father from the family no. 2. Analysis of the ATP7B gene has led to no significant findings, given that all family members have a very similar ATP7B genotype, including a few common, benign missense mutations. Therefore, I have moved on to analyzing other genes which might be responsible for pathogenic development in the proband.

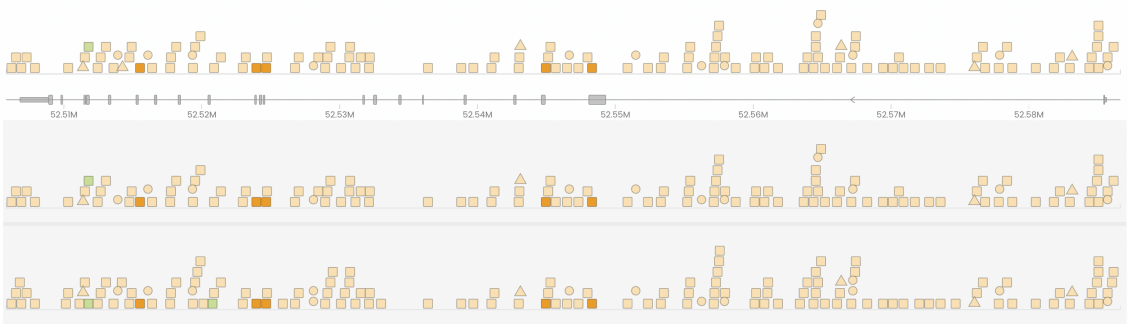

Figure 17 - User interface of the gene.iobio.io tool. The image shows the variants in the ATP7B gene as colored shapes. Squares represent missense mutations, circles are insertions, triangles are deletions. The shapes are colored based on their predicted coding consequence. The green represents low consequence, yellow is intermediate, orange is moderate, and red is high. The top track is of the proband and below it are the tracks for mother and father respectively.

Using genepanel.iobio.io, I gave selected a panel of genes that might be connected to the essential tremor and Parkinson's disease in the proband's father. I have analyzed the following genes: SNCA, FBXO7, DCTN1, TSC1, TSC2, PARK2, UCHL1, LRRK2, PINK1, SNCAIP, GAA, VPS35, PARK7, PRKAG2, LAMP2, HTRA2. No pathogenic or any other high-consequence variants were found in any of the family members.

I have entered Wilson's disease phenotype description in the genepanel.iobio.io tool and have received the following list of genes:

Upon analyzing them, I have found multiple high-impact de novo mutations in the CTBP2 gene of the proband. Those variants are not present in the genomes of parents. I have searched journals for mentions of why the CTBP2 gene should be associated with Wilson’s disease but have not found any results. The genepanel.iobio.io tool works by literature mining, so that it might have probably picked up the gene in some article describing various diseases and made this connection. Nevertheless, the variants are significant and could be further studied. The variant rs968566248,

NC_000010.11:g.124994597C>A results in stop gain in in the CTBP2 gene. The sister of the proband has only the mutation rs78681531.

Table 6 - De novo mutations in the CTBP2 gene of the proband from the family no. 2.

rs968566248	heterozygous	NC_000010.11:g.124994597C>A	Glu758Ter
rs78681531	heterozygous	NC_000010.11:g.124994621C>A	Asp750Tyr
rs75839774	heterozygous	NC_000010.11:g.124989543T>C	Asn978Ser
rs201671356	heterozygous	NC_000010.11:g.124994581A>G	Val763Ala

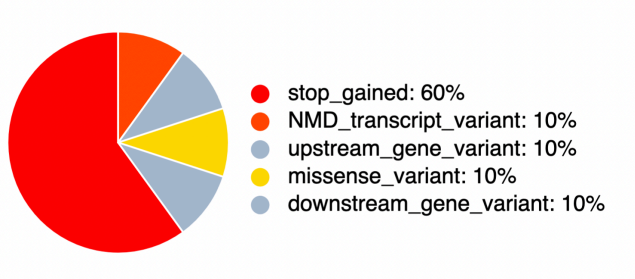

Figure 18 - Pie chart showing the coding consequences of the CTBP2 variants in the proband from the family no. 2. Generated by Ensembl. (46)

Again, I implemented the same process to analyze variants in the proband and her mother from the family no. 3. Unfortunately, the genomic data of the father are unavailable for this study. The search did not yield any variants in the ATP7B gene, which could be causal to Wilson's disease. The proband is homozygous for various missense mutations in the ATP7B gene, but all of them are common in the population (approx. frequency 0.4 to 0.6).

5. Discussion

Even though genetics are a major decisive factor in the diagnosis of Wilson's disease, many studies have shown that the genotype-phenotype correlation is much looser than expected. Still, the presence of a pathogenic genotype can predict a higher risk of developing Wilson's disease with varying levels of severity. With the use of the current whole-genome sequencing technologies, which are becoming increasingly affordable, a more comprehensive screening could be done. The genetic prevalence of homozygotes is more than four times higher than the number of clinically diagnosed cases. (3)

The cases in this study were outliers because their symptoms have appeared de novo, without any specific family history. They all received a negative result in targeted genetic testing, which looked for the most common mutations in the ATP7B gene after all developed symptoms characteristic of Wilson's disease. Their families have very little to no history of symptoms that would usually be associated with the typical manifestation of Wilson's disease. (20)

Family no. 1 had some history of neurological symptoms from the father's side of the proband. The father of the proband has died of amyotrophic lateral sclerosis (ALS), but given that around 90 to 95% of cases of this disease have no known cause and are therefore considered sporadic ALS, this may have been the case as well.

In the case of family no. 2, the proband has primarily psychiatric manifestation typical to Wilson's disease. Previous genetic testing has not found any causal variant in the ATP7B gene, and neither did analysis in this study. However, high-impact variants were found to appear de novo only in proband in the CTBP2 gene. In this case, the parents and the sister of the proband were sequenced. The analysis was made on her genome too. Four heterozygous variants in the gene might be pathogenic based on their effect prediction. Unfortunately, the variants are very rare, so no other documented cases exist. These variants can be considered a causal candidate, which might play some role in the development of clinical manifestation in this patient, given that the CTBP2 is a corepressor gene that might target various other genes and modify their expression.

In the case of **family no. 3**, there are no known hepatic or neurological symptoms in both the mother and the father of the proband, including in their respective grandparents. Unfortunately, there was no access to data from the father of the proband from that family, but given the autosomal recessive inheritance pattern, it may not have yielded any more insights into the problem.

Even though the findings of this study remain inconclusive, the results are an example of how Wilson's disease can appear spontaneously and atypically even in individuals with no prior known family risk. Several studies have tried to find a genotype-phenotype correlation but without much success. People with the same genotype often present with a wide variety of manifestations and with various levels of disease severity. (40) The age of onset also varies greatly, but most people are usually diagnosed until they reach age 40. A rare case of a woman who was first diagnosed at age 72 has also been studied. (41) All of this raises the question of whether other factors may affect the development and presentation of Wilson's disease. There have been studies about modifier genes and epigenetic factors affecting Wilson's disease.

A study has been done on 248 patients with Wilson's disease and identified that variants in the ESD and INO80 gene might be associated with the risk of neurological manifestation (42). Also, APOE and MBD6 variants have been linked to later onset of symptoms in Wilson's disease. (43) Therefore, it seems that even though it is a disease with a monogenic cause, it has polygenic risk and protective factors that influence or even completely prevent its clinical manifestation from developing. The case from **family no. 2** may also be linked to the mutations in the corepressor gene CTBP2, but the issue is that there are no underlying changes in the ATP7B, so it would seem that some other mutation affects the gene's functionality and therefore causes the manifestation in the patient.

Epigenetic factors affecting Wilson's disease are even less studied than modifier genes. There seems to be a link between environmental and epigenetic factors that are related to copper accumulation and methionine metabolism. Copper accumulation plays a role in the inhibition of enzymes that lead to decreased methylation reactions that are important in providing a supply of methyl groups for modifications of DNA, RNA, and proteins. Copper can also induce oxidative stress damage in mitochondria. One study has found

that methylated regions in patients affected by Wilson's disease have differentiated them into two cohorts with hepatic and neurologic symptoms. (40)

Future investigation in the cases included in this study might be focused on epigenetic factors and wider research on possible modifier genes. Also, sequencing other living relatives and children of the affected patients might provide deeper insights. All the children from the included families are still young, which makes it unlikely that they will present some symptoms of Wilson's disease. Nevertheless, they should be monitored in order to prevent complications from even arising. Early detection remains one of the most effective methods in managing Wilson's disease today. (2)

References

1. **Ronald F. Pfeiffer,** Wilson's Disease. 2007, Semin Neurol, pp. 123-132.

2. **Anna Członkowska, Michael L. Schilsky.** Wilson Disease, Handbook of Clinical Neurology, Volume 142. s.l. : Elsevier Science, 2017.

3. **Leung M.** The Present and Future Challenges of Wilson's Disease Diagnosis and Treatment, Clin Liver Dis (Hoboken), 2021, pp. 267-270.

4.Wilson's disease for patients and families. EuroWilson. [Online] http://www.eurowilson.org/data/pdf/EN-adults-eurowilson-for-adults.pdf.

5. **Vrabelova**, Frequency of Wilson Disease recessive mutation in the Czech Population, Molecular Genetics and Metabolism 2005, pp. 277-285.

6. **Dedoussis, Gomes A. & George V**. Geographic distribution of ATP7B mutations in Wilson disease, Annals of Human Biology, 2016, pp. 43:1, 1-8.

7. **Landrum MJ, Lee JM, Benson M et al.** ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Research, 2018

8. **Tümer, Z., & Møller**, Menkes Disease., , European Journal of Human Genetics, 2009, pp. Vol. 18, Issue 5, pp. 511–518.

9. **Cox, D. W., & Moore, S. D.** Copper transporting P-type ATPases and human disease, Journal of bioenergetics and biomembranes, 2002, pp. 34(5), 333–338.

10. **Lenartowicz, M., & Krzeptowski,** W. Budowa i funkcja białek ATP7A i ATP7B--ATPaz transportujacych jony miedzi, Postepy biochemii, 2010, pp. 56(3), 317–327.

11. **National Library of Medicine**, ATP7B ATPase [Online], 2022. https://www.ncbi.nlm.nih.gov/gene/540.

12. **AlphaFold**. AlphaFold Protein Structure Database. Copper-transporting ATPase [Online] https://alphafold.ebi.ac.uk/entry/P35670.

13**. Schushan M., Bhattacharjee A., Nir Ben-Tal, Lutsenko S.**, A structural model of the copper ATPase ATP7B to facilitate analysis of Wilson disease-causing mutations and studies of the transport mechanism, Metallomics, 2012, pp. Volume 4, Issue 7, Pages 669–678.

14. **Fanni D, Pilloni L, Orru S et al.** Expression of ATP7B in normal human liver, Eur J Histochem, 2005, pp. 49 (4): 371–378.

15. **Chang, I. J., & Hahn, S. H.** The genetics of Wilson disease, Handbook of clinical neurology, 2017, pp. 142, 19–34.

16. **Przybyłkowski, A., Gromadzka, G., Chabik, G., Wierzchowska, A., Litwin, T., & Członkowska, A.** Liver cirrhosis in patients newly diagnosed with neurological phenotype of Wilson's disease. Functional neurology, 2014, pp. 29(1), 23–29.

17. **Herbert L. Fred, MD, Hendrik A. van Dijk.** Kayser-Fleischer ring. [Online] September 14, 2007. http://cnx.org/content/m15007/latest/.

18. **Samarth Kabbur, Emmanuel Roy, Samhith Komatreddy, Harish Jawahar.** Github. [Online] 2020. https://emmanuel-roy.github.io/Killer-Eye/.

19. **NHGRI. National Human Genome Research Institute**. [Online] 2021. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.

20. **genomeWeb**. As MGI Preps US Launch of Sequencers, Customers From Europe and Canada Report Initial Results. genomeWeb. [Online] 2020. https://www.genomeweb.com/sequencing/mgi-preps-us-launch-sequencers-customers-europe-and-canada-report-initial-results#.YuYoksFBzrw.

21. **Coffey AJ, Durkie M, Hague S et al.** A genetic study of Wilson's disease in the United Kingdom. 2013, Brain, pp. 136(Pt 5):1476-87.

22. **Leung, M., Aronowitz, P. B., & Medici, V.** The Present and Future Challenges of Wilson's Disease Diagnosis and Treatment, Clinical liver disease, 2021, pp. 17(4), 267–270.

23. **Ferenci P, Stremmel W, Czlonkowska A, et al.** Age and sex but Not ATP7B genotype effectively influence the clinical phenotype of Wilson disease, Hepatology, 2019, pp. 69:1464-1476.

24. **Mareček Z, Brůha R.** Wilsonova choroba [Wilsons disease], Vnitřní lékařství, IV. interní klinika 1. LF UK, 2013, Vols. 59(7):578-83.

25. **Centrael Tyson Evans**, EyeWiki. Wilson's Disease/Kayser Fleischer Ring. [Online] April 24, 2022. https://eyewiki.aao.org/Wilson%27s_Disease/Kayser_Fleischer_Ring#cite_note-1.

26. **Coni P, Pichiri G, Lachowicz JI et al.** Zinc as a Drug for Wilson's Disease, Non-Alcoholic Liver Disease and COVID-19-Related Liver Injury, Molecules, 2021, 26(21):6614.

27. **Mulligan C, Bronstein JM.** Wilson Disease: An Overview and Approach to Management, Neurol Clin, 2020, pp. May;38(2):417-432.

28. **NCBI**, LiverTox: Clinical and Research Information on Drug-Induced Liver Injury, National Institute of Diabetes and Digestive and Kidney Diseases, [Online] 2020, https://www.ncbi.nlm.nih.gov/books/NBK548119/#_NBK548119_pubdet_.

29. **National Center for Advancing Translational Sciences**, Inxight Drugs, [Online] August 5th, 2022, https://drugs.ncats.io/substance/206J6X63BE.

30. **Ahmad A, Torrazza-Perez E, Schilsky ML.** Liver transplantation for Wilson disease, Handbook of Clinical Neurology, 2017, pp. 142:193-204.

31. **Ultragenyx**, CYPRUS 2+ study update, www.wilsondisease.org, [Online] February 17, 2022, https://wilsondisease.org/wp-content/uploads/2021/08/Ultragenyx-Wilson-Disease-Update-Feb-2022.pdf.

32. **Moini M, To U, Schilsky ML.** Recent advances in Wilson disease, Transl. Gastroenterol. Hepatol, 2021

33. **Lodish, Harvey F.** Molecular Cell Biology, 4th ed. New York : W.H. Freeman, 2000.

34. **Van der Auwera GA, Carneiro M, Hartl C et al.** From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline, Curr Protoc Bioinformatics, 2013, Vols. 43:11.10.1-11.10.33.

35. **Illumina**. FASTQ files explained. [Online], 2021. [Cited: 25.7.2022.] https://support.illumina.com/bulletins/2016/04/fastq-files-explained.html.

36. **Petr Danecek, James K Bonfield, Jennifer Liddle et al.** Twelve years of SAMtools and BCFtools, GigaScience, 2021, Issue 2, Volume 10 .

37. **R. Li H. and Durbin. s.l.** Fast and accurate short read alignment with Burrows-Wheeler Transform, 2009, Bioinformatics.

38. **Sherry, S.T., Ward,M. and Sirotkin,K.** dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation, Genome Res, 1999, 677–679, Vol. 9.

39. **Andrews, S.** FastQC: A Quality Control Tool for High Throughput Sequence Data. [Online] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

40. **S., Wingett SW and Andrews. s.l.** FastQ Screen: A tool for multi-genome mapping and quality control, 2018, F1000Research

41. **Broad Institute**, Picard Toolkit. [Online] GitHub Repository, 2019. https://broadinstitute.github.io/picard/.

42. **James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov.** Integrative Genomics Viewer, Nature Biotechnology, 2011, Vols. 29, 24–26 .

43. **Di Sera, T., Velinder, M., Ward, A. et al. s.l.** Gene.iobio: an interactive web tool for versatile, clinically-driven variant interrogation and prioritization, 2021, Sci Rep

44. **Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC.** PLINK: a toolset for whole-genome association and population-based linkage analysis.: American Journal of Human Genetics, 2007, Vol. 81.

45. **Samuel A. Lambert, Laurent Gil, Simon Jupp, Scott C. Ritchie, Yu Xu, Annalisa Buniello, Aoife McMahon, Gad Abraham, Michael Chapman, Helen Parkinson, John Danesh, Jacqueline A. L. MacArthur and Michael Inouye. s.l.** The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation, Nature Genetics, 2021.

46. **McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. s.l.** *The Ensembl Variant Effect Predictor,* Genome Biology, 2016, Vol. 17(1):122.

47. **Christos Kopanos, Vasilis Tsiolkas, Alexandros Kouris, Charles E Chapple, Monica Albarca Aguilera, Richard Meyer, Andreas Massouras.** *VarSome: the human genomic variant search engine,* Bioinformatics, 2019, Vol. 35.

48. **Medici V, LaSalle JM. s.l.** *Genetics and epigenetic factors of Wilson disease,* Ann Transl Med, 2019.

49. **Cao C, Colangelo T, Dhanekula RK, Brandt D, Laothamatas I, Thapar M, Herrine SK, Civan JM.** *A Rare Case of Wilson Disease in a 72-Year-Old Patient,* ACG Case Rep Journal, 2019

50. **Kluska A, Kulecka M, Litwin T, et al. s.l.** *Whole-exome sequencing identifies novel pathogenic variants across the ATP7B gene and some modifiers of Wilson's disease phenotype,* Liver Int, 2019

51. **Litwin T, Gromadzka G, Członkowska A.** *Apolipoprotein E gene (APOE) genotype in Wilson's disease: impact on clinical presentation,* Parkinsonism Related Disorders, 2012.

Specifications of the used hardware

Local Computer:

Motherboard: Gigabyte Aorus Ultra X570

CPU: AMD Ryzen 9 3900X 12-core, 24 threads

RAM: 64GB DDR4

Graphics card: MSI GeForce RTX 3070 Ti, 8GB

Operating System: Ubuntu 20.04.3 LTS

Remote Server (provided by First Faculty of Medicine at Charles University):

Processor: 32x AMD Ryzen Threadripper 2950X 16-core

RAM: 128 GB

Operating System: Ubuntu 20.04.1 LTS