POMORSKI UNIWERSYTET MEDYCZNY W SZCZECINIE

Konrad Podsiadło Mgr.

Analiza związku mutacji nonsensownych z wybranymi fenotypami w populacji polskiej

Analysis of the relationship between nonsense mutations and selected phenotypes in the Polish population

Rozprawa doktorska w dziedzinie nauk medycznych i nauk o zdrowiu

Dyscyplina nauki medyczne

Promotor: dr hab. n. med. Jeremy Clark, prof. PUM

Pracę wykonano w Zakładzie Biochemii Klinicznej i Molekularnej.

Szczecin, 2023

Podziękowania

Dear dr hab. n. med. Jeremy Clark, prof. PUM,

I would like to express my deepest gratitude for being my doctoral thesis advisor and for sharing with me the incredible journey we have embarked on together for the past 8 years. Your guidance and support were instrumental in helping me achieve my academic goals, and your expertise and dedication to teaching have been an inspiration to me. Throughout the years, your mentorship has not only helped me develop my research skills, but has also taught me the importance of hard work, perseverance, and discipline. Your insightful feedback and constructive criticism have challenged me to become a better scholar, and your encouragement has motivated me to strive for excellence. I am grateful for the countless hours you have spent with me, discussing ideas, reviewing drafts, and offering guidance. I will forever cherish the memories and the knowledge gained during our journey. I cannot express how much I appreciate your contributions to my personal and professional growth, and I am honored to have had you as my mentor and advisor.

Drogi Prof. dr hab. n. med. Andrzeju Ciechanowiczu,

Chciałbym złożyć serdeczne podziękowania za okazane mi wsparcie i pomoc w trakcie realizacji mojej pracy doktorskiej w Pana zakładzie. Jestem niezwykle wdzięczny za możliwość korzystania z Pana wiedzy i doświadczenia, które pozwoliło mi pogłębić swoje umiejętności oraz rozwinąć swoje zainteresowania badawcze. Pragnę podkreślić, że Pana nieoceniona pomoc w rozwiązywaniu problemów, a także cierpliwe wyjaśnienia i wskazówki były kluczowe dla mojego sukcesu. Jeszcze raz pragnę złożyć serdeczne podziękowania za okazane mi wsparcie i zaangażowanie w moją pracę naukową. Cieszę się, że miałem okazję pracować pod Pana kierunkiem

Drogi dr.n.med. Thierry Van de Wetering'u

Chciałbym wyrazić moje najszczersze podziękowania za okazaną mi pomoc naukową oraz wsparcie w trakcie moich badań. Pana wiedza i doświadczenie były nieocenione w mojej pracy, a Pana pozytywne podejście do nauki i zaangażowanie w moje projekty zainspirowało mnie do dalszego rozwoju. Nie sposób przecenić Pana wkładu w moje badania i osiągnięcia naukowe. Pana ekspercka wiedza i doświadczenie pozwoliły mi rozwiązać wiele trudnych problemów i osiągnąć wysoki poziom naukowy. Dodatkowo, Pana wsparcie i rady były dla mnie bardzo ważne w trudnych momentach, kiedy to potrzebowałem dodatkowego wsparcia.

Droga Kamilo Rydzewska,

Chciałbym wyrazić moje ogromne podziękowania za Twoje wsparcie, pomoc i motywację w trakcie pisania pracy doktorskiej. Wiedziałem od samego początku, że praca ta będzie wymagała ogromnego nakładu pracy i zaangażowania, ale Twoja obecność i wsparcie sprawiły, że cały proces był zdecydowanie łatwiejszy. Nie sposób przecenić Twojej roli w ciężkich sytuacjach, kiedy to wraz z Tobą pokonywaliśmy trudności i przeszkody, które napotkaliśmy na naszej drodze. Twoja cierpliwość, wiedza i doświadczenie pomogły nam wiele razy, gdy czuliśmy się przytłoczeni i zagubieni.

Index

Glossary:

CBC	cap-binding complex
EJC	exon junction complex
eRF1/eRF3	eukaryotic release factor
GWAS	genome-wide association study
NMD	nonsense-mediated mRNA decay
PABP	poly(A)-binding protein
PTC	premature termination codon
SMG	suppressor with morphogenetic effects on genitalia
SNPs	single nucleotide polymorphisms
UPF	up-frame shift
UTR	untranslated region
NCI	number of children per individual

1. Introduction

Mutations are often associated with negative phenomena; however mutations can play a key role in adaptation to new environmental conditions, and quite often are desirable [1]. Mutations provide an essential and vital basis for evolutionary adaptation on many levels such as, for example, involving changes in protein functionality, which is crucial to the proper working of biochemical pathways.

Many diseases are caused by mutations which introduce a premature stop codon, which often leads to a loss of range of functions or lower mRNA levels due to mRNA nonsense decay. Examples of such diseases are: cystic fibrosis, Duchenne muscular dystrophy, β-thalassemia, and many types of cancer. These kinds of changes can arise from the occurrence of germline or somatic DNA mutations, inaccurate or inefficient pre-mRNA splicing, or lack of optimization of RNA editing [2].

Currently, the significance of the occurrence of nonsense mutations (including pretermination codons) is not fully understood. The considerable range of genetic variation in human populations may partly reflect characteristic processes of adaptation to changing environmental conditions. However, genomic signatures of adaptation have not yet been fully elucidated. Understanding the extent of allelic variation in human genes within and outside populations, based on the action of demographic and evolutionary factors, is one of the main research goals of human genetics. In recent years, epigenetic factors have also been studied and these also have a significant impact on modern humans. Recent genetic studies have revealed that some genes can undergo strong selection for new alleles [3–6]. These include genes involved in lactase persistence [7], altered bitter taste [8], reduced olfactory receptors [9] and malaria resistance [5,6]. These examples of recent selection in humans have been discovered using candidate gene studies with an a priori hypothesis of selection. More recently, whole-genome approaches have been used to identify candidate genomic loci that may be the target of positive selection during human evolution. However, much is still unknown about the types of genes or biological processes commonly involved in the adaptation of modern humans. One area, the study of which might increase knowledge and give a better understanding of the processes involved, lies in the study of the occurrence of nonsense mutations and their possible association with predisposition to obesity, overweight or infertility, as is discussed below. This might also lead to the identification of new treatment options [10].

1.1 Diseases caused by nonsense mutations

Nonsense mutations are sometimes referred to as any mutations which result in, for example, a nonsense protein (e.g. by a frame shift or a pretermination codon), and this more loose or generalized definition is used in this section only. (In the remainder of the thesis “nonsense mutation” is used synonymously with “pretermination codon”.)

It was estimated that nonsense mutations (as defined using the generalized definition) are involved in about 10% of patients with genetic disorders, in particular they account for 20% of all disease-associated single-base-pair mutations, and are three times more likely to come to clinical attention than missense mutations [11]. Genetic diseases caused by nonsense mutations belong to extreme classification i.e. they are either rare or very common. They either provide a rare pathology category (Duchenne muscular dystrophy (DMD), cystic fibrosis (CF), spinal muscular fibrosis (SMA)) or a frequent disease class (cancer, metabolic diseases, neurological disorders). Nonsense mutations are also found in some relevant oncogenes of many cancer patients, resulting in a complete lack of full-length protein products. This fact makes nonsense mutations a point of interest for a significant number of patients and medical researchers [12].

Types of point mutation are shown in Table 2, and each of these types might produce a nonsense mutation (including the possibility of a pretermination codon). Nonsense mutations can have severe consequences. For example, deletion in an ion channel protein causes a substantial fraction of cystic fibrosis (CF) cases, a chronic disease affecting the lungs and the digestive system. In this case one or more nucleotides are "skipped" during replication or otherwise excised, often resulting in a frameshift [13].

Table 1. Types of Mutations and Their Impact- examples.

POINT CLASS OF MUTATION		
HUMAN DISEASE LINKED TO THE MUTATION	TYPE OF MUTATION	DESCRIPTION
Sickle-cell anemia	SUBSTITUTION	One base is incorrectly added during replication and replaces the pair in the corresponding position on the complementary strand
beta-thalassemia	INSERTION	One or more extra nucleotides are inserted into replicating DNA, often resulting in a frameshift
Cystic fibrosis	DELETION	one or more nucleotides is omitted during replication

Recently, new potential therapeutics for human diseases resulting from nonsense mutation are being developed. These new approaches are based on, primarily, nucleic acids and include (**Figure 6**) [12]:

- antisense oligonucleotides to alter the processing of pre-mRNA and to modulate the expression of essential factors for NMD and translation termination; antisense oligos with exon-skipping ability to splice out pretermination codon (PTC)-harboring exons in frame
- suppressor-tRNAs to read a PTC, acting by incorporation of a cognate amino acid at the PTC position
- RNA editing,
- box-H/ACA -guide RNAs to directly modify the PTC, thus converting it back to a sense codon by targeted conversion of uridine in the first position of a PTC into pseudouridine
- CRISPR technology gene editing

Besides those mentioned, nonsense suppression by small molecules can be added, e.g. aminoglycoside antibiotics are a way to promote PTC recognition by near cognate tRNA, essentially competing with translation termination and enabling the synthesis of a full functional protein [14,15].

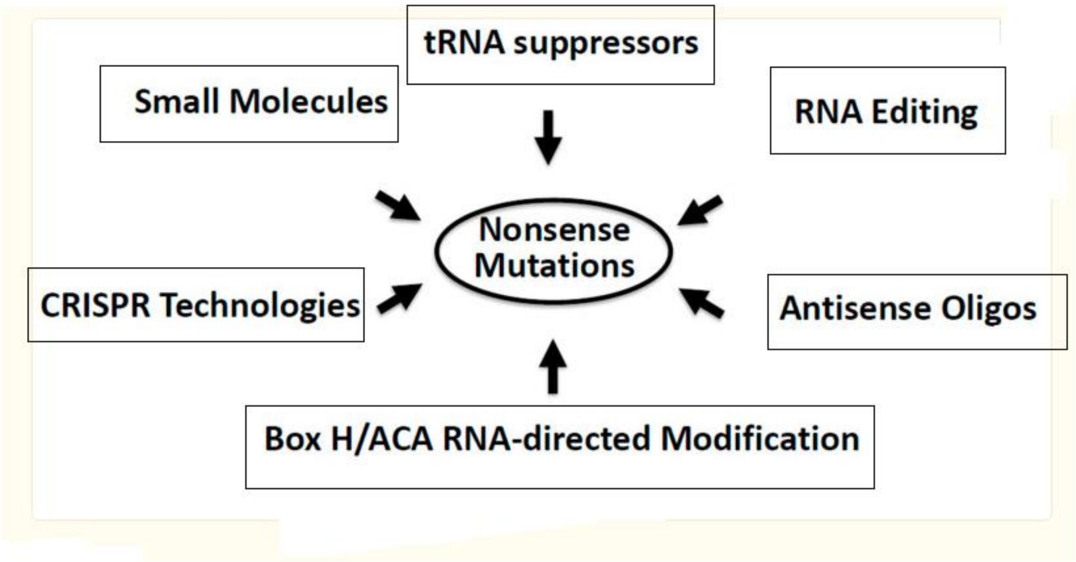

Figure 1. Nonsense suppression by various approaches. Source of information: Morais et al (2020) [12].

The above section uses a generalized definition of nonsense mutations – whereas in the rest of this thesis the definition of "nonsense mutation" is restricted to those which produce pretermination codons.

1.2 Nonsense mutations which cause pretermination codons.

A nonsense mutation is a point mutation of a single base pair (A, G, C, or T). One type of nonsense mutation results in a codon that specifies termination of translation: UAA, UGA, or UAG and it is this type of nonsense mutation that is referred to in the rest of this thesis.

A nonsense mutation involving standard bases results in the premature termination of translation of the mRNA; thus a truncated protein is formed (Figure 1). Truncated proteins, or proteins lacking parts of their original structure, might not be expressed at all or, if they are expressed, are usually unable to function properly and can lead to various genetic disorders. The relationship between these nonsense mutations and various phenotypes is that the protein truncation can either result in reduced or complete lack of expression, or drastically alter the normal working of the proteins which can cause a wide range of phenotypes, from mild to severe. These phenotypes can be determined by the sequence and three-dimensional structure of the truncated protein created as a result of the mutation. To further elaborate — a specific protein truncation by a nonsense mutation is determined by the newly created stop codon, and the structure of the protein, if expressed, is altered by the amount of amino acid chain that was removed from its original intended structure. Due to premature termination several consequences can be considered. Firstly, the mRNA carrying a premature mutation is often targeted for rapid degradation (through a cellular process known as nonsense-mediated mRNA decay), so translation might not be possible. Secondly, even if the mRNA is stable enough to be translated, the truncated protein is usually so unstable that it can be rapidly degraded through cellular mechanisms [16].

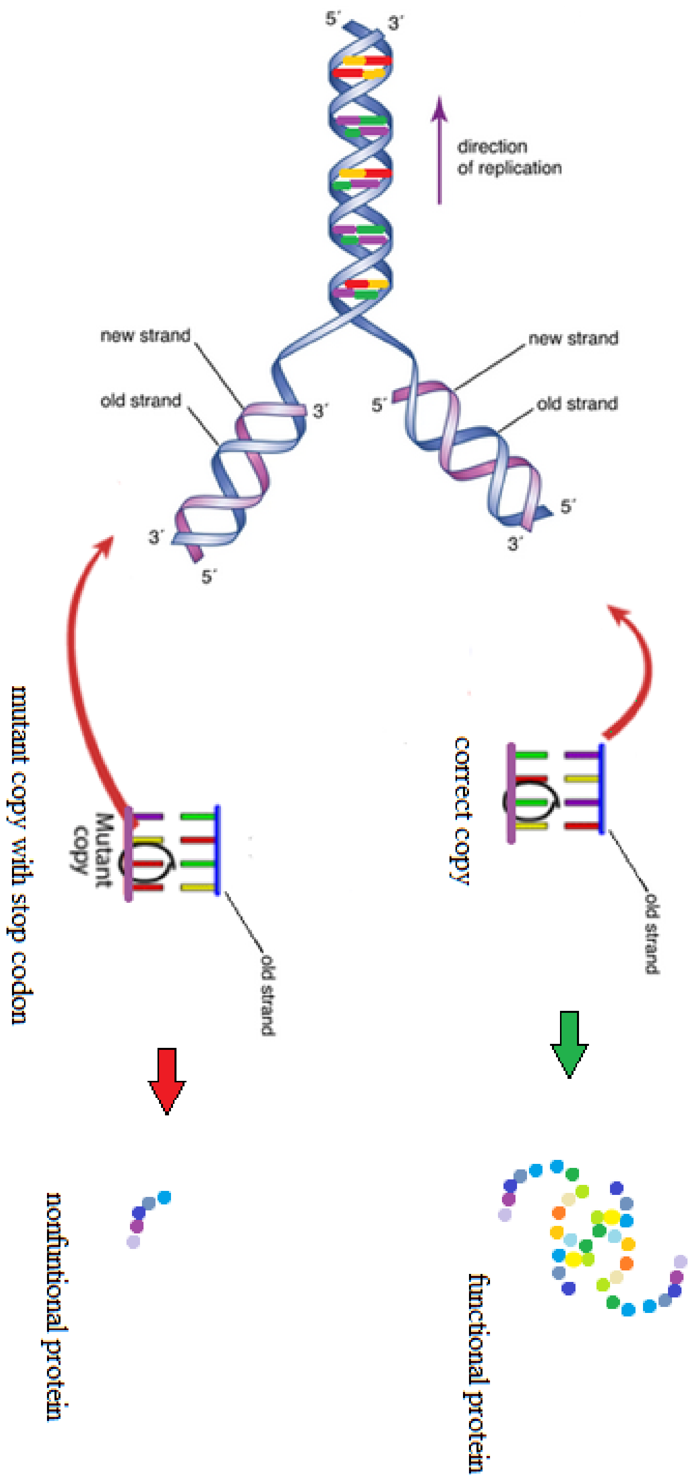

Figure 2. ***The nonsense mutation****. A nonsense mutation involving standard bases occurs in DNA when a sequence change gives rise to a stop codon rather than a codon specifying an amino acid. The presence of the new stop codon results in the production of a shortened, unfinished protein that is likely non-functional.*

Source of image: [https://www.genome.gov/genetics-glossary/Nonsense-Mutation]

Mutations can be classified for the ease of understanding their different types and their roles (**Figure 2**) and can be inherited or can occur de novo [17]. The earlier a mutation affects physiological development, the more severe the effects it will likely have on the phenotype of an organism [18].

An additional factor that can enhance a mutation's impact is its class. There are 3 basic classes of mutations. The first class are the mutations that cause a change in a single nucleotide, which are called point mutations. Within that class 3 types of mutations can be differentiated: substitution, insertion and deletion. Point mutations can have a variety of effects on an organism's phenotype, depending on where the mutation occurs and what type of change to the nucleotide sequence is made (**Table 1**). Some point mutations are silent, meaning they do not alter the amino acid sequence of a protein and have no effect on its function. Other point mutations can result in missense or nonsense mutations, which can lead to changes in protein structure or function, or even premature termination of protein synthesis [19]. Although nonsense mutations are thought to be mostly due to replication error (see Figure 2), they could be produced by any of the other mechanisms shown.

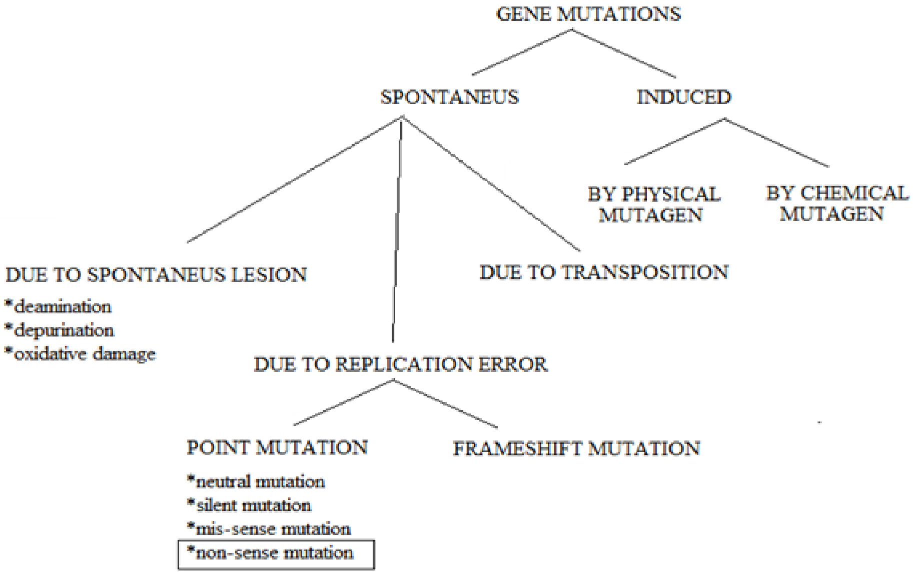

Figure 3. ***Classification of mutations.*** *A nonsense mutation is placed in the 'mutation due to replication error' sub-category of mutation spontaneous mutations. It is further placed in the 'point mutation' sub-category of mutation due to replication error (specifically, in the sub-category of point mutations – coding region). Modified from [https://www.biologyonline.com/dictionary/nonsense-mutation]*

Table 2. Differences between nonsense mutations and other types of point mutations. [https://www.biologyonline.com/dictionary/nonsense-mutation]

POINT MUTATIONS				
CHARACTERISTICS	NONSENESE MUTATION	NEUTRAL MUTATION	MISSENSE MUTATION	SILENT MUTATION = a type of neutral mutation
TYPE OF CHANGE	Nonsense codons/ premature termination codons/ premature stop codon developed	Non-synonymous or synonymous codon developed	Non-synonymous codon developed	Synonymous codon developed
CHANGES IN THE DNA SEQUENCE	YES	YES	YES	YES
CHANGES AT THE AMINO ACIDS LEVEL	YES	YES	YES	NO
AMINO ACID ENCODING	Stop codon/ Premature termination codon	SAME or different	DIFFERENT	SAME
PROTEINENCODING	Incomplete/ Non-functional protein product	SAME or different	DIFFERENT	SAME

In a study by Fujikura et al. on Premature termination codons in modern human genomes, 246 PTCs were found in which natural s-election resulted in new alleles with high frequencies (1% to 96%) of derived alleles and varying levels of population diversity . In the National Heart, Lung, and Blood Institute (NHLBI) and the 1000 Genomes (1000G) projects - two large sets of population exome sequences were used to detect recently-arising PTCs in the human genome. 16,281 segregating PTCs were discovered by a comprehensive search from a total of 7,595 people chosen from 16 ethnicities. Data about alleles with PTC mutations were obtained from the UCSC genome browser and the NCBI dbSNP database. As a reference point, a derived allele frequency (DAF) of 1% was used.

PTC genes formed protein and regulatory networks restricted to 15 biological processes or gene families, of which seven categories were previously undescribed.

For the present analysis, 141 SNPs leading to a premature stop codon were selected from the 246 described in Fujikura, K. et al. Seven of the proteins coded for were shown to be possibly involved during spermatogenesis [20]. These mutations are located in the genes: *DYX1C1, ZAN, ODF3L1, PLA2G2C, SPATA8, SPERT, STARD6* and are reported to be strongly expressed in the gonads [21]. Some of the genes harbouring the mutation have been ontologically categorised as follows [22]: metabolism (*PRAMEF2, TRPM1, CALML4, MATK, FASTKD1, ERVMER34, UNC93A, PSCA*); metabolism of drugs (*KRTAP1-1, ABCA10, ABCC12, CYP2C18, SULT1C3, UGT2A1*), immune system (*LAIR2, PXDNL, IFNE, TLR5*), zinc finger (*ZNF860, ZIM3, ZNF727, ZNF77, ZNF80*), keratin *(FLG2, KRT83, KRTAP1-1*). Based on the ontology of some genes connected with metabolism or gonads, it was decided to analyse the phenotypes: obesity, overweight, fertility and life expectancy.

1.3 Factors co-responsible for mutations

There are numerous factors which can play a role in the occurrence of nonsense mutations. Due to the fact that nonsense mutations are placed in the 'spontaneous mutation' broad category, and mostly arise as point mutations (Figure 2.), these kind of mutations do not have to arise from biological, chemical or physical mutagens. The most common cause of these mutations is thought to be spontaneous DNA changes, e.g. the DNA spontaneously breaks down or is not copied accurately (Figure 1). A nonsense mutation can be caused without mutagenic agent activity following particular processes during replication of genetic material. There are various mechanisms by which nonsense mutations can arise e.g. due to replication processes: error in DNA repair, error in transcription, error in polymerisation. However, nonsense mutations might also be caused by mutagens which can increase the rate of mutation by damaging DNA. Physical mutagens, such as UV rays and X-rays, cause damage to DNA by breaking chemical bonds and altering the structure of DNA molecules. Chemical mutagens, on the other hand, interfere with DNA replication by inserting themselves between base pairs or causing base substitutions and deletions [23].

1.4 Effects of mutations

Occurrence of a nonsense mutation in a gene can result in deleterious, neutral or beneficial outcomes. Deleterious nonsense mutations are the most common and are usually linked to genetic disorders as they lead to the truncation of a protein that performs a vital function in the body. The existence of a nonsense mutation causes an overall decline in reproductive fitness. Deleterious nonsense mutations are harmful because, even if it is expressed, the truncation of the protein can lead to the loss of its functional domains. This can result in a malfunctioning or non-functional protein, which can have serious consequences for the body. For example, if the protein is an enzyme that catalyzes a critical biochemical reaction, its truncation could lead to the accumulation of toxic metabolites and metabolic disorders. Similarly, if the protein is a receptor that mediates important signaling pathways, its truncation could disrupt cellular communication and lead to developmental defects or diseases. Therefore, nonsense mutations that cause premature termination codons (PTCs) in vital proteins are generally considered harmful and can be associated with genetic disorders or diseases. [24,25]

Neutral mutations that go undetected due to no apparent change in protein functioning (or with changes that are compensated for by other proteins) and effects are recognized as neutral in nature. However, just because a mutation is considered neutral does not mean it cannot have any impact on the organism. For instance, a neutral mutation may alter the rate of transcription or translation, which could affect the overall expression of the gene. Additionally, neutral mutations can accumulate over time and lead to genetic drift, which can have significant evolutionary consequences [25]. Many nonsense mutations have apparently near-neutral selection.

Beneficial mutations are crucial for the survival and evolution of species. They provide a selective advantage to organisms, allowing them to adapt to changing environments and improve their chances of reproducing successfully. One of the most significant benefits of beneficial mutations is that they can help organisms resist diseases and environmental stressors. For example, a mutation that confers resistance to a particular pathogen can help an organism survive in an environment where that pathogen is prevalent. Another benefit of beneficial mutations is that they can lead to the development of new traits or characteristics that enhance the fitness of organisms. This can include changes in behavior, morphology, or physiology that allow an organism to better exploit its environment or compete with other organisms for resources [26]. Some nonsense mutations appear to be beneficial.

1.5 Nonsense-mediated decay

NMD is a metabolic pathway operating at the interface between transcription and translation. NMD often has the ability to distinguish an mRNA carrying a PTC rather than the normal stop codon. It also regulates the abundance of a large number of cellular RNAs. In other words, NMD also targets non-mutant transcripts, and its regulation of normal gene expression impacts a wide range of physiological processes including cell differentiation, responsiveness to stress and development of disease [27].

Maintaining intracellular homeostasis requires precise and tightly controlled gene expression mechanisms, which are controlled by multiple levels of regulation so that damaged genes and redundant transcripts are removed. NMD (nonsense-mediated mRNA decay) is one of the most essential RNA quality control processes and gene regulatory mechanisms, which recognises and degrades PTC-containing mRNAs. NMD is a metabolic pathway operating at the interface between transcription and translation. The course of correct translation termination is shown in the **Figure 3**. NMDs exert control not only over defective transcripts, but also over normal transcripts or non-coding RNAs, and genes containing miRNA and snoRNA sequences and its regulation of normal gene expression impacts a wide range of physiological processes including cell differentiation, responsiveness to stress and development of disease. The miRNAs (microRNAs) are involved in RNA silencing and post-transcriptional regulation of gene expression. snoRNAs are involved after transcription to pre-rRNA molecules. The pre-rRNA undergoes a complex pattern of nucleoside modifications, include methylations and pseudouridylations, guided by snoRNAs, prior to cleavage by exo- and endonucleases, NMD activity engages different enzymes to destroy the transcript.

Translation of aberrant mRNAs could, in some cases, lead to deleterious gain-of-function or dominant-negative activity of the resulting proteins. As well as this, NMD uses multiple proteins at multiple stages and is an extremely complex process that also affects the development and adaptation of organisms to changing environmental conditions through the regulation of gene expression [28–30].

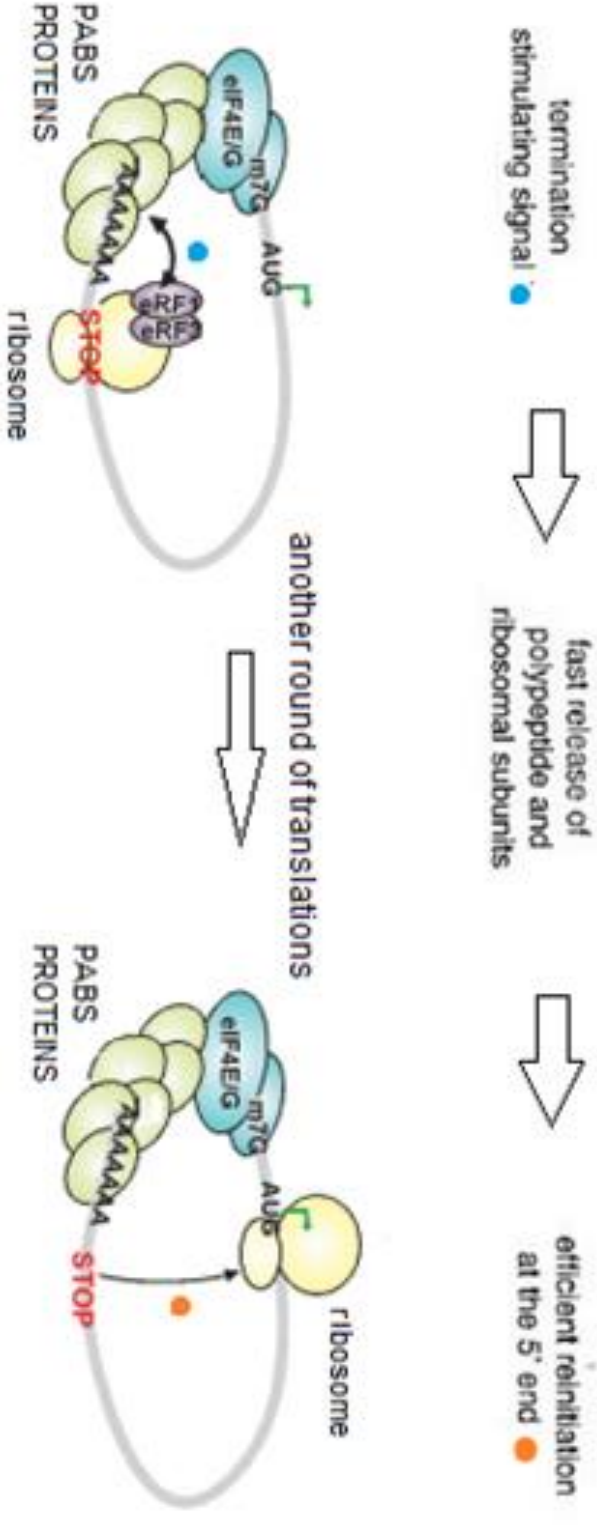
termination stimulating signal ●
fast release of polypeptide and ribosomal subunits
efficient reinitiation at the 5' end ●
another round of translations
PABS PROTEINS
eIF4E/G
m7G
AUG
eRF1
eRF3
AAAAAA
STOP
ribosome
PABS PROTEINS
eIF4E/G
m7G
AUG
AAAAAA
STOP
ribosome

Figure 4. The course of correct translation termination. Proper translation termination depends on a stimulating signal from the poly(A) tail region. Source of image Mühlemann et al (2008) and Sulkowska et al (2017) [11,31].

The mechanism for recognizing mRNA molecules to be degraded through the NMD pathway is not fully understood. The signal to initiate NMD appears to be the stopping of the ribosome at a mislocated stop codon or a translation termination that is too late. Most often, the recognition of transcripts depends on the context of the 3' end (3'UTR, untranslated region), including the poly(A) tail and the poly(A)-binding protein (PABP) that binds this structure, and the course of translation. During translation of normal mRNA, protein factors stabilizing the non-coding end of the 3' mRNA are likely to interact with components of the ribosome retained at the stop codon. This interaction induces conformational changes of ribonucleoprotein molecules that stabilize the transcript and direct it to subsequent rounds of translation. Disruption of this interaction by an elongated non-coding end of the 3'UTR, e.g. in yeast, or in mammals the presence of an intron downstream of the stop codon, results in NMD activation [28,29].

Several versions of the NMD metabolic pathway description are available in the literature. One model of NMD is an exon junction complex (EJC)-based metabolic pathway, **Figure 4**. The EJC is a multiprotein complex that binds to mRNAs of more than 24 base pairs in size, resulting from intron excision. During normal translation, the EJC complex is removed from the mRNA by the ribosome, but when a defective EJC transcript is recognized, it becomes anchored to the mRNA and provides a platform on which the NMD complex is built [11,31].

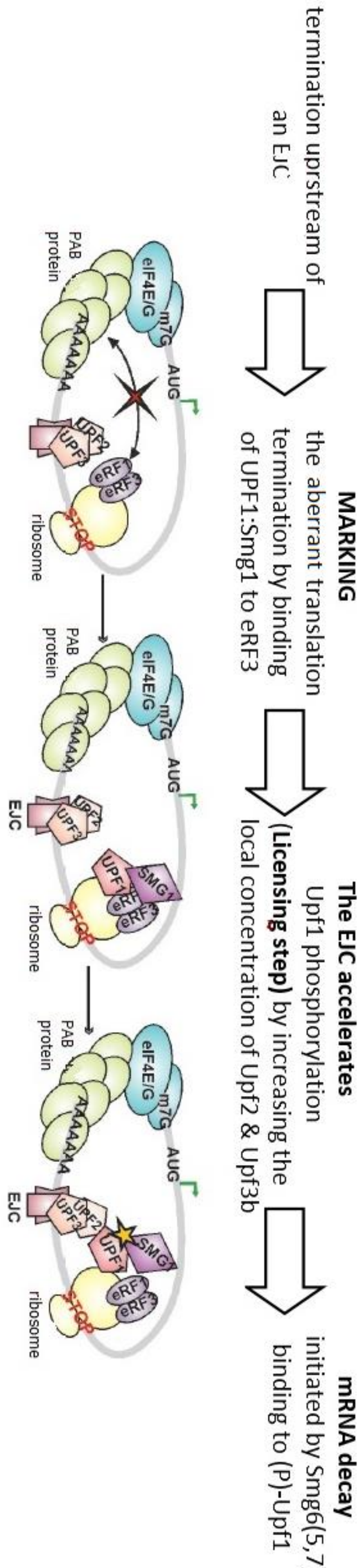
termination uprstream of an EJC
MARKING
the aberrant translation termination by binding of UPF1:Smg1 to eRF3
The EJC accelerates
Upf1 phosphorylation
(Licensing step) by increasing the local concentration of Upf2 & Upf3b
mRNA decay
initiated by Smg6(5,7) binding to (P)-Upf1
eIF4E/G
m7G
AUG
PAB protein
AAAAAAA
UPF2
UPF3
eRF1
eRF3
STOP
ribosome
EJC
SMG1
UPF1

Figure 5. Incorrect termination of translation. Intron based model dependent on EJC. Source of image Mühlemann et al (2008) and Sulkowska et al (2017) [11,31].

In the case of a PTC located 50-55 base pairs upstream of an exon-exon binding site, the ribosome is immobilized together with the translation termination factors eRF1 and eRF3 (eukaryotic release factor), which do not interact with PABP, as occurs during normal translation. As a result of this process, the eRF1 and eRF3 proteins can bind to the SMG1 kinase (suppressors with morphogenetic effects on genitalia) and the UPF1 (up-frame shift) helicase to form the SURF complex (SMG1-eRF1-eRF3-UPF1 complex). The UPF2 and UPF3 proteins present in the EJC complex form a physical connection with SURF, leading to phosphorylation of UPF1 by SMG1 This reaction is a key step and allows the formation of phosphorylated amino acid residues, which are recognized and bound by SMG5-SMG7 proteins through a characteristic 14-3-3 domain. Previous data indicated that the recognition of transcripts from PTCs occurs during the first round of translation, when the mRNA is still attached to the EJC and to the nuclear cap structure-binding complex (CBC). However, more recent studies using mammalian cells have revealed that NMD activation is not only restricted to the first round of translation and also occurs for mRNAs with the cytoplasmic protein eIF4E present on the cap structure. More common in yeast cells, but also present in mammals and in the plant world, a context-dependent model of NMD is known for the 3'UTR sequence, named faux 3'UTR (Figure 5). During the normal course of translation, the cytoplasmic PABPC proteins stabilizing the poly(A) tail interact with the stop codon of the right stop codon of the ribosome and the eRF3 and eRF1 factors to allow translation termination and release of newly formed polypeptides. Excessive distance between the stalled ribosome and PABPC proteins, resulting in a loss of interaction between them provides information about transcript abnormalities triggering NMD activation and subsequently leading to the recruitment of UPF and SMG proteins [11,31].

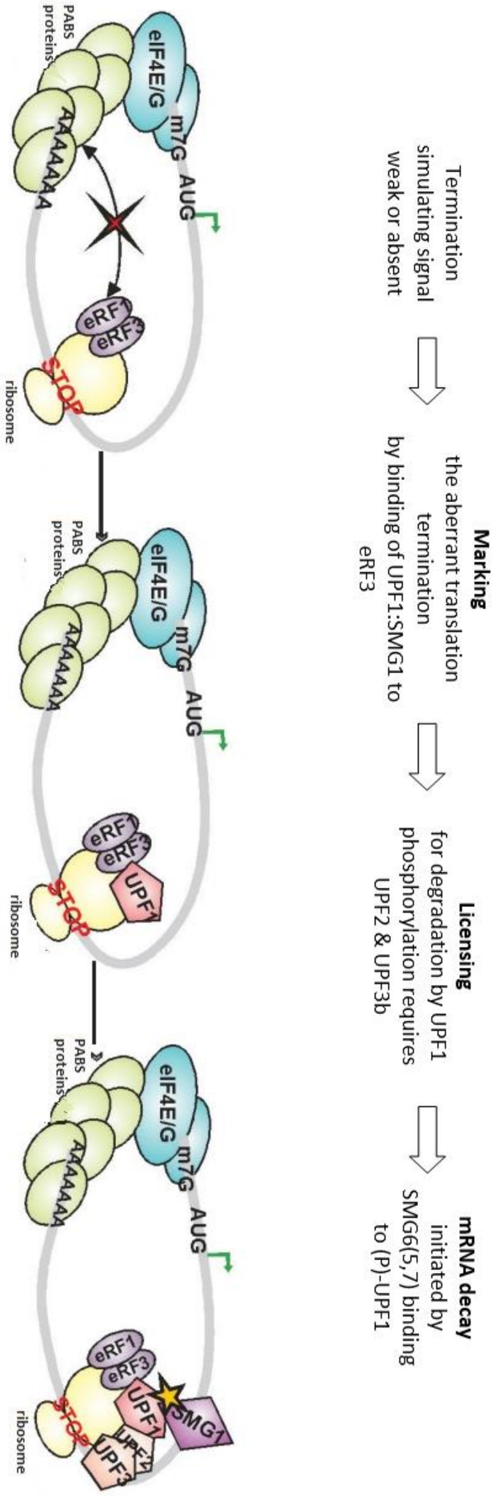
Termination simulating signal weak or absent
Marking the aberrant translation termination by binding of UPF1:SMG1 to eRF3
Licensing for degradation by UPF1 phosphorylation requires UPF2 & UPF3b
mRNA decay initiated by SMG6(5,7) binding to (P)-UPF1
eIF4E/G
m7G
AUG
PABS proteins
AAAAAA
eRF1
eRF3
STOP
ribosome
UPF1
SMG1
UPF2
UPF3

Figure 6. . Incorrect termination of translation. A context-dependent model of NMD based on the 3'UTR sequence. Source of image Mühlemann et al (2008) and Sulkowska et al (2017) [11,31].

1.6 Association methods

Increased availability of high genotyping technology together with advances in DNA sequencing and the development of statistical methodology appropriate for genome-wide association scan mapping is meaningful for the improvement of worldwide healthcare. The progress of new technologies in the field of medical and genetic studies is enabling the utilization of new, advanced methodologically and complex studies. Besides new types of equipment this requires bioinformation technology and in particular biostatistical methodologies and algorithms.

Rapid progress in the development of genomic tools, including genome sequencing and high-density single nucleotide polymorphism (SNP) genotyping has enabled the development of new powerful approaches to the mapping of complex traits and to the subsequent identification of causal genes. The availability of large-scale genomic data has enabled the development of personalized medicine approaches. By analyzing an individual's genetic makeup, theoretically, and sometimes in practice, researchers can tailor treatments to specific patients based on their unique genetic profile. This has the potential to improve patient outcomes and reduce healthcare costs by avoiding ineffective treatments [32].

GWAS (a genome-wide association study) has revolutionized the field of genetics and has helped in identifying genetic variants that are associated with various diseases such as cancer, diabetes, and heart disease. A GWAS enables the testing of hundreds of thousands of genetic variants across many genomes in order to find those statistically associated with a specific trait or disorder (which are then called genomic risk loci or genetic susceptibility loci) [33].The goal of GWASs is to identify associations between genotypes and phenotypes using copy number variants or sequence variations in the human genome (**Figure 7**); however the most frequently used genetic variants are SNPs (single nucleotide polymorphisms). During GWASs testing differences are analyzed in the allele frequency of genetic variants between individuals who are ancestrally similar but distinct according to the phenotypes taken into consideration.

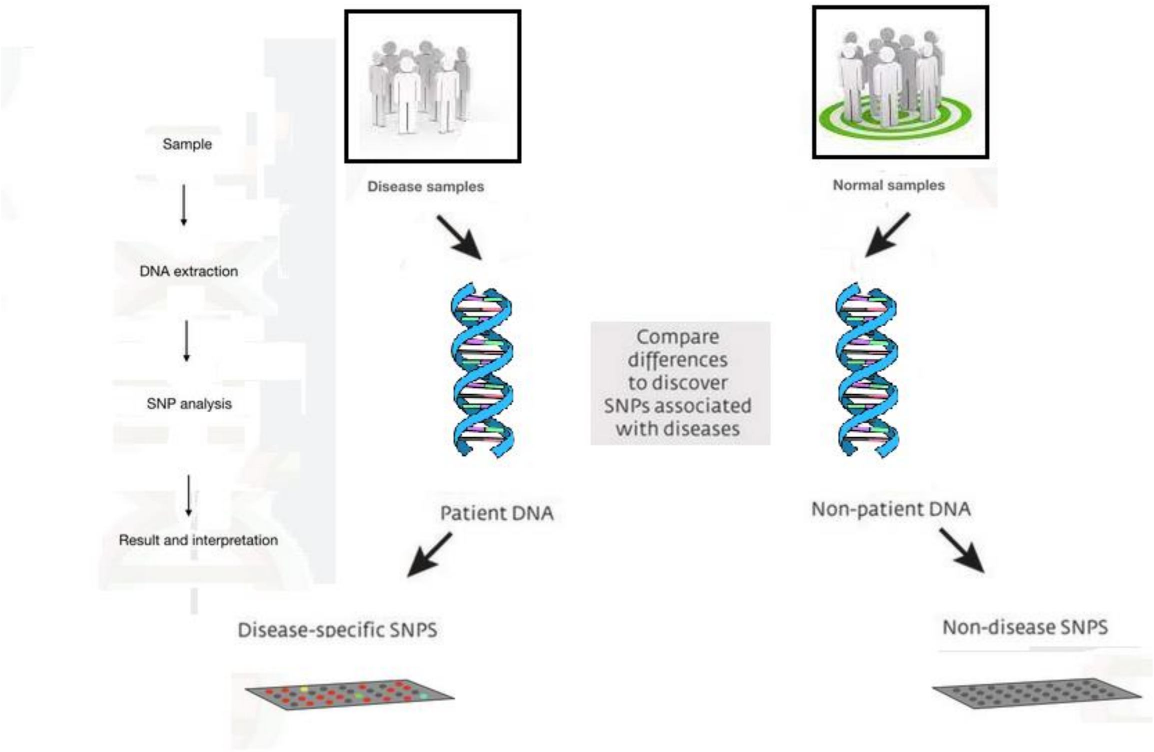

Figure 7. GWAS study. Source of image Palsson et al (2019) [34].

The results obtained thanks to GWASs have a range of applications in many fields including social sciences, ecology, genetics, medicine and more. Collected data can be used for better understanding of a phenotype's biology, estimating its heritability and also calculating genetic correlations or making clinical risk prediction. GWAS results can inform drug development by identifying potential targets for therapeutic interventions, such as for example drug development programs, inferring potential causal relationships between risk factors and associated health outcomes. GWASs have the potential to greatly impact our understanding and treatment of complex diseases. By understanding the genetic basis of a disease, researchers can develop drugs that target specific pathways or proteins involved in the disease process. They can also acquire data from trait-associated genetic variants that can be used as control variables in epidemiology. Further, results can be used to predict an individual's risk for physical and mental disease based on their genetic profile. Finally, results can be used to create biological markers for many diseases and can be used for making a better diagnosis of many diseases [33,35].

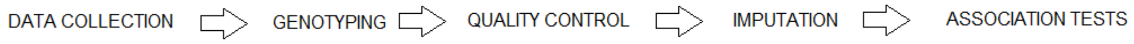

Figure 8. GWAS algorithm

The experimentation for a GWAS is associated with several steps (**Figure 7**, **Figure 8**) and requires proper preparation and tools. The phases of GWAS include:

1) collecting of DNA probes and phenotypic information (age, sex, disease status, etc.)
2) genotyping (GWAS arrays)
3) quality control
4) imputation of untyped variants using haplotype phasing and reference
5) statistical test for association (and, if necessary, a meta-analysis)
6) an independent replication
7) interpretation of the results by conducting multiple post-GWAS analyses

Errors may enter the a GWAS during several steps and therefore carefulness and appropriate tools are required while setting up the examination. To avoid several problems, an error- and bias-standardized quality and analysis protocol is advised to perform a GWAS [33,35].

In February 2023 more than 12 GWAS reports were made with broad scope (**Figure 10**) and were published at the NCBI pages. In recent available data from 2019, the GWAS catalogue contained 157 000 associations and **Figure 9** demonstrates the increasing number of associations from 2006 to 2019.

The present study in this thesis is related to a GWAS (genome-wide association study) as it utilized a database obtained through an agreement with the University of Lodz, which contained 5,600 samples from healthy people in Poland, with 500,000 SNPs, including 141 PTC SNPs. The Lodz study's original objective was to determine, by means of case-control association analysis, how polymorphic risk variants in the FTO/IRXB region affect obesity and/or overweight in the Polish population, and this was successfully published in 2017 [36].

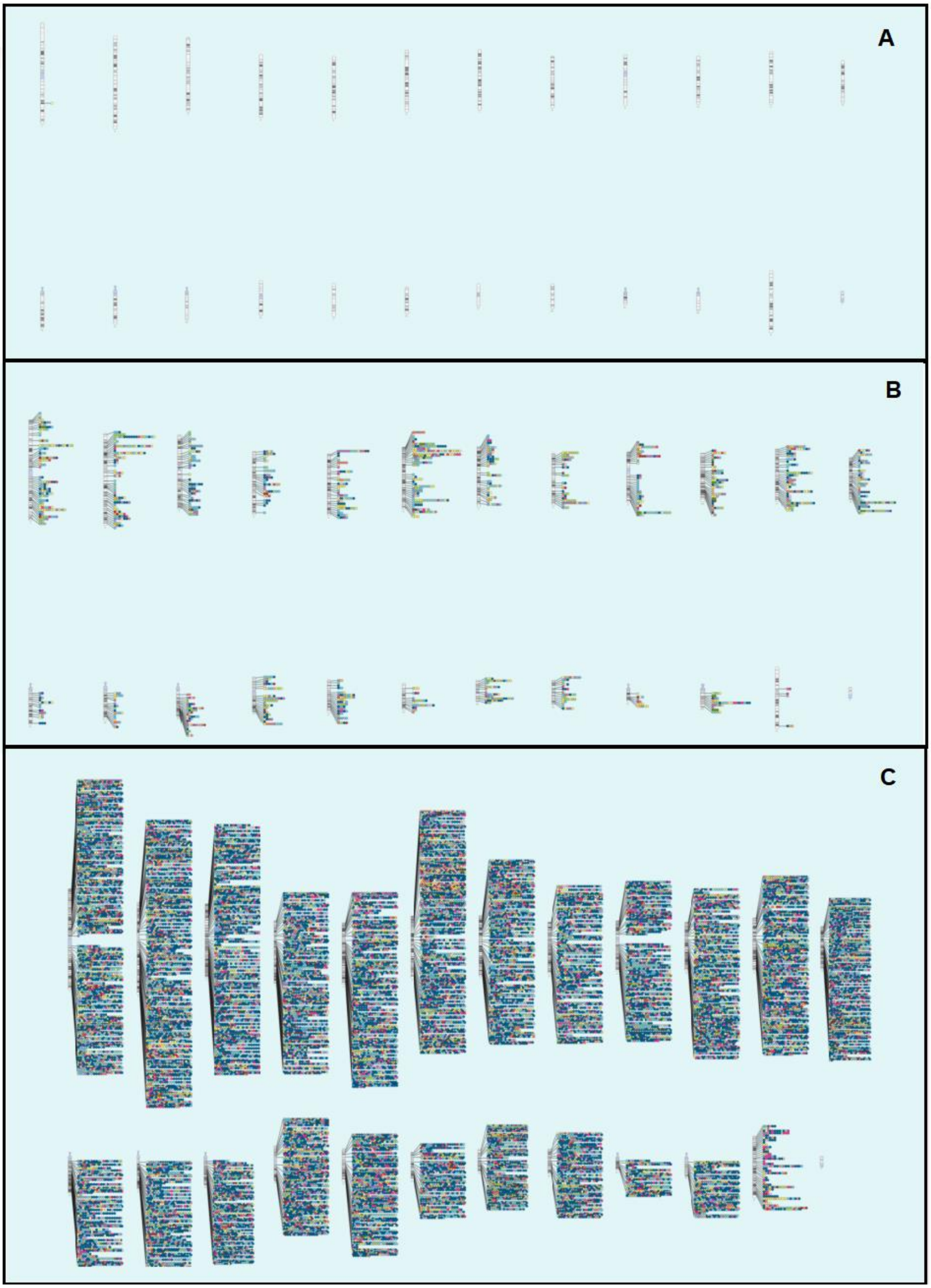

Figure 9. Published GWASs in particular years: A) June 2006, B) June 2011, C) July 2019. Source of image: Buniello et al (2019) [37]

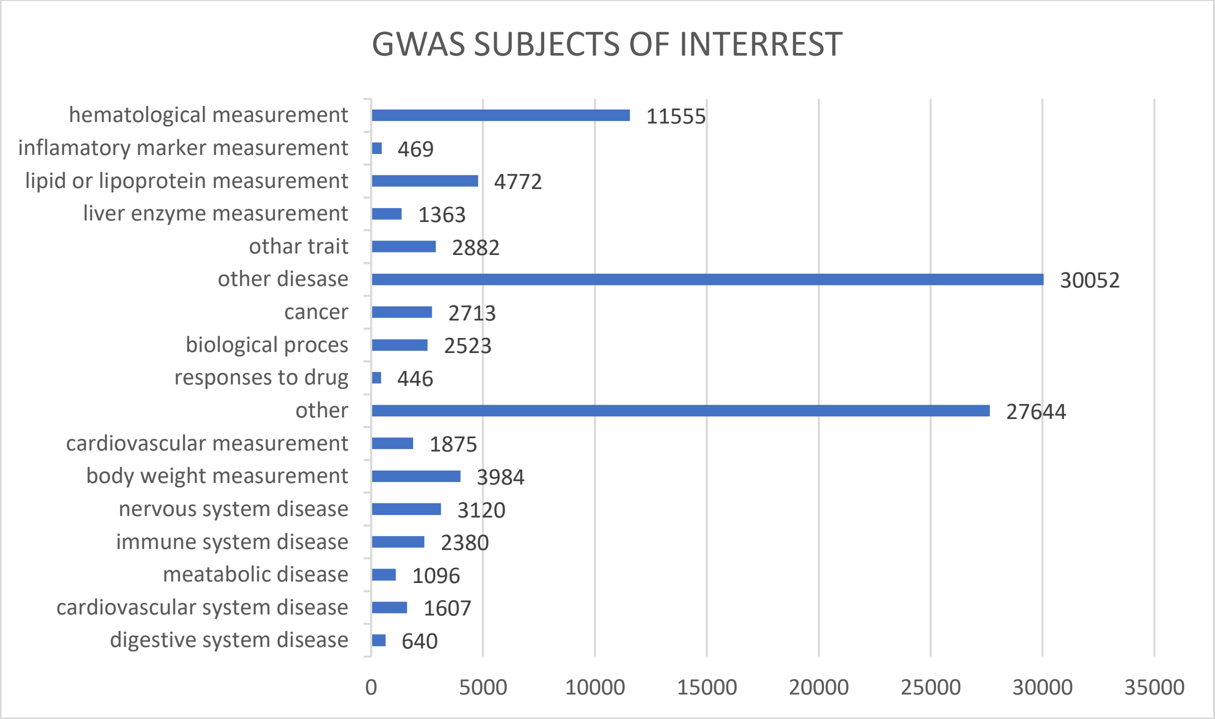

Figure 10. GWAS associations subjects. Source of image: [https://www.ebi.ac.uk/gwas/diagram]

2. Aims and goals

The aims of the present thesis were to perform regressions between genetic models of selected pretermination codon SNPs and several phenotypes (listed below), using data obtained from a large database of 500 000 SNPs each from a large sample with almost 6000 healthy subjects from the entire geographical region of Poland:

a) Analysis of possible associations between pretermination codons and age (life span);

b) Analysis of possible associations between pretermination codons and the number of children (fertility);

c) Analysis of possible associations between pretermination codons and body mass index (obesity and overweight).

3. Methodology

3.1 Sample and data collection

All procedures in section 3.1.1. were performed or coordinated by members of the Biobank Lab of the University of Łódź. All further procedures, from 3.1.2. onwards, were performed by the author of this thesis.

3. 1. 1. Data collection.

The data was collected as a part of the TESTOPLEK study, which was carried out between 2010 and 2012. The data was registered as the "POPULOUS" collection at the Biobank Lab of The Department of Molecular Biophysics of The University of Łódź [36]. A reputable public opinion polling and surveying firm carried out the sampling (SMG/KRC Poland, a Millward Brown subsidiary, Warsaw, Poland). Each participant completed the questionnaire and provided written informed consent. Each person's saliva was put into an Oragene OG-500 DNA collection/storage container (from DNA Genotek, Ontario, Canada). The Review Board of the University of Lodz (KBBN-UL/II/2014) approved the TESTOPLEK study. The latest Helsinki Declaration (World Medical Association Declaration of Helsiniki; www.wma.net), which establishes ethical guidelines for medical research involving humans, was followed.

Initially more than 10 000 people in Poland were surveyed, and from successful saliva and questionnaire collection, an initial study group was formed with a total of 6047 participants. Bone marrow transplantation, diabetes, leukemia, and malignancy were exclusion criteria and participants (n = 488) who disclosed any of these illnesses were excluded: 5559 subjects declared themselves to be in good health and this study group was made up of 2747 men and 2812 women.

DNA material storage and isolation:

The samples of saliva were kept at room temperature until the first processing step. 500 mL of saliva were manually extracted for DNA isolation using the manufacturer's method (PrepitL2P, PD-PR-052,DNA Genotek). The amount of elution was 50 µl. The broad range Quant-iTTMdsDNA Broad Range Assay Kit (InvitrogenTM, Carlsbad, CA, USA) was used to measure the amount of DNA. As part of the Biobank Lab of the Łódź University laboratory's regular protocol for DNA quality control, all DNA samples underwent a PCR reaction to verify their sex [32]. Only DNA samples with the right concentration and purity were used, and these were diluted in sterile DNase-free water to 50 ng/l. The same fluorometric approach was used to recheck the concentration. The standard operating procedures of the Biobank Lab were followed for all laboratory processes relating to sample management.

Genotyping:

According to the manufacturer's recommended procedure, a total of 5559 DNA samples were genotyped for more than 550 000 SNPs using 24x1 InfiniumHTS Human Core Exome BeadChips (Illumina Inc., San Diego, CA, USA). DNA samples were amplified, then fragmented enzymatically and hybridized to the BeadChips. The BeadChips underwent extension and X-staining procedures after that. In another step, iScan was used to scan the BeadChips (Illumina Inc., SanDiego, CA, USA). The Genotyping Module was used to transfer unprocessed fluorescence intensities into GenomeStudio. All of the data underwent rigorous quality control, including sample deletion (n = 141) if the call rate and 10% GenCall parameter fell below 0.94 and 0.40, respectively.

3.1.2. **Genotype quality control:**

Before any files were generated for additional examination, a complete control analysis was performed. Data from SNPs that met all quality control standards were chosen for association studies, and are referred to as "populous data". Further exclusion criteria were: genotyping efficiency lower than 98% and a minimum minor allele frequency (MAF) of 1%. The genotype data were exported from the Genome Studio software using a PLINK INPUT report plug-in into the files .MAP and.PED, which could be easily imported into R and/or PLINK. The manufacturer's website [http://emea.support.illumina.com/downloads/infinium-coreexome-24-v1-1-support-files.html?langsel=/pl/] provides another file type (a compressed .txt file). This file was downloaded and included the names of the Illumina SNPs as well as, if available, the corresponding rs number (produced at the NCBI dbSNP Short Genetic Variations database; https://www.ncbi.nlm.nih.gov/SNP). The Populous database exported a .txt file containing phenotypic information, including the patient's status (healthy (1) or sick (2)), sex (female (1), male (2)), year of birth, district, and the number of children. Age was determined by deducting the birth year from 2012 and phenotype information related to the patient's condition in 2012 was available.

3.2 Statistical analysis

3.2.1 Data preparation

The data file was extracted from the POPULOUS database with a script, kindly provided by Dr. Thierry van de Wetering.

The script used external sources to create a list of Single Nucleotide Polymorphisms (SNPs) to be compared with Illumina data, and assigned base-pair positions according to the Genome Reference Consortium Human Build 37 (GRCh37). The R statistical platform, Rstudio, PLINK, and Notepad++ were used to develop the Analysis Script, which contained R coding and a PLINK coding shell. The script allowed the user to define various parameters, such as whether R warnings should be kept on or off, whether deletions or insertions should be included or excluded from the analysis, and whether subjects with unknown sex should be included or excluded. The script also allowed the user to define the minimum percentage of subjects which have to be genotyped and the years of birth (YoB) groups. The script included a Region Analysis and a SNP Analysis, and PLINK was run within R to analyze the data. Finally, the produced .ped and .map files were imported and processed [38].

The script withdrew the SNPs from POPULOUS database from a .txt file provided by the author of this thesis, which had list of 246 SNPs that lead to the introduction of a PTC, identified by Fujikura et al. [22]. A total of 141 SNPs were found in the POPULOUS database. In the next stage the genotype data was merged with available metadata (age, sex, district , BMI, NCI). The patients with all phenotype data missing were removed, leaving a total of 5095 subjects for analysis.

BMI categorization:

The patients were categorized using WHO (World Health Organization) BMI (Body Mass Index) classification criteria for adults:

BMI below 18.5 – Underweight, BMI between 18.5 and 24.9 - Normal weight, BMI between 25 and 29.9 – Overweight, BMI greater/equal 30 – Obesity.

NCI subject subgroup

Previous researchers have suggested that an association analysis with numbers of children (NCI) should be conducted in a group who had completed their reproductive period. Barban et al. (2016) defined this as : age ≥ 45 women; age ≥ 55 men [39]. The research of Barban et al. (2016) involved a genome-wide association study (GWAS) of reproductive behavior in over 250,000 individuals of European ancestry. The study found 12 loci that were significantly associated with reproductive behavior, including age at first birth and number of children. A second dataset, with the NCI subject subgroup, using the Barban et al. age criteria, was therefore created.

3.2.2 The graphical representation of PTC locations on chromosomes

The graphical presentation of PTC's location on chromosomes was performed using the R [chromoMap] package.

The R chromoMap package provides interactive and customizable visualization of chromosomes or chromosomal regions, allowing users to map chromosome features and associated data. The package allows for visualizing polyploidy, creating high-resolution maps, annotating groups of elements with distinct colors, and visualizing multi-omics data using a variety of plot types. Other features include creating 2D chromosome plots, visualizing correlations between genomic features, adjusting chromosome range or visualizing specific regions such as genes/SNPs, adding labels and hyperlinks to the plot, and saving plots as HTML documents. The package documentation is also available as an R vignette.

The function required 2 data input files; chromosome name and length file and annotation (SNP) chromosomal position file.

All data were obtained from NCBI databases [https://ncbi.nlm.nih.gov/] .

Coding:

```
A<- read.table('chrdata.txt')
B<-read.table('annot.txt', sep="\t")
chromoMap(list(A),list(B), chr_color = 'gray', anno_col = "black", labels= "T",
        label_angle = -65,
        ch_gap= 9,
        chr_length = 10,
        chr_width = 45,
        canvas_width = 1150,label_font= 12,
        segment_annotation=T,
        text_font_size=c(15))
dev.off()
```

3.2.3 Standardized effect sizes

The standardized effect sizes used in this thesis were: Cohen’s d, Spearman’s r and odds ratio. Standardized effect size is a statistical measure that quantifies the magnitude of a difference or relationship between two groups or variables relative to measured variance (in the discussion below it is referred to simply as “effect size”). It provides information on the practical or clinical significance of the difference or relationship, which is often not apparent from p-values or statistical significance alone [40].

The effect size used to measure the difference in means between two quantitative variables in this thesis was **Cohen’s d**. Cohen's d is a widely used measure of standardized effect size, typically used to quantify the difference between two means. It is calculated by dividing the difference between the means of two groups by the pooled standard deviation of the two groups. Cohen's d is expressed in units of standard deviation, which makes it easy to interpret and compare across studies. A small Cohen's d effect size is typically considered to be 0.2, a medium effect size is 0.5, and a large effect size is 0.8 or greater. It was further expanded in Sawilowsky’s work to: very small 0.01, very large 1.2 and huge 2.0 [41]. Cohen's d is a useful tool in meta-analyses and in determining the practical significance of statistical findings. It provides a standardized measure of the difference between two groups, allowing researchers to compare standardized effect sizes across studies and to draw more general conclusions about the strength of an effect [42].

The effect size used to measure the strength of correlation in this thesis was **Spearman’s “r”**. The "r" effect size refers specifically to the correlation coefficient, which measures the strength and direction of the linear relationship between two variables. The value of r ranges from -1 to +1, with -1 indicating a perfect negative correlation, 0 indicating no correlation, and +1 indicating a perfect positive correlation. Cohen (1988) proposed a widely used classification system for interpreting the magnitude of r values. According to this system, an r value of 0.1 indicates a small effect size, 0.3 indicates a medium effect size, and 0.5 indicates a large effect size. However, it should be noted that the interpretation of effect sizes may vary depending on the context and research question. Overall, the r effect size is a useful statistical measure that provides information on the strength and direction of the linear relationship between two variables. [40,42,43].

The effect size used to measure the risk of obesity as a qualitative variable in this thesis was the odds ratio. The odds ratio (OR) is a statistical measure used to quantify the strength of the association between two variables in a case-control or cohort study. It represents the ratio of the odds of an event occurring in one group compared to the odds of the event occurring in another group. An OR of 1 indicates no association between the variables, while an OR greater than 1 indicates a positive association and an OR less than 1 indicates a negative association. The odds ratio is a useful tool in medical research and epidemiology, allowing researchers to determine the risk of a disease or condition in a certain group compared to another group. It is particularly useful in case-control studies, where the odds of exposure to a risk factor can be compared between cases and controls [44].

3.2.4 Adjustor analyses

Adjustor analysis was performed using the R [smplot] package [45]. The smplot package provides several functions for creating scatterplots with smoothed trend lines, including *sm.scatterplot* and *sm.densityplot*. These functions allow the user to specify various parameters such as the type of smoothing algorithm to use, the degree of smoothing, and the colors and shapes of the data points. The graphs created with this package include Spearman’s rho correlation coefficient value and corresponding p.value.

All statistical analyses were two-tailed and statistical significance level was set to p.value < 0.05, with or without Bonferroni correction.

3.2.5 Association analyses

All statistical association analyses for this thesis were performed using the R statistical platform (version 3.4; https://cran.r-project.org) [46]

For association analysis with quantitative variables, linear and logistic regression models, implemented in the R SNPassoc package, were used. All statistical tests were two-tailed and two statistical significance p.value thresholds were set, nominal p < 0.05 and after Bonferroni correction < (0.05/141 SNPs analyzed) = 3.55 x 10^{-4}.

SNPassoc is an R package designed for genetic association studies. It provides a suite of tools for performing various types of genetic association analyses, including single SNP tests, haplotype analysis, gene-based tests, and rare variant analysis. The package also includes functions for quality control and data cleaning of genetic data.

Some of the key features of SNPassoc include:

- Single SNP analysis: SNPassoc provides several tests for single SNP analysis, including chi-square tests, logistic regression, and linear regression. These tests can be performed on both binary and continuous traits.
- Haplotype analysis: SNPassoc allows users to perform haplotype analysis using several methods, including haplo.stats, THESIAS, and PLINK.
- Gene-based analysis: SNPassoc provides several gene-based tests, including VEGAS, GATES, and GRASS. These tests can be used to identify genes that are associated with a trait of interest.
- Rare variant analysis: SNPassoc includes functions for performing rare variant analysis using several methods, including SKAT, SKAT-O, and burden tests.
- Quality control: SNPassoc includes functions for performing quality control and data cleaning of genetic data, such as checking for missing data, checking for genotyping errors, and checking for population stratification.

Overall, SNPassoc is a powerful and flexible package for performing genetic association analyses in R. It provides a wide range of tools for analyzing genetic data, making it a useful tool for researchers in the field of genetic epidemiology [47].

Linear regression is a commonly used model in genetic association studies, as it can be used to test for association between a single genetic variant and a phenotype [48–51]. The association analysis results for 5 genetic models (dominant, recessive, overdominant, log-additive) were tabulated (available as supplementary tables on CD), and a Manhattan type plot was generated with the use of that data.

A Manhattan plot is a type of plot that is commonly used in genome-wide association studies (GWAS) to visualize the results of SNP association tests across the genome. The plot shows the negative log10 transformed p-values on the y-axis and the SNP name x-axis. Additionally two horizontal lines, representing p.value thresholds were added ($-\log_{10}(0.05) = 1.3$, $-\log_{10}(3.55 * 10^{-4}) = 3.45$). Potentially significant clinical results were highlighted as blue, if the p.value passed nominal p.value threshold and red if the p.value passed Bonferroni correction p.value threshold. The results that passed at least nominal p.value threshold were tabulated, sample size, mean, standard errors, the mean difference and its 95% confidence interval compared to the most frequent homozygous genotype, p-value for an overall gene effect and the Akaike Information Criterion (AIC) for each genetic model were presented.

Akaike Information Criterion (AIC) is a statistical measure used to evaluate the relative goodness of fit of different models to a given set of data. The AIC is based on the principle of maximum likelihood, which seeks to find the model parameters that maximize the likelihood of the observed data. AIC estimates the amount of information lost when a given model is used to approximate the true data-generating process, with the aim of identifying the model that strikes the best balance between goodness of fit and parsimony (i.e., simplicity). According to Burnham and Anderson (2002), AIC is defined as follows [52]:

$AIC = -2 \log(L) + 2k$

where L is the maximized likelihood of the model, and k is the number of parameters in the model. The first term, -2 log(L), is a measure of the goodness of fit of the model, while the second term, 2k, penalizes the model for having too many parameters. The model with the lowest AIC value is considered to be the best model for the given data. AIC has several advantages over other model selection criteria, such as the Bayesian Information Criterion (BIC), including its ability to handle small sample sizes and its focus on predictive accuracy rather than parameter estimation. However, it also has some limitations, such as its reliance on asymptotic theory and the fact that it assumes the true model is among the candidate models being compared [53].

The main genetic models thought to be most relevant to potential clinical significance that were taken under consideration in this study were mostly: the dominant model[54] and the recessive model[55–57]. There are no reports of PTC inheritance with overdominant or additive genetic models.

Coding used for the regression models was:

```
ANALYSIS <-setupSNP(DATA_SET,colSNPs = c(13:ncol(DATA)),sep="")

#

TABLE_PHENOTYPE<-WGassociation(PHENOTYPE(continuous
variable)~Adjustor1+Adjustor2…,data=ANALYSIS,genotypingRate = 0.1)
```

e.g

```
TABLE_BMI<-WGassociation(BMI~YoB+GENDER,data=ANALYSIS,genotypingRate = 0.1)
```

BMI category – analysis

All statistical association analyses were performed in an R environment. For the qualitative variable of BMI categories versus SNP for dominant and recessive models, the Fisher exact test and Cochran Armitage trend test were used. Statistical significance level was set to $p < 0.05$.

4. RESULTS

4.1 Group characteristics

4.1.1 Summary of the group

From the initial 5559 healthy subjects, patients with missing data (in age, BMI and NCI) were removed, giving a study group with a total of 5095 subjects from all 16 districts in Poland.

Table 3 shows that 2605 (51.1%) subjects were female and 2490 (48.9%) male. The average age was 42.4 years old (sd =14.8), the number of children per individual was 1.58 (sd= 1.52) and mean body mass index (BMI) was 25.3 (sd = 4.55). The subjects were divided into four BMI groups: 185 (3.63%) subjects were qualified to the underweight group, 2472 (48.5%) to the normal weight group, 1698 (33.3%) to the overweight group and 742 (14.6%) to the obesity group. The qualification criteria for the BMI groups are described in the methodology Data preparation section. There were 45 subjects with missing data of patient's age and 1 subject without number of children.

In total 141 SNPs that lead to nonsense mutations were found and data for these was drawn from the POPULOUS database.

Table 3.Subject Characteristics

Variable	Number of subjects with data available (N)	Number of subjects in sub-groups, n (%)	Mean (s.d.)
Total number of subjects	5095		
SEX: Female		2605 (51.1%)	
Male		2490 (48.9%)	
Age	5050		42.4 (14.8)
Number of children per individual	5094		1.58 (1.52)
Body Mass Index (BMI)	5095		25.3 (4.55)
BMI GROUP:	5095		
Underweight		185 (3.63%)	
Normal weight		2472 (48.5%)	
Overweight		1696 (33.3%)	
Obesity		742 (14.6%)	

N- sample size; for qualitative variable n (%) of the group is presented, for quantitative mean and (standard deviation).

4.1.2 Adjustors

An analysis of correlations between the variables representing the analyzed phenotypes was conducted in order to determine any intercorrelation and possible interference on later obtained results.

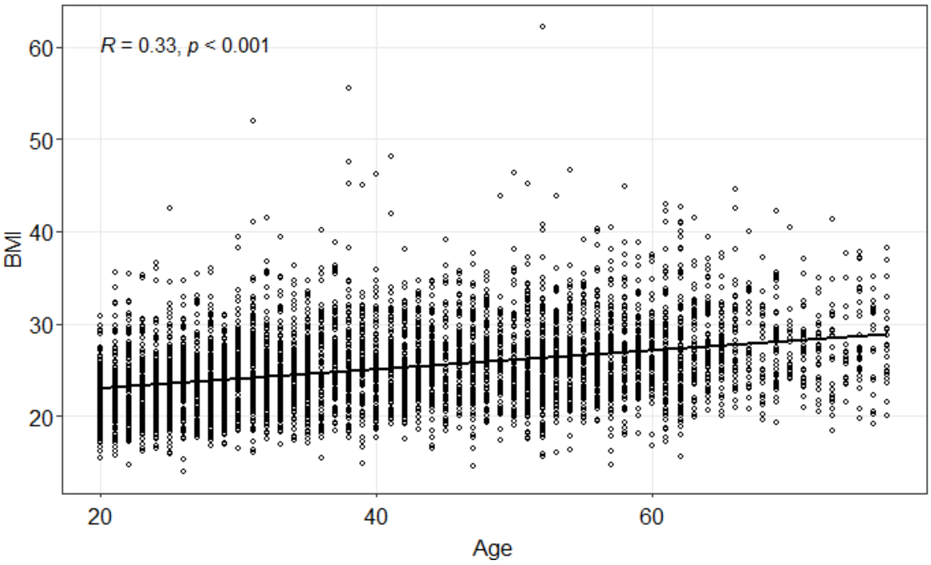

Figure 11. Comparison of age to body mass index in the Polish population. Line drawn from linear model approximation (not used in statistical analysis).

A correlation of medium (r = 0.33, p< 0.05) effect size was detected between patient's age and body mass index, **Figure 11**. Additionally, upon separate analyses in females and males, trends with different effects were observed **Figure 12**.

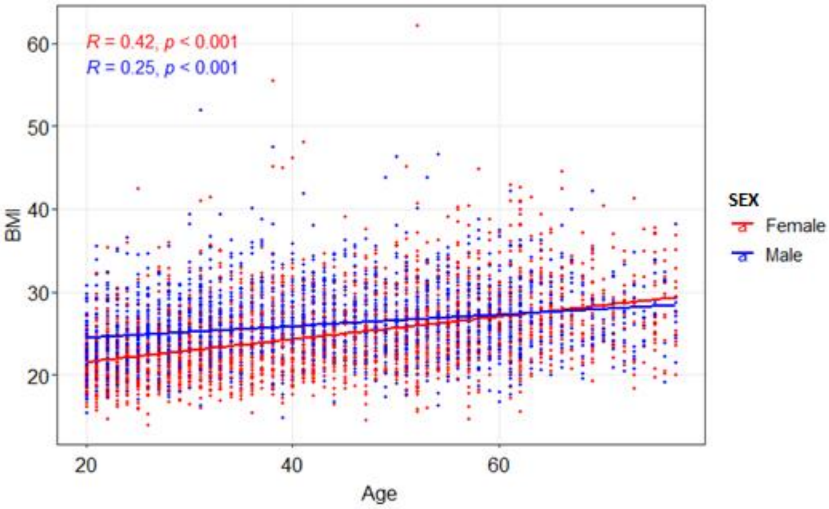

Figure 12. Comparison of age to body mass index in female and male groups. Lines drawn from linear model approximations (not used in statistical analysis).

The mean BMI value in the female group was 24.67 (sd= 4.86) kg/m^2 and 26.00 (sd = 4.09) kg/m^2 in the male group. The difference in means (female minus male data) was -1.32 kg/m^2 (95% CI: -1.57 lower, -1.07 upper, p.value < 0.05) **Table 4**.

Table 4. Summary of body mass indices in the female and male populations (all values in units of kg/m^2)

Sex	n	mean	sd	median	min	max	range	se	T test statistics for difference in means
Female	2605	24.67	4.86	23.74	14.01	62.09	48.09	0.10	diff 95% CI : -1.57 -1.08 t = -10.55, p-value < 2.2e-16
Male	2490	26.00	4.09	25.61	14.85	51.93	37.08	0.08	

n- sample size; sd- standard deviation; se – standard error; CI- confidence intervals lower-upper

The total population was not homogeneous due to differences between the groups. The effect of age on BMI was stronger in the female group (r= 0.42) than in the male group (r = 0.25) (**Figure 12).** Additionally, the difference in mean values lies on the edge of a BMI group qualification criterion, which might impact the qualitative analysis. following these discoveries, it was determined to perform the genotype association analyses with BMI using adjustments for age and sex.

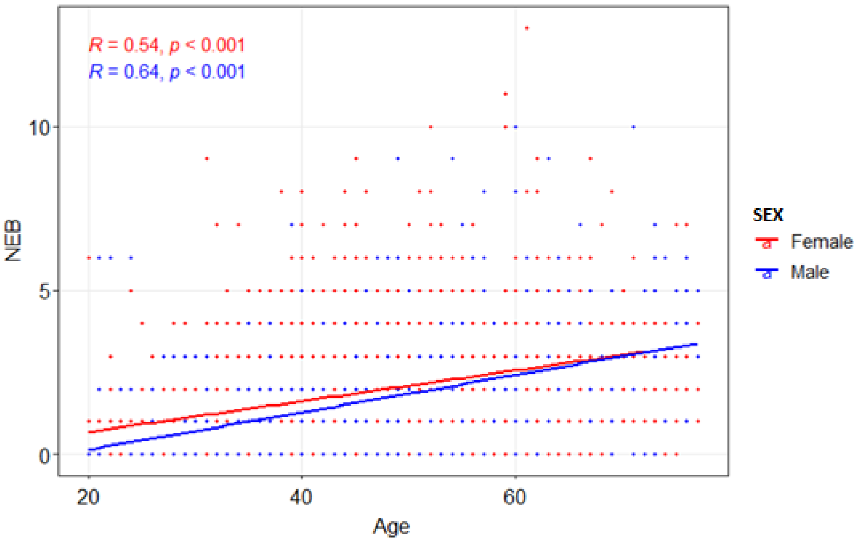

Figure 13. Comparison of age to number of children born (NCI) in female and male groups. Lines drawn from linear model approximations (not used in statistical analysis).

Another dependency was graphically presented in **Figure 13.** There was strong (r > 0.5) correlation between the subject's age and number of children born (NCI) per individual, r = 0.54 for females and r= 0.64 for males. The mean value of NCI in Female group was 1.75 (sd= 1.54) and 1.4 (sd = 1.47) in male group **Table 5**.

Table 5. Summary of number of children per individual (NCI) in the female and male populations

Sex	n	mean	sd	median	min	max	range	se	T test statistics
Female	2604	1.75	1.54	2	0	13	13	0.03	diff 95% CI : 0.27 0.44 t = 8.36 p-value < 2.2e-16
Male	2490	1.40	1.47	1	0	10	10	0.03	

n- sample size; sd- standard deviation; se – standard error; CI- confidence intervals lower-upper

The difference in means of number of children born was significantly different across districts **Table 6**. The lowest mean of number of children born was in Opolskie district, 1.08 children (sd=1.19), the highest was in Dolnośląskie district 1.76 children (sd=1.69).

Table 6. Comparison of number of children born in different districts

District	n	mean	sd	median	min	max	range	se	Kruskal Wallis test statistics:
Dolnośląskie	196	1.76	1.69	1	0	9	9	0.12	Kruskal-Wallis chi-squared = 34.11, df = 15, p-value = 0.0033
Kujawsko-Pomorskie	286	1.7	1.52	2	0	9	9	0.09	
Łódzkie	373	1.72	1.67	2	0	13	13	0.09	
Lubelskie	194	1.69	1.66	2	0	10	10	0.12	
Lubuskie	154	1.64	1.47	2	0	7	7	0.12	
Małopolskie	115	1.59	1.59	1	0	6	6	0.15	
Mazowieckie	462	1.56	1.46	1.5	0	8	8	0.07	
Opolskie	196	1.08	1.19	1	0	5	5	0.08	
Podkarpackie	397	1.68	1.71	1	0	10	10	0.09	
Podlaskie	231	1.63	1.41	2	0	6	6	0.09	
Pomorskie	396	1.46	1.37	1	0	8	8	0.07	
Śląskie	922	1.57	1.51	1	0	10	10	0.05	
Świętokrzyskie	79	1.66	1.42	2	0	6	6	0.16	
Warmińsko-Mazurskie	254	1.48	1.46	1	0	7	7	0.09	
Wielkopolskie	542	1.60	1.53	2	0	10	10	0.07	
Zachodniopomorskie	250	1.46	1.39	1	0	8	8	0.09	

Following this, it was decided to perform genotype association analysis with fertility (NCI) with age, sex and district adjustments.

4.1.3 SNP data quality

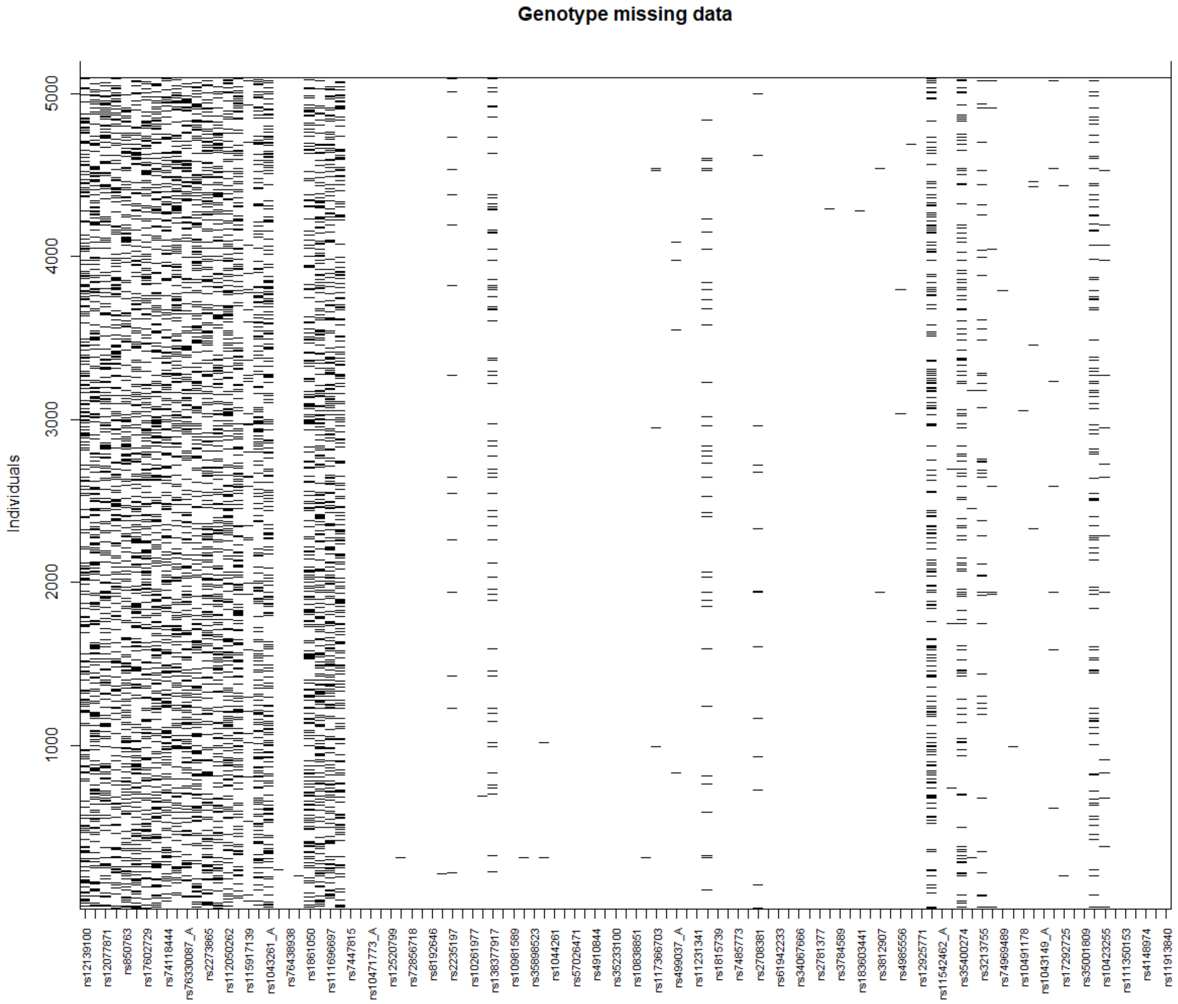

Figure 14. Missing genotype readings for selected RSs

Data quality is presented graphically, **Figure 14** and tabulated as a percentage of missing genotype readings for individual RSs in Table 5. SNPs missing the most data were: rs36078704 (66.1%, *DDX49*), rs76330087 (26.2%, *ATP6V1G3*) and rs62154921 (25.9%,*VWA3B*)**, Table 7**. 72 RSs did not have any missing readings.

Table 7. Summary of missing genotype readings data for selected RS

RS number	Missing % of genotype	RS number	Missing % of genotype	RS number	Missing % of genotype
rs34291832	0	rs76101114	0	rs4788587	0.2
rs4639011	0	rs34960436	0	rs118004742	0.2
rs3732781	0	rs78283108	0	rs117366703	0.3
rs79892855	0	rs34931752	0	rs71377306	0.3
rs76438938	0	rs3784589	0	rs545652	0.4
rs74437357	0	rs57809907	0	rs8072510	0.5
rs7447815	0	rs28413581	0	rs28602966	0.6
rs1023840	0	rs150843673	0	rs499037	0.7
rs10471773	0	rs36102575	0	rs2270416	0.7
rs35391433	0	rs13338754	0	rs41291550	0.8
rs12520799	0	rs11542462	0	rs2233919	0.8
rs17184009	0	rs34381648	0	rs55727303	0.9
rs3130453	0	rs74969489	0	rs10423255	1.5
rs6907580	0	rs1043149	0	rs2708381	1.7
rs34672740	0	rs35699176	0	rs61753375	1.7
rs10237332	0	rs61751875	0	rs2235197	2.5
rs67047829	0	rs17001893	0	rs115917139	3.5
rs2293766	0	rs35001809	0	rs61750839	4.6
rs10261977	0	rs61737751	0	rs1790218	4.9
rs328	0	rs111350153	0	rs3213755	5.2
rs117752382	0	rs41282820	0	rs138377917	7.5
rs2039381	0	rs4148974	0	rs541169	11.8
rs10981589	0	rs28502153	0	rs35400274	12.1
rs45579335	0	rs35032582	0	rs7499011	20.6
rs1476860	0	rs62239058	0	rs114730569	24.5
rs35898523	0	rs75411676	0.1	rs1043261	24.5
rs1044261	0	rs72856718	0.1	rs6671527	24.6
rs7904983	0	rs45621032	0.1	rs863362	24.6
rs57026471	0	rs8192646	0.1	rs5744168	24.6
rs2647574	0	rs34427887	0.1	rs112050262	24.7
rs16930998	0	rs9886752	0.1	rs114429815	24.8
rs4910844	0	rs12240276	0.1	rs116389032	24.9
rs61730422	0	rs1815739	0.1	rs74118444	24.9
rs35233100	0	rs61942233	0.1	rs2273865	25
rs7120775	0	rs80072371	0.1	rs12077871	25.1
rs10838851	0	rs41281112	0.1	rs850763	25.1
rs1459101	0	rs2781377	0.1	rs1861050	25.1

rs11228710	0	rs183603441	0.1	rs61731313	25.1
rs2298553	0	rs3812907	0.1	rs111696697	25.2
rs11231341	0	rs12925771	0.1	rs112033303	25.2
rs77002186	0	rs10491178	0.1	rs12568784	25.4
rs35231465	0	rs17292725	0.1	rs12139100	25.5
rs497116	0	rs74830030	0.1	rs2176186	25.6
rs7485773	0	rs11913840	0.1	rs62154921	25.7
rs16910526	0	rs2272754	0.2	rs76330087	25.9
rs146753414	0	rs11071990	0.2	rs17602729	26.2
rs34067666	0	rs4985556	0.2	rs36078704	66.1

4.1.4 Chromosomal localization of the SNPs

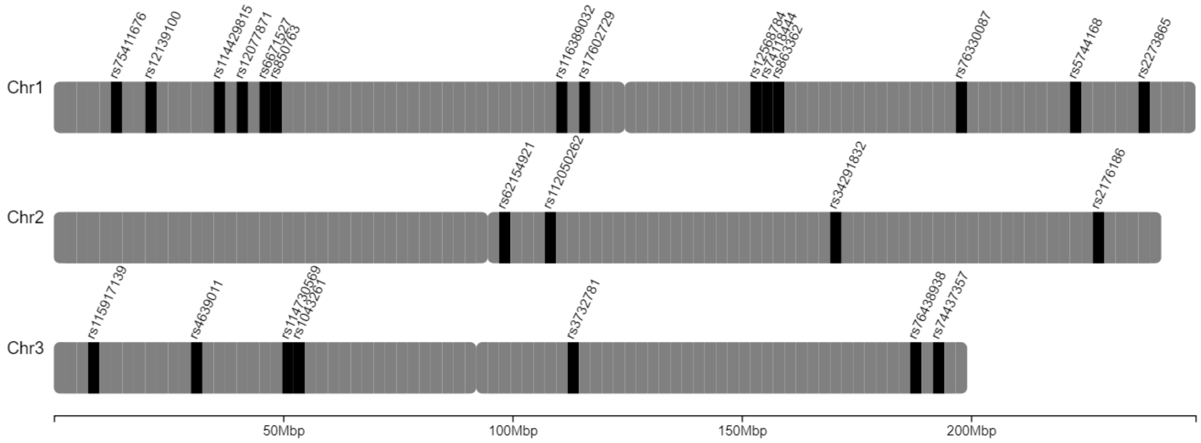

Figure 15. Localization of the SNPs on chromosomes 1,2,3

There were 14 PTC SNPs which were located on chromosome 1 (**Figure 15**):

rs75411676 Chr1:12919891; rs12139100 Chr1:20501582; rs114429815 Chr1:35227093;
rs12077871 Chr1:40773150; rs6671527 Chr1:47080679; rs850763 Chr1:48708228;
rs116389032 Chr1:109823457; rs17602729 Chr1:115236057; rs12568784 Chr1:152323132;
rs74118444 Chr1:155291263; rs863362 Chr1:158549492; rs76330087 Chr1:198505831;
rs5744168 Chr1:223285200; rs2273865 Chr1:236706300.

There were 4 PTC SNPs located on chromosome 2:

rs62154921 Chr2:98779439, rs112050262 Chr2:108863758, rs34291832 Chr2:170387886,
rs2176186 Chr2:228476140.

There were 8 PTC SNPs located on chromosome 3:

rs115917139 Chr3:9874914, rs4639011 Chr3:32030998, rs114730569 Chr3:52005638,
rs1043261 Chr3:53899276, rs3732781 Chr3:113955187, rs79892855 Chr3:113955726,
rs76438938 Chr3:186461524, rs74437357 Chr3:193052769.

Location of the rs79892855 overlaps with graphical grid of rs3732781, therefore only the second is presented in **Figure 15** .

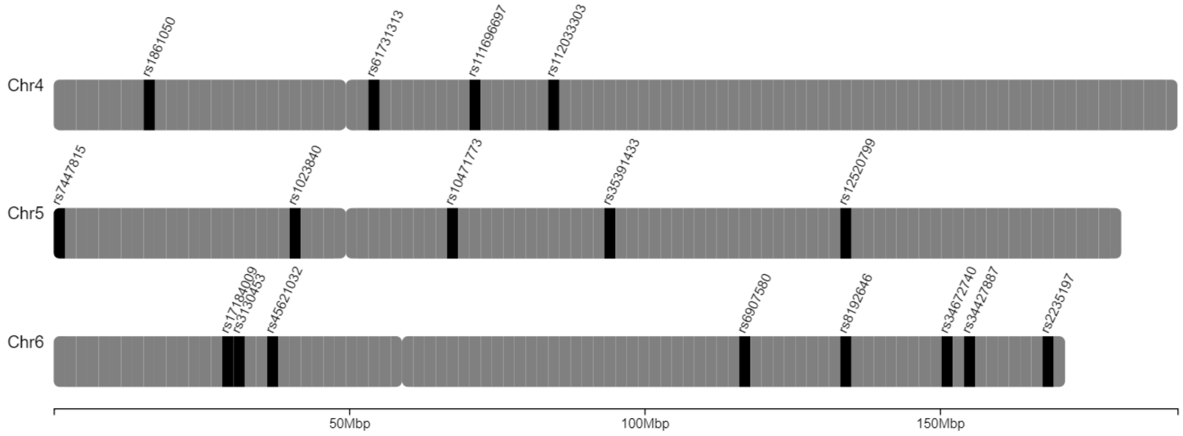

Figure 16. Localization of the SNPs on chromosomes 4,5,6

There were 4 SNPs located on chromosome 4 (**Figure 16**):

rs1861050 Chr4:15482360, rs61731313 Chr4:53611484, rs111696697 Chr4:70512787,
rs112033303 Chr4:84206004.

There were 5 SNPs located on chromosome 5:

rs7447815 Chr5:1240757, rs1023840 Chr5:41061715, rs10471773 Chr5:68616079,
rs35391433 Chr5:94749787, rs12520799 Chr5:134782450.

There were 9 SNPs located on chromosome 6:

rs17184009 Chr6:29407955, rs3130453 Chr6:31124849, rs72856718 Chr6:31125257,
rs45621032 Chr6:36274148, rs6907580 Chr6:117150008, rs8192646 Chr6:132938842,
rs34672740 Chr6:150387059, rs34427887 Chr6:154567863, rs2235197 Chr6:167709702.

Location of rs72856718 overlaps with graphical grid of rs3130453, therefore only the second is presented in **Figure 16**.

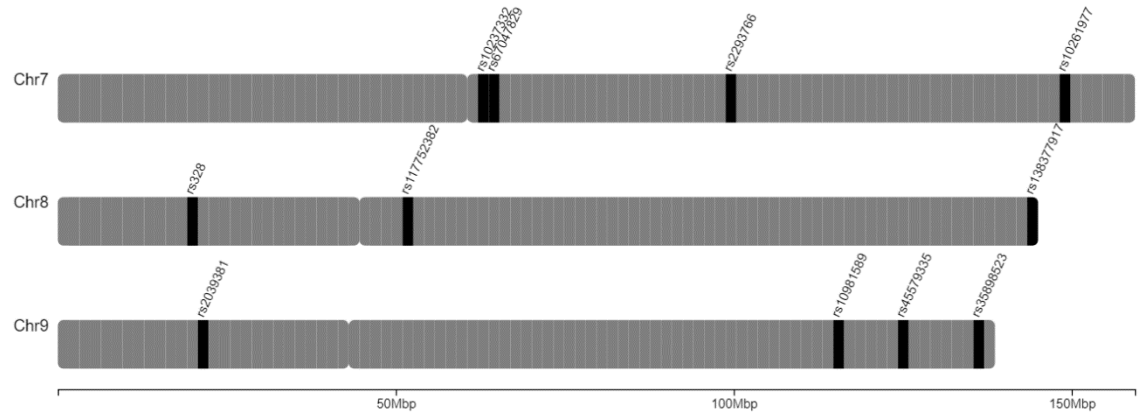

Figure 17. Localization of the SNPs on chromosomes 7,8,9

There were 4 SNP located on chromosome 7 (**Figure 17**):

rs10237332 Chr7:63529269, rs67047829 Chr7:64452738, rs2293766 Chr7:100371358,
rs10261977 Chr7:149528262.

There were 4 SNPs located on chromosome 8:

rs328 Chr8:19819724, rs117752382 Chr8:52284560, rs138377917 Chr8:143763531,
rs2272754 Chr8:144522387.

Location of rs2272754 overlaps with graphical grid of rs138377917, therefore only 2nd is presented on **Figure 17**.

There were 7 SNPs located on chromosome 9:

rs2039381 Chr9:21481483, rs10981589 Chr9:115759519, rs45579335 Chr9:125239501,
rs1476860 Chr9:125391241, rs35898523 Chr9:136029645, rs55727303 Chr9:136131576,
rs9886752 Chr9:139634495.

Location of rs1476860 overlaps with graphical grid of rs45579335, therefore only 2nd is presented on **Figure 17**. Rs9886752 and rs55727303 location overlaps with graphical grid of rs35898523, therefore only last one is presented on **Figure 17**.

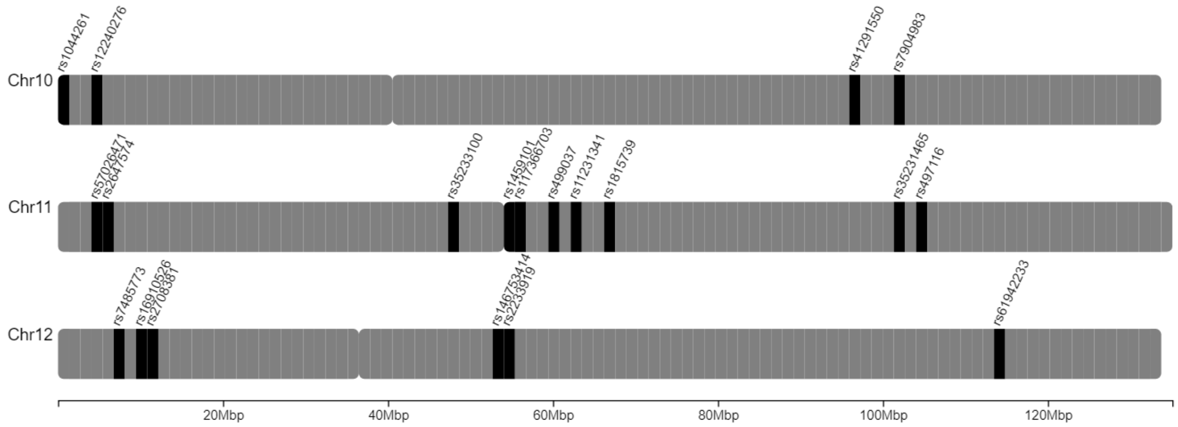

Figure 18. Localization of the SNPs on chromosomes 10,11,12

There were 4 SNPs located on chromosome 10 (**Figure 18**) :

rs1044261 Chr10:1065710, rs12240276 Chr10:4889403, rs41291550 Chr10:96447562,
rs7904983 Chr10:102056016.

There were 19 SNPs located on chromosome 11:

rs57026471 Chr11:5068662, rs2647574 Chr11:5444136, rs16930998 Chr11:5462702,
rs4910844 Chr11:5776484, rs61730422 Chr11:5989223, rs35233100 Chr11:47306630,
rs7120775 Chr11:48266736, rs10838851 Chr11:48286231, rs1459101 Chr11:55339652,
rs117366703 Chr11:56086560, rs11228710 Chr11:56431216, rs499037 Chr11:59480952,
rs2298553 Chr11:60265002, rs11231341 Chr11:62848487, rs77002186 Chr11:62850775,
rs1790218 Chr11:63057925, rs1815739 Chr11:66328095, rs35231465 Chr11:102584135,
rs497116 Chr11:104763117.

Rs16930998, rs4910844, rs61730422 localization overlaps with graphical grid of rs2647574, rs11228710 overlaps with rs117366703, rs2298553 overlaps with rs499037.Rs77002186, rs1790218 overlaps with rs11231341, therefore they’re not displayed on **Figure 18.**

There were 6 SNPs located on chromosome 12:

rs7485773 Chr12:7475081, rs16910526 Chr12:10271087, rs2708381 Chr12:11214145,
rs146753414 Chr12:52711747, rs2233919 Chr12:54577718, rs61942233 Chr12:113403675.

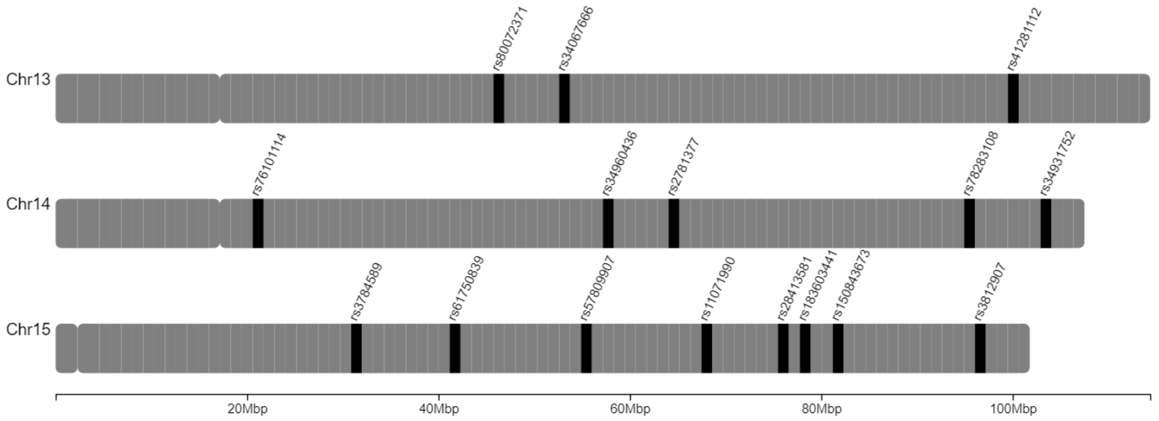

Figure 19. Localization of the SNPs on chromosomes 13,14,15

There were 3 SNPs located on chromosome 13 (**Figure 19**):

rs80072371 Chr13:46287373, rs34067666 Chr13:53617309, rs41281112 Chr13:100518634.

There are 5 SNPs located on chromosome 14:

rs76101114 Chr14:21500218, rs34960436 Chr14:57947421, rs2781377 Chr14:64560092,
rs78283108 Chr14:94935628, rs34931752 Chr14:102729886.

There are 8 SNPs located on chromosome 15:

rs3784589 Chr15:31294714, rs61750839 Chr15:42162467, rs57809907 Chr15:55722882,
rs11071990 Chr15:68497597, rs28413581 Chr15:76016583, rs183603441 Chr15:78807407,
rs150843673 Chr15:81624929, rs3812907 Chr15:97327393.

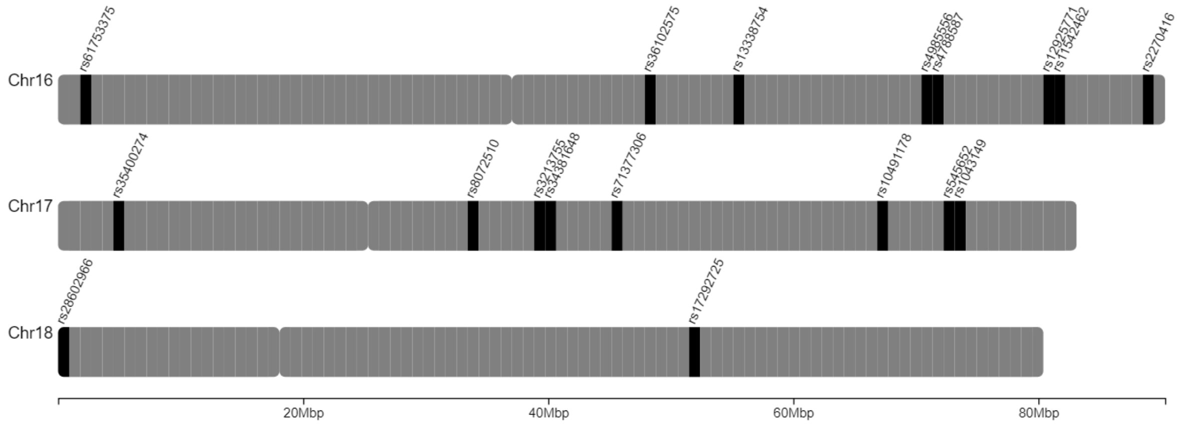

Figure 20. Localization of the SNPs on chromosomes 16,17,18

There were 9 SNPs located on chromosome 16 (**Figure 20**)**:**

rs61753375 Chr16:1825982, rs36102575 Chr16:48130781, rs13338754 Chr16:55903554,
rs4985556 Chr16:70694000, rs4788587 Chr16:72001136, rs12925771 Chr16:81199544,
rs7499011 Chr16:81242198, rs11542462 Chr16:82033810, rs2270416 Chr16:89261482.
Location of rs7499011 overlaps with graphical grid of rs12925771, therefore only 2nd is presented on **Figure 20**.

There were 10 SNPs located on chromosome 17:

rs35400274 Chr17:4803711, rs8072510 Chr17:33772658, rs3213755 Chr17:39197499,
rs34381648 Chr17:39880966, rs71377306 Chr17:45425287, rs74969489 Chr17:45452257,
rs118004742 Chr17:45468858, rs10491178 Chr17:67149973, rs545652 Chr17:72588806,
rs1043149 Chr17:74077797.
Location of rs118004742, rs74969489 overlaps with graphical grid of rs71377306, therefore only the last one is displayed on **Figure 20.**

There were 2 SNPs located on chromosome 18:

rs28602966 Chr18:658170, rs17292725 Chr18:51880889.

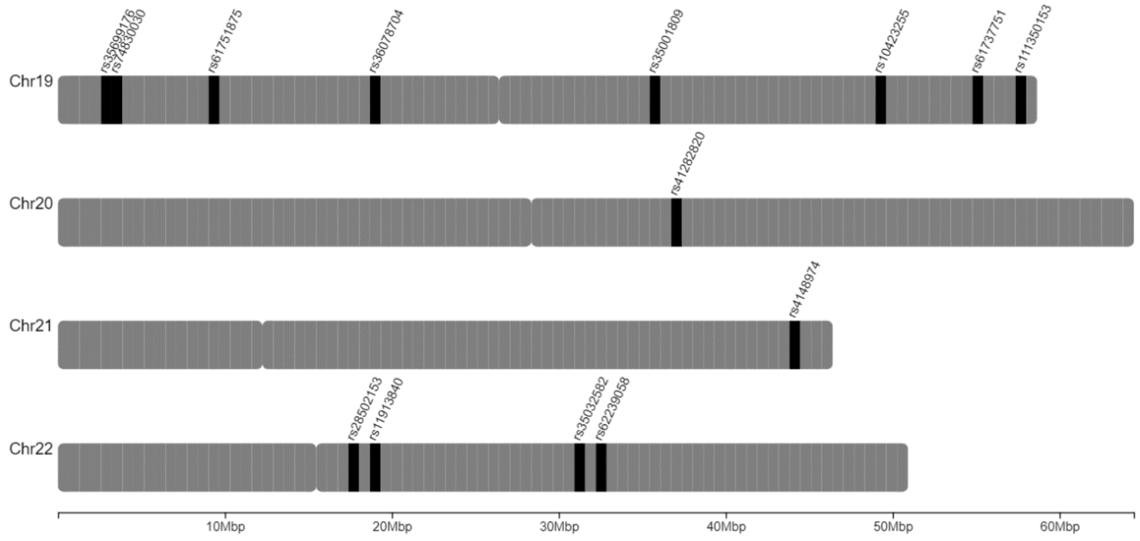

Figure 21. Localization of the SNPs on chromosomes 19,20,21,22

There were 10 SNPs located on chromosome 19 (**Figure 21**):

rs35699176 Chr19:2936535, rs74830030 Chr19:3789321, rs61751875 Chr19:9236969,
rs17001893 Chr19:9237263, rs36078704 Chr19:19039030, rs35001809 Chr19:35718891,
rs541169 Chr19:35719020, rs10423255 Chr19:49445774, rs61737751 Chr19:55019261,
rs111350153 Chr19:57646393.

Location of rs17001893 overlaps with graphical grid of rs61751875, therefore only 2nd is presented on **Figure 21.** Location of rs541169 overlaps with graphical grid of rs35001809, therefore only 2nd is presented on **Figure 21.**

There was 1 SNP located on chromosome 20 :

rs41282820 Chr20:36869005.

There was 1 SNP located on chromosome 21:

rs4148974 Chr21:44323720.

There were 4 SNPs located on chromosome 22:

rs28502153 Chr22:17469049 rs11913840 Chr22:18912677 rs35032582 Chr22:30891264
rs62239058 Chr22:32643460

There were no SNPs located on the sex chromosomes.

4.1.5 The frequency of particular nonsense mutations in the Polish population

Table 8. Genotype and allele frequencies of selected PTCs in the Polish population - calculated from data from the POPULOUS database.

	alleles	major allele (%)	Major allele homozygote	Heterozygote	Minor allele homozygote
rs75411676	G/T	94.8	4559(89.603%)	529(10.397%)	-
rs12139100	A/G	50.2	1267(33.36%)	1279(33.676%)	1252(32.965%)
rs114429815	T/C	50.4	1294(33.768%)	1278(33.351%)	1260(32.881%)
rs12077871	G/A	50.6	1276(33.456%)	1311(34.373%)	1227(32.171%)
rs6671527	A/G	50.2	1318(34.287%)	1223(31.816%)	1303(33.897%)
rs850763	T/G	50.5	1288(33.761%)	1279(33.526%)	1248(32.713%)
rs116389032	T/A	50.1	1285(33.595%)	1259(32.915%)	1281(33.49%)
rs17602729	A/G	50.5	1257(33.44%)	1280(34.052%)	1222(32.509%)
rs12568784	G/T	50.4	1283(33.745%)	1263(33.219%)	1256(33.035%)
rs74118444	T/G	50.1	1259(32.915%)	1316(34.405%)	1250(32.68%)
rs863362	T/C	50.2	1268(32.986%)	1324(34.443%)	1252(32.57%)
rs76330087	G/A	50.1	1248(33.042%)	1291(34.181%)	1238(32.777%)
rs5744168	G/A	51.1	1290(33.576%)	1347(35.06%)	1205(31.364%)
rs2273865	A/T	51	1311(34.319%)	1272(33.298%)	1237(32.382%)
rs62154921	T/G	50.4	1264(33.377%)	1286(33.958%)	1237(32.664%)
rs112050262	A/G	50.3	1287(33.524%)	1287(33.524%)	1265(32.951%)
rs34291832	G/C	99.2	5018(98.489%)	77(1.511%)	-
rs2176186	C/T	51.1	1287(33.94%)	1302(34.335%)	1203(31.725%)
rs115917139	C/T	96.6	4592(93.371%)	314(6.385%)	12(0.244%)
rs4639011	C/T	98	4887(95.918%)	208(4.082%)	-
rs114730569	A/G	50.5	1297(33.732%)	1288(33.498%)	1260(32.77%)
rs1043261	T/C	50.1	1294(33.654%)	1264(32.874%)	1287(33.472%)
rs3732781	A/C	70.4	2539(49.853%)	2090(41.037%)	464(9.111%)
rs79892855	G/A	100	5094(99.98%)	1(0.02%)	-
rs76438938	C/T	96.9	4784(93.896%)	309(6.065%)	2(0.039%)
rs74437357	G/A	97.9	4876(95.739%)	216(4.241%)	1(0.02%)
rs1861050	T/C	50	1286(33.683%)	1246(32.635%)	1286(33.683%)
rs61731313	C/T	50.5	1271(33.298%)	1311(34.346%)	1235(32.355%)
rs111696697	A/T	50.2	1269(33.316%)	1285(33.736%)	1255(32.948%)
rs112033303	A/T	50.4	1278(33.552%)	1283(33.683%)	1248(32.765%)
rs7447815	C/G	63.7	2076(40.762%)	2332(45.788%)	685(13.45%)
rs1023840	C/T	80.3	3286(64.495%)	1610(31.6%)	199(3.906%)
rs10471773	G/A	97	4791(94.033%)	302(5.927%)	2(0.039%)
rs35391433	C/T	97.8	4879(95.761%)	209(4.102%)	7(0.137%)
rs12520799	T/A	59.6	1831(35.944%)	2408(47.271%)	855(16.784%)
rs17184009	C/T	99	4989(97.92%)	106(2.08%)	-

rs3130453	C/T	52.4	1399(27.469%)	2542(49.912%)	1152(22.619%)
rs72856718	C/A	90.6	4182(82.193%)	855(16.804%)	51(1.002%)
rs45621032	T/A	98.9	4978(97.761%)	114(2.239%)	-
rs6907580	G/A	95.1	4601(90.304%)	486(9.539%)	8(0.157%)
rs8192646	C/T	95.2	4614(90.631%)	463(9.094%)	14(0.275%)
rs34672740	C/A	100	5094(99.98%)	1(0.02%)	-
rs34427887	C/T	94.3	4527(88.957%)	548(10.768%)	14(0.275%)
rs2235197	G/A	91.7	4186(84.225%)	740(14.889%)	44(0.885%)
rs10237332	C/T	100	5092(99.941%)	3(0.059%)	-
rs67047829	G/A	90.6	4198(82.395%)	841(16.506%)	56(1.099%)
rs2293766	G/A	99.5	5044(99.018%)	50(0.982%)	-
rs10261977	C/T	83.4	3542(69.519%)	1416(27.792%)	137(2.689%)
rs328	C/G	91.6	4282(84.06%)	767(15.057%)	45(0.883%)
rs117752382	A/T	98.7	4964(97.467%)	129(2.533%)	-
rs138377917	G/A	98.7	4589(97.39%)	119(2.525%)	4(0.085%)
rs2272754	G/T	84.4	3634(71.437%)	1314(25.831%)	139(2.732%)
rs2039381	G/A	99.2	5009(98.312%)	86(1.688%)	-
rs10981589	G/A	95.5	4649(91.282%)	432(8.482%)	12(0.236%)
rs45579335	G/T	98.7	4957(97.311%)	137(2.689%)	-
rs1476860	G/A	69.6	2468(48.449%)	2154(42.285%)	472(9.266%)
rs35898523	G/T	94.1	4513(88.594%)	559(10.974%)	22(0.432%)
rs55727303	C/T	98.5	4899(96.991%)	152(3.009%)	-
rs9886752	G/A	88	3941(77.411%)	1078(21.175%)	72(1.414%)
rs1044261	C/T	93.9	4487(88.067%)	596(11.698%)	12(0.236%)
rs12240276	C/T	83	3493(68.598%)	1464(28.751%)	135(2.651%)
rs41291550	T/A	99.7	5026(99.426%)	29(0.574%)	-
rs7904983	G/A	100	5092(99.961%)	2(0.039%)	-
rs57026471	C/T	86.9	3856(75.697%)	1140(22.379%)	98(1.924%)
rs2647574	C/T	58	1715(33.667%)	2481(48.704%)	898(17.629%)
rs16930998	G/A	98.1	4903(96.232%)	192(3.768%)	-
rs4910844	A/T	78.4	3138(61.614%)	1707(33.517%)	248(4.869%)
rs61730422	G/A	91.5	4264(83.69%)	796(15.623%)	35(0.687%)
rs35233100	C/T	94.2	4521(88.734%)	554(10.873%)	20(0.393%)
rs7120775	C/G	86.4	3817(74.931%)	1169(22.949%)	108(2.12%)
rs10838851	A/T	78.4	3112(61.079%)	1761(34.563%)	222(4.357%)
rs1459101	C/T	75.3	2855(56.046%)	1959(38.457%)	280(5.497%)
rs117366703	C/T	98.5	4924(96.948%)	153(3.012%)	2(0.039%)
rs11228710	T/C	63.3	2045(40.145%)	2363(46.388%)	686(13.467%)
rs499037	G/A	98.8	4938(97.57%)	123(2.43%)	-
rs2298553	T/C	50.6	1367(26.835%)	2426(47.625%)	1301(25.54%)
rs11231341	C/A	79.5	3227(63.337%)	1649(32.365%)	219(4.298%)
rs77002186	G/A	98.2	4910(96.369%)	185(3.631%)	-
rs1790218	A/G	57.5	1558(32.164%)	2458(50.743%)	828(17.093%)
rs1815739	C/T	59.3	1815(35.651%)	2413(47.397%)	863(16.951%)
rs35231465	G/A	98.1	4904(96.251%)	188(3.69%)	3(0.059%)
rs497116	A	100	5095(100%)	-	-
rs7485773	C/T	97.3	4820(94.64%)	272(5.341%)	1(0.02%)

rs16910526	A/C	92.7	4375(85.885%)	697(13.683%)	22(0.432%)
rs2708381	C/T	73.3	2642(52.735%)	2063(41.178%)	305(6.088%)
rs146753414	C/A	98.7	4961(97.389%)	132(2.591%)	1(0.02%)
rs2233919	G/A	100	5049(99.921%)	4(0.079%)	-
rs61942233	C/T	98.7	4960(97.427%)	130(2.554%)	1(0.02%)
rs80072371	C/A	93.7	4469(87.817%)	601(11.81%)	19(0.373%)
rs34067666	C/T	98.7	4966(97.468%)	128(2.512%)	1(0.02%)
rs41281112	C/T	98.3	4917(96.639%)	169(3.322%)	2(0.039%)
rs76101114	C/G	99.9	5086(99.843%)	8(0.157%)	-
rs34960436	G/A	97.6	4850(95.191%)	245(4.809%)	-
rs2781377	G/A	93.7	4469(87.765%)	601(11.803%)	22(0.432%)
rs78283108	G/A	99.9	5087(99.882%)	6(0.118%)	-
rs34931752	G/A	95.5	4637(91.047%)	450(8.836%)	6(0.118%)
rs3784589	C/A	91.6	4270(83.808%)	796(15.623%)	29(0.569%)
rs61750839	G/A	94.4	4321(88.891%)	540(11.109%)	-
rs57809907	C/A	94.4	4540(89.142%)	535(10.505%)	18(0.353%)
rs11071990	G/A	99.6	5038(99.115%)	45(0.885%)	-
rs28413581	C/T	100	5093(99.98%)	1(0.02%)	-
rs183603441	T/A	98.9	4974(97.721%)	115(2.259%)	1(0.02%)
rs150843673	G/T	98.3	4926(96.721%)	164(3.22%)	3(0.059%)
rs3812907	C/T	89.3	4053(79.642%)	979(19.238%)	57(1.12%)
rs61753375	C/T	99.3	4933(98.522%)	74(1.478%)	-
rs36102575	C/T	97.3	4824(94.681%)	266(5.221%)	5(0.098%)
rs13338754	G/A	100	5092(99.941%)	3(0.059%)	-
rs4985556	C/A	89.9	4127(81.16%)	889(17.483%)	69(1.357%)
rs4788587	G/A	82.2	3442(67.689%)	1476(29.027%)	167(3.284%)
rs12925771	G/A	72.5	2656(52.181%)	2064(40.55%)	370(7.269%)
rs7499011	G/A	54.5	1099(27.169%)	2209(54.611%)	737(18.22%)
rs11542462	G/A	88.3	3976(78.037%)	1049(20.589%)	70(1.374%)
rs2270416	C/A	97.3	4789(94.7%)	264(5.22%)	4(0.079%)
rs35400274	G/A	85.9	3289(73.464%)	1116(24.927%)	72(1.608%)
rs8072510	G/T	87.9	3907(77.107%)	1095(21.61%)	65(1.283%)
rs3213755	G/A	85.5	3458(71.594%)	1344(27.826%)	28(0.58%)
rs34381648	G/T	100	5092(99.941%)	3(0.059%)	-
rs71377306	C/T	95.3	4618(90.941%)	446(8.783%)	14(0.276%)
rs74969489	A/T	98.4	4930(96.8%)	162(3.181%)	1(0.02%)
rs118004742	T/G	94.4	4526(89.042%)	544(10.702%)	13(0.256%)
rs10491178	G/A	95.3	4635(91.025%)	439(8.621%)	18(0.353%)
rs545652	C/A	84.1	3573(70.376%)	1393(27.437%)	111(2.186%)
rs1043149	C/T	82.6	3492(68.538%)	1435(28.165%)	168(3.297%)
rs28602966	G/T	97.6	4822(95.24%)	240(4.74%)	1(0.02%)
rs17292725	G/A	96.5	4736(93.082%)	348(6.84%)	4(0.079%)
rs35699176	G/A	96.2	4710(92.462%)	377(7.401%)	7(0.137%)
rs74830030	G/A	100	5088(99.98%)	1(0.02%)	-
rs61751875	G/A	100	5090(99.902%)	5(0.098%)	-
rs17001893	G/A	100	5093(99.961%)	2(0.039%)	-
rs36078704	C/T	95.7	1578(91.372%)	149(8.628%)	-

rs35001809	C/T	96.6	4764(93.503%)	319(6.261%)	12(0.236%)
rs541169	C/T	64.9	1859(41.375%)	2116(47.095%)	518(11.529%)
rs10423255	C/T	92.6	4299(85.62%)	696(13.862%)	26(0.518%)
rs61737751	C/T	95.8	4673(91.717%)	417(8.184%)	5(0.098%)
rs111350153	T/A	98.2	4911(96.426%)	179(3.515%)	3(0.059%)
rs41282820	G/A	98.2	4913(96.428%)	179(3.513%)	3(0.059%)
rs4148974	C/T	94.9	4588(90.049%)	495(9.715%)	12(0.236%)
rs28502153	C/A	63.9	2095(41.119%)	2323(45.594%)	677(13.288%)
rs11913840	C/T	97.7	4862(95.521%)	224(4.401%)	4(0.079%)
rs35032582	C	100	5094(100%)	-	-
rs62239058	C/A	99	4993(98.037%)	100(1.963%)	-

The most frequent minor allele homozygotes in the Polish population were: rs6671527 (33.9%) , rs1861050 (33.68%), and rs116389032 (33.49%). rs6671527 is located in chr1:46615007 in the MOB kinase activator 3C (*MOB3C)* gene. Rs1861050 is located in chr4:15480736 in the coiled-coil and C2 domain containing 2A (*CC2D2A*) gene. Rs116389032 is located in chr1:109280835 in the proline and serine rich coiled-coil 1 (*PSRC1)* gene.

Minor allele homozygotes of 22 SNPs occurred with a frequency over 30%. Only 2 RSs were not present in the analyzed dataset – as these were monomorphic (rs35032582 and rs497116) **Table 8**.

The number of SNPs with minor allele frequencies above 5% was 141 – 64 = 77 (55%) and these SNPs should probably be regarded as having near-neutral selection.

4.2 Association analysis – longevity

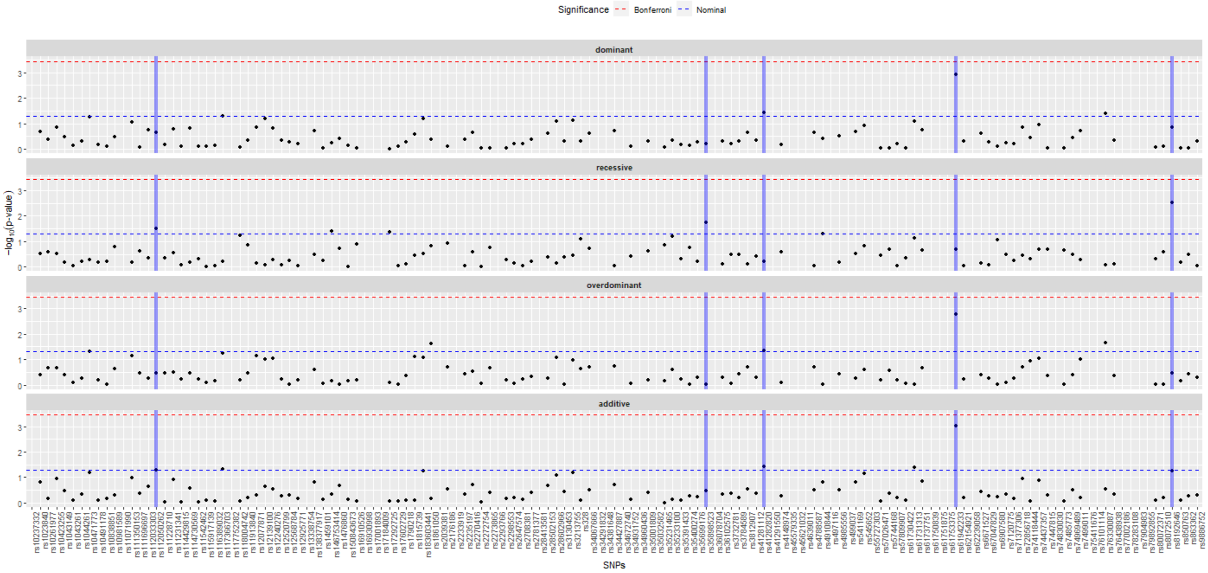

Figure 22. Manhattan-type plot of the association p-value between selected SNPs versus patient's age for different genetic models. The analysis presented here shows that 5 SNPs passed the nominal p.value threshold (p<0.05) for either recessive or dominant model (blue stripes). A high-resolution graph is available as electronic version on attached cd (named Figure 22). More detailed analyses for these 5 RSs were conducted.

Table 9. Detailed ***rs112050262*** *association analysis with patient's age for selected genetic models*

	n	mean	se	dif	lower	upper	p-value	AIC
Dominant								
A/A	1280	42	0.413	0			0.22873	31324
G/A-G/G	2524	42.61	0.2962	0.6133	-0.38523	1.6119		
Recessive								
A/A-G/A	2551	42.04	0.2913	0			0.03032	31321
G/G	1253	43.15	0.4267	1.1093	0.105824	2.1127		
Overdominant								
A/A-G/G	2533	42.57	0.297	0			0.34117	31325
G/A	1271	42.08	0.4109	-0.4859	-1.48625	0.5145		
log-Additive								
0,1,2				0.5737	-0.00438	1.1517	0.05183	31322

n- sample size, se – standard error, dif- means difference, lower/upper 95% Confidence Interval for dif, AIC- Akaike information criterion

The first SNP from the longevity analysis (from the Manhattan plot in **Figure 22)** which was analyzed that passed the nominal p.value threshold for the recessive model was rs112050262. This polymorphism is located in the *SULT1C3* gene on chromosome 2. Based on the lowest AIC value, the recessive model was the best fitted model for this dataset

Table 9. This SNP did not exhibit Hardy Weinberg equilibrium (p <0.05). The mean value of age in the G/G (minor allele homozygote) group was 43.15 years old (y) (sd = 15.1, n= 1253). The mean value of age in A/A+G/A (major allele and heterozygote) group was 42.04 y (sd= 14.7, n= 2551). The difference in means was 1.11 years with 95% CI lower 0.106, upper 2.113. The mean age of subjects with the G/G variant was slightly lower than the second group (Cohen's d= 0.074, very small effect size due to the large variance).

Table 10. Detailed rs41282820 association analysis with patient's age for selected genetic models

	n	mean	se	dif	lower	upper	p-value	AIC
Dominant								
G/G	4870	42.34	0.2124	0			0.03833	41570
G/A-A/A	180	44.67	1.1125	2.332	0.1259	4.538		
Recessive								
G/G-G/A	5047	42.42	0.2088	0			0.59301	41574
A/A	3	47	4	4.579	-12.212	21.37		
Overdominant								
A/A-G/G	4873	42.34	0.2123	0			0.04367	41570
G/A	177	44.63	1.1298	2.289	0.0656	4.513		
log-Additive								
0,1,2				2.295	0.14267	4.447	0.03668	41570

n- sample size, se – standard error, dif- means difference, lower/upper 95% Confidence Interval for dif, AIC- Akaike information criterion

The second SNP in the longevity analysis that was analyzed that passed the nominal p.value threshold for the dominant, overdominant and log-additive models was rs41282820. This polymorphism is located in the *KIAA1755* gene on chromosome 20. We could not determine which model fitted our data better from the AIC values (the AIC values for all the models were the same **Table 10**), although there were only 3 subjects with the minor allele homozygote which may diminish the effect of this analysis, except with the dominant model, which was therefore the only reliable model in this case.

This SNP followed Hardy-Weinberg equilibrium (p=0.23). The mean value of age in the G/G (major allele homozygote) group was 42.34 y (sd=14.82, n= 4870). The mean value of age in G/A-A/A (heterozygote, minor allele homozygote) group was 44.67 y (sd =14.84, n= 180). The difference in means was 2.33 years with 95% CI lower 0.126, upper 4.538. The mean age of subjects with at least one minor allele was slightly higher than for the group with major allele homozygote (Cohen's d = 0.157, very small effect size).

Table 11. Detailed rs61942233 association analysis with patient's age for selected genetic models

	n	mean	se	dif	lower	upper	p-value	AIC
Dominant								
C/C	4915	42.32	0.2115	0			0.001231	41531
T/C-T/T	131	46.56	1.2766	4.242	1.671	6.813		
Recessive								
C/C-T/C	5045	42.42	0.2088	0			0.210466	41540
T/T	1	61	0	18.578	-10.495	47.651		
Overdominant								
C/C-T/T	4916	42.32	0.2114	0			0.001735	41532
T/C	130	46.45	1.2815	4.127	1.546	6.708		
log-Additive								
0,1,2				4.288	1.746	6.83	0.000951	41531

n- sample size, se – standard error, dif- means difference, lower/upper 95% Confidence Interval for dif, AIC- Akaike information criterion

The third SNP in the longevity analysis which was analyzed passed the nominal p.value threshold for almost all tested models (except recessive): rs61942233. This polymorphism is located in the *OAS3* gene on chromosome 12. The best genetic model (based on AIC value) for this data was either dominant or log-additive **Table 11**. The same situation occurred as in the previous analysis i.e. there was a very small sample size in the minor allele homozygote group (with only 1 subject). This SNP followed Hardy-Weinberg equilibrium (p = 0.577). The mean value for age in the C/C (major allele homozygote) group was 42.32 y (sd =14.82, n= 4915). The mean value for age in T/C-T/T group was 46.56 y (sd= 14.55, n =131. The mean age of subjects with at least one minor allele was slightly higher than the group's with major allele homozygote (Cohen's d = 0.28, small effect size).

Table 12. Detailed rs8192646 association analysis with patient's age for selected genetic models

	n	mean	se	dif	lower	upper	p-value	AIC
Dominant								
C/C	4573	42.33	0.2186	0			0.143899	41540
T/C-T/T	473	43.37	0.7043	1.0471	-0.357	2.451		
Recessive								
C/C-T/C	5032	42.39	0.2091	0			0.00307	41533
T/T	14	54.14	2.8243	11.7502	3.97533	19.525		
Overdominant								
C/C-T/T	4587	42.36	0.2183	0			0.347306	41541
T/C	459	43.05	0.7155	0.6826	-0.7407	2.106		
log-Additive								
0,1,2				1.3096	-0.0334	2.653	0.056134	41538

n- sample size, se – standard error, dif- means difference, lower/upper 95% Confidence Interval for dif, AIC- Akaike information criterion

The fourth SNP in the longevity analysis that passed the nominal p.value threshold for codominant and recessive models was rs8192646. This polymorphism is located in the *TAAR2* gene on chromosome 6. Based on AIC the best genetic model for this analysis was the recessive model **Table 12**. This SNP followed Hardy-Weinberg equilibrium (p=0.447). The mean value of age in the T/T (minor allele homozygote) group was 54.14 y (sd= 10.18, n=14). The mean value of age in C/C+T/C group was 42.39 y (sd= 14.82, n =5032. The mean age of subjects with 2 minor alleles was significantly higher than the second group's (Cohen's d= 0.94, large effect size).

Table 13. Detailed rs35898523 association analysis with patient's age for selected genetic models

	n	mean	se	dif	lower	upper	p-value	AIC
Dominant								
G/G	4474	42.39	0.2201	0			0.6177	41565
T/G-T/T	575	42.72	0.6526	0.328	-0.9599	1.616		
Recessive								
G/G-T/G	5027	42.39	0.2091	0			0.01772	41560
T/T	22	49.91	3.0436	7.5142	1.3059	13.723		
Overdominant								
G/G-T/T	4496	42.43	0.2197	0			0.99386	41565
T/G	553	42.43	0.6653	0.00514	-1.305	1.315		
log-Additive								
0,1,2				0.58154	-0.6356	1.799	0.34909	41564

n- sample size, se – standard error, dif- means difference, lower/upper 95% Confidence Interval for dif, AIC- Akaike information criterion

The last SNP in the longevity analysis that passed the nominal p.value threshold for the recessive model was rs35898523. This polymorphism is located in the *GBGT1* gene on chromosome 9. Based on AIC (the lowest AIC value), the recessive model was the best fitted model for this dataset **Table 13**. This SNP followed Hardy-Weinberg equilibrium (p=0.312). The mean value for age in the T/T (minor allele homozygote) group was 49.91 y (sd = 14.28, n =22), 22 subjects were in this group. The mean value of age in the G/G-T/G (major allele homozygote, heterozygote) group was 42.39 y (sd= 14.82, n =5027). The mean age of subjects with 2 minor alleles was higher than the second group’s (Cohen’s d= 0.51, medium effect size).

The raw p.value data used to generate **Figure 22** is available as the excel spreadsheet “YOB_ANALYSIS” on attached CD.

4.3 Association analysis – number of children (NCI)

The association analysis with the number of children per individual (NCI) was conducted for two data sets: the whole group, characterized in section **4.1,** and the subset of subjects above the age of reproductive activity. The criteria for selection of the second group were described in section **3.2.1** .

4.3.1 Whole group – association with NCI

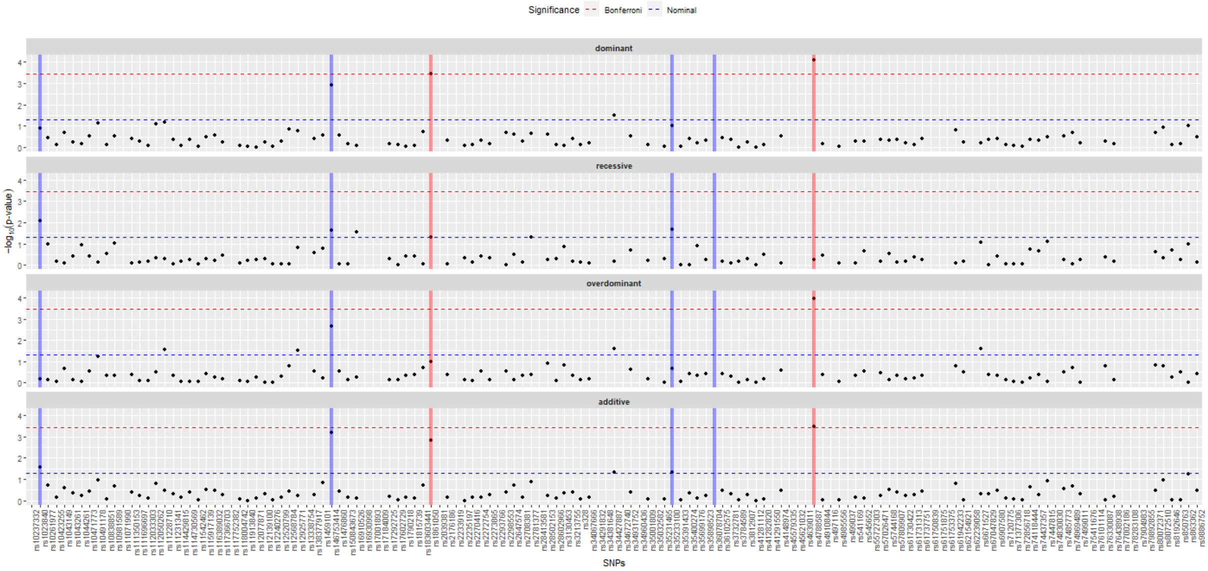

*Figure 23. Manhattan-type plot of the association p-values between selected SNPs with number of children per individual for different genetic models. The analysis presented above (**Figure 23**) shows that 3 SNPs passed the nominal (p <0.05) p.value threshold for either dominant or recessive genetic models (blue stripes). 2 SNPs passed the Bonferroni corrected (red stripes) p.value threshold. Overall 6 SNPs were selected for further, more detailed analysis. A high-resolution graph is available as an electronic version on the attached cd (named Figure 23).*

The first RS from the NCI study (from the Manhattan plot in Figure 23) analyzed that passed the nominal p.value threshold for the reccesive model was rs1023840. This polymorphism is located in the *MROH2B* gene on chromosome 5. Based on AIC (the lowest AIC value), the recessive model was the best fitted model for this dataset **Table 14**. This SNP followed Hardy-Weinberg equilibrium (p = 0.92). The mean value of NCI in T/T (minor homozygote) group was 1.421 children (sd= 1.3, n= 197). The mean value of NCI in C/C-T/C (major homozygote, heterozygote) group was 1.59 children (sd = 1.52, n =4805). The mean value of number of children was slightly lower in the minor homozygote group than in second group (Cohen’s d= 0.12, very small effect size).

Table 14. Detailed rs1023840 association analysis with NCI for selected genetic models

	n	mean	se	dif	lower	upper	p-value	AIC
Dominant								
C/C	3218	1.59	0.0270	0			0.1309	16824
T/C-T/T	1784	1.571	0.0355	-0.0580	-0.1333	0.0172		
Recessive								
C/C-T/C	4805	1.59	0.0221	0			0.0081	16820
T/T	197	1.421	0.0927	-0.2502	-0.4354	-0.0650		
Overdominant								
C/C-T/T	3415	1.58	0.0260	0			0.6540	16826
T/C	1587	1.59	0.0382	-0.0177	-0.0952	0.0597		
log-Additive								
0,1,2				-0.0717	-0.1356	-0.0078	0.0280	16822

n- sample size, se – standard error, dif- means difference, lower/upper 95% Confidence Interval for dif, AIC- Akaike information criterion

Table 15. Detailed rs146753414 association analysis with NCI for selected genetic models

	n	mean	se	dif	lower	upper	p-value	AIC
Dominant								
C/C	4872	1.573	0.0218	0			0.0012	16813
A/C-A/A	129	1.992	0.1372	0.3763	0.1492	0.6034		
Recessive								
C/C-A/C	5000	1.583	0.0215	0			0.0243	16819
A/A	1	6	0.0000	2.9300	0.3809	5.4791		
Overdominant								
C/C-A/A	4873	1.574	0.0218	0			0.0022	16815
A/C	128	1.961	0.1347	0.3557	0.1277	0.5836		
log-Additive								
0,1,2				0.3904	0.1660	0.6149	0.0007	16812

The second RS in the NCI study analyzed that passed the nominal p.value threshold for dominant genetic model was rs146753414. This SNP is located in *KRT83* gene on chromosome 12. Based on AIC the dominant model was a better model than recessive **Table 15**. This SNP followed Hardy-Weinberg equilibrium (p=0.588). The mean value of NCI in the C/C (major homozygote) group was 1.573 children (sd= 1.51, n=4872). The mean value of NCI in the A/C-A/A (heterozygote, minor homozygote) group was 1.992 children (sd= 1.56, n =129). The mean value of NCI in minor homozygote-heterozygote group was slightly higher (Cohen's d =0.27, very small effect size).

Table 16. Detailed rs35233100 association analysis with NCI for selected genetic models

	n	mean	se	dif	lower	upper	p-value	AIC
Dominant								
C/C	4435	1.569	0.0227	0			0.0975	16824
T/C-T/T	567	1.691	0.0663	0.0961	-0.0175	0.2097		
Recessive								
C/C-T/C	4982	1.582	0.0215	0			0.0217	16821
T/T	20	1.85	0.4309	0.6692	0.0981	1.2403		
Overdominant								
C/C-T/T	4455	1.571	0.0227	0			0.2227	16825
T/C	547	1.686	0.0670	0.0718	-0.0436	0.1873		
log-Additive								
0,1,2				0.1104	0.0026	0.2182	0.0448	16823

n- sample size, se – standard error, dif- means difference, lower/upper 95% Confidence Interval for dif, AIC- Akaike information criterion

The third RS in the NCI study analyzed that passed the nominal p.value threshold for recessive model was rs35233100. This SNP is located in the *MADD* gene on chromosome 12. Based on AIC (the lowest AIC value), the recessive model was the best fitted model for this dataset **Table 16**. This SNP followed Hardy-Weinberg equilibrium (p=0.445). The mean value of NCI in T/T (minor homozygote) group was 1.85 children (sd= 1.93, n=20). The mean value of NCI in the C/C-T/C (major homozygote, heterozygote) group was 1.569 children (sd= 1.51, n = 4982). The mean value of number of children born was slightly higher in the minor allele homozygote group than in second group (Cohen's d= 0.162, very small effect size).

Table 17. Detailed rs1861050 association analysis with NCI for selected genetic models

	n	mean	se	dif	lower	upper	p-value	AIC
Dominant								
C/C	1259	1.463	0.03964	0			0.000356	12573
T/C-T/T	2490	1.64	0.03136	0.1599	0.07221	0.2476		
Recessive								
C/C-T/C	2485	1.564	0.02954	0			0.051022	12582
T/T	1264	1.612	0.04493	0.08734	-0.0003	0.175		
Overdominant								
C/C-T/T	2523	1.538	0.03	0			0.103985	12583
T/C	1226	1.668	0.04373	0.07335	-0.0151	0.1618		
log-Additive								
0,1,2				0.08199	0.03149	0.1325	0.001473	12575

n- sample size, se – standard error, dif- means difference, lower/upper 95% Confidence Interval for dif, AIC- Akaike information criterion

The fifth SNP from the NCI study (from the Manhattan plot in Figure 23) , that passed p.value after Bonferroni correction for dominant model, was rs1861050. This SNP is located in *CC2D2A* on chromosome 4. Based on AIC the dominant model was the best fitted model for this dataset **Table 17**. This SNP did not follow Hardy-Weinberg equilibrium (p < 0.05). The mean value of NCI in the C/C (major allele homozygote) group was 1.463 children- (sd= 1.403, n= 1259. The mean value of NCI in the T/C-T/T (heterozygote, minor allele homozygote) group was 1.64 children (sd= 1.56, n =2490). The mean value of NCI in the T/C – T/T group was slightly higher than in major allele homozygote group (Cohen's d =0.12, very small effect size).

Table 18. Detailed rs4788587 association analysis with NCI for selected genetic models

	n	mean	se	dif	lower	upper	p-value	AIC
Dominant								
G/G	3383	1.646	0.02664	0			7.72E-05	16785
A/G-A/A	1610	1.452	0.03623	-0.1556	-0.2327	-0.0785		
Recessive								
G/G-A/G	4829	1.586	0.02196	0			0.577698	16800
A/A	164	1.53	0.10887	-0.0575	-0.2599	0.1449		
Overdominant								
G/G-A/A	3547	1.641	0.0259	0			0.000116	16785
A/G	1446	1.443	0.03841	-0.1564	-0.2358	-0.0769		
log-Additive								
0,1,2				-0.1215	-0.1879	-0.0551	0.000337	16787

n- sample size, se – standard error, dif- means difference, lower/upper 95% Confidence Interval for dif, AIC- Akaike information criterion

The last analyzed SNP from the NCI study, that passed p.value Bonferroni correction threshold for dominant model, was rs4788587. This SNP is located in *PKD1L3* gene on chromosome 16. Based on AIC the dominant model was the best fitted model for this dataset **Table 18**. This SNP followed Hardy-Weinberg equilibrium (p= 0.45). The mean value of NCI in the G/G (major allele homozygote) group was 1.646 children (sd=1.54, n =3383). The mean value of NCI in the A/G-A/A (heterozygote, minor allele homozygote) group was 1.452 children (sd= 1.45, n =1610). The mean value of NCI was slightly lower in A/G-A/A group than in major allele homozygote group (Cohen's d=0.13, very small effect size).

All of the raw p.value data used to generate **Figure 23** is available as the excel spreadsheet "NCI_ANALYSIS" an on CD.

4.3.2 Subset characteristics

Table 19. Summary of the subset

	[ALL]N=1722	N
SEX:		1722
Female	1132 (65.7%)	
Male	590 (34.3%)	
Age	59.1 (7.89)	1722
Number of children born	2.41 (1.52)	1722

N- sample size; for qualitative variable n (%) of the group is presented, for quantitative mean and (standard deviation).

The subset analyzed with NCI consisted of 1722 subjects. **Table 19** shows that 1132 of them were female (65.7%) and 590 were male (34.3%). The average age was 59.1 y (sd= 7.89) years, and mean value of NCI was 2.41 children (sd= 1.52).

4.3.3 Subset – NCI analysis

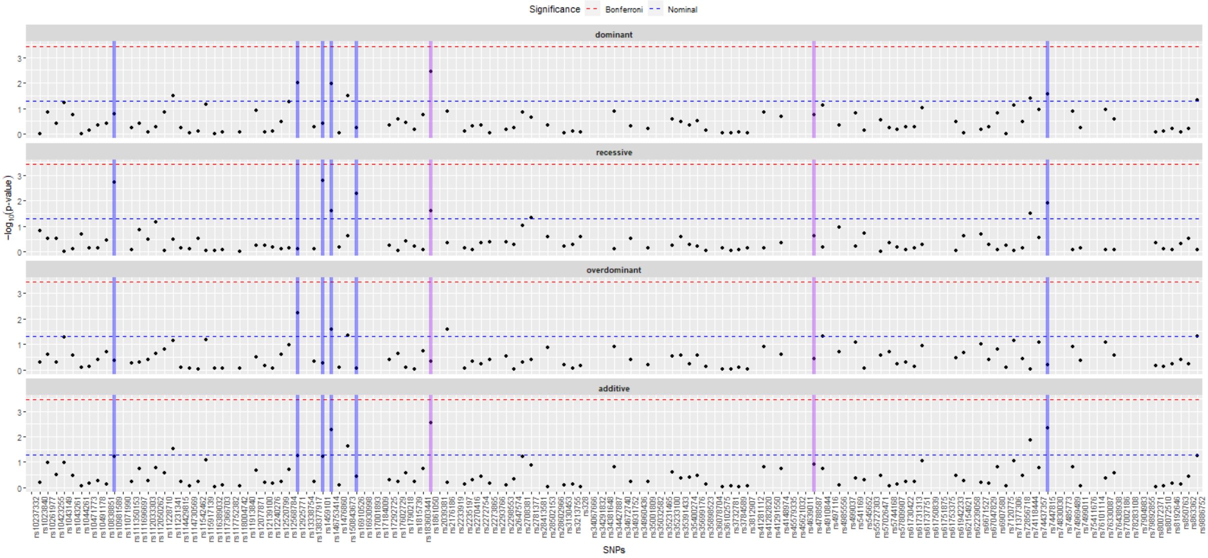

Figure 24. Manhattan-type plot of association p-values between selected PTCs with numbers of children per individual for different genetic models – subset. The analysis presented above shows that 6 SNPs passed the nominal p.value threshold (p<0.05) for either the recessive or dominant model (blue stripes). Purple stripes represent RS association p.values that passed Bonferroni correction in the previous analysis. A more detailed analysis for these 8 RSs was conducted. A high-resolution graph is available as an electronic version on the attached cd (named Figure 24).

Table 20. Detailed rs10981589 association analysis with NCI subset for selected genetic models

	n	mean	se	dif	lower	upper	p-value	AIC
Dominant								
G/G	1564	2.393	0.0383	0			0.1699	6293
A/G-A/A	143	2.573	0.1402	0.1832	-0.0783	0.4447		
Recessive								
G/G-A/G	1702	2.402	0.0367	0			0.0019	6285
A/A	5	4.6	1.6912	2.1250	0.7879	3.4621		
Overdominant								
G/G-A/A	1569	2.4	0.0386	0			0.4360	6294
A/G	138	2.5	0.1302	0.1057	-0.1602	0.3715		
log-Additive								
0,1,2				0.2383	-0.0100	0.4866	0.0601	6291

The first RS from the NCI subset study (from the Manhattan plot in **Figure 24)** analyzed that passed the nominal p.value threshold for the recessive model was rs10981589. Based on AIC this is the best fitted model for this dataset **Table 20.** This SNP is located in the *ZNF883* gene on chromosome 9. This SNP followed Hardy-Weinberg equilibrium (p=0.142). The mean value of NCI in the A/A (minor allele homozygote) group was 4.6 children (sd = 3.54, n =5). The mean value of NCI in the G/G-A/G (major allele homozygote, heterozygote) group was 2.4 children (sd = 1.51, n =1702). The mean value of NCI was significantly higher in the A/A (minor allele homozygote) group than in the G/G-A/G group (Cohen's d = 0.8, large effect size).

Table 21. Detailed rs12925771 association analysis with NCI subset for selected genetic models

	n	mean	se	dif	lower	upper	p-value	AIC
Dominant								
G/G	889	2.314	0.0475	0			0.0100	6282
A/G-A/A	816	2.512	0.0573	0.1907	0.0457	0.3356		
Recessive								
G/G-A/G	1573	2.412	0.0389	0			0.8020	6289
A/A	132	2.371	0.1170	-0.0347	-0.3059	0.2365		
Overdominant								
G/G-A/A	1021	2.321	0.0440	0			0.0057	6281
A/G	684	2.539	0.0645	0.2086	0.0608	0.3563		
log-Additive								
0,1,2				0.1121	-0.0022	0.2264	0.0547	6285

n- sample size, se – standard error, dif- means difference, lower/upper 95% Confidence Interval for dif, AIC- Akaike information criterion

The second RS in the NCI subset study that passed the nominal p.value threshold for dominant and overdominant models was rs12925771. Based on AIC values the overdominant model was the best fitted model for this dataset **Table 21**. This SNP is located in *PKD1L2* gene on chromosome 16. This SNP followed Hardy-Weinberg equilibrium (p = 0.95). The mean value of NCI in the A/G (heterozygote) group was 2.539 children (sd= 1.68, n =684). The mean value of NCI in the A/A-G/G (minor allele homozygote, major allele homozygote) group was 2.321 children (sd=1.45, n =1021). The mean value of NCI in A/G was slightly higher than in the A/A-G/G group (Cohen's d = 0.139, very small effect size).

Table 22. Detailed rs1459101 association analysis with NCI subset for selected genetic models

	n	mean	se	dif	lower	upper	p-value	AIC
Dominant								
C/C	959	2.376	0.0488	0			0.4030	6294
T/C-T/T	748	2.449	0.0566	0.0624	-0.0838	0.2085		
Recessive								
C/C-T/C	1620	2.382	0.0375	0			0.0016	6284
T/T	87	2.897	0.1900	0.5304	0.2018	0.8590		
Overdominant								
C/C-T/T	1046	2.42	0.0477	0			0.5666	6294
T/C	661	2.39	0.0587	-0.0435	-0.1925	0.1054		
log-Additive								
0,1,2				0.1165	-0.0056	0.2386	0.0616	6291

n- sample size, se – standard error, dif- means difference, lower/upper 95% Confidence Interval for dif, AIC- Akaike information criterion

The third RS in the NCI subset study analyzed that passed the nominal p.value threshold for recessive model was rs1459101. Based on AIC values this was the best fitted model for this dataset **Table 22**. This SNP is located in *OR4C16* gene on chromosome 11. This SNP followed Hardy-Wenberg equilibrium (p= 0.06). The mean value of NCI in the T/T (minor allele homozygote) group was 2.897 children (sd=1.75, n =87). The mean value of NCI in the C/C-T/C (major allele homozygote, heterozygote) group was 2.382 children (sd= 1.51, n =1620). The mean value of NCI was slightly higher in the T/T group than in the C/C-T/C group (Cohen's d = 0.31 , small effect size).

Table 23. Detailed rs146753414 association analysis with NCI subset for selected genetic models

	n	mean	se	dif	lower	upper	p-value	AIC
Dominant								
C/C	1661	2.392	0.0375	0			0.0109	6288
A/C-A/A	46	3	0.2062	0.5808	0.1341	1.0276		
Recessive								
C/C-A/C	1706	2.406	0.0370	0			0.0249	6289
A/A	1	6	0.0000	3.4285	0.4353	6.4216		
Overdominant								
C/C-A/A	1662	2.394	0.0376	0			0.0255	6290
A/C	45	2.933	0.1995	0.5150	0.0634	0.9665		
log-Additive								
0,1,2				0.6171	0.1844	1.0498	0.0052	6287

n- sample size, se – standard error, dif- means difference, lower/upper 95% Confidence Interval for dif, AIC- Akaike information criterion

The fourth RS in the NCI subset study analyzed that passed the nominal p.value threshold for the dominant model was rs146753414. Only 1 person had the minor allele homozygote variant, and therefore the dominant model will be described

Table 23. This SNP is located in *KRT83* gene on chromosome 12. This SNP followed Hardy-Weinberg equilibrium (p= 0.28). The mean value of NCI in the C/C (major allele homozygote) group was 2.39 children (sd=1.52, n =1661). The mean value of NCI in the A/C-A/A (heterozygote, minor allele homozygote) group was 3 children (sd=1.36, n = 46). The mean value of NCI was slightly higher in the A/C-A/A group than in the major allele homozygote group (Cohen's d=0.42 ,small effect size).

Table 24. Detailed rs16910526 association analysis with NCI subset for selected genetic models

	n	mean	se	dif	lower	upper	p-value	AIC
Dominant								
A/A	1460	2.398	0.0404	0			0.5881	6291
C/A-C/C	246	2.463	0.0912	0.0570	-0.1492	0.2632		
Recessive								
A/A-C/A	1702	2.402	0.0369	0			0.0052	6283
C/C	4	4.5	1.3229	2.1321	0.6377	3.6264		
Overdominant								
A/A-C/C	1464	2.404	0.0405	0			0.8741	6291
C/A	242	2.43	0.0891	0.0168	-0.1909	0.2245		
log-Additive								
0,1,2				0.0927	-0.1084	0.2938	0.3666	6290

n- sample size, se – standard error, dif- means difference, lower/upper 95% Confidence Interval for dif, AIC- Akaike information criterion

The fifth RS in the NCI subset study analyzed that passed the nominal p.value threshold for the recessive model was rs16910526. Based on AIC values the recessive model was the best fitted (from statistically significant models) model for this dataset **Table 24**. This SNP is located in *CLEC7A* gene on chromosome 12. This SNP followed Hardy-Weinberg equilibrium (p=0.07). The mean value of NCI in the C/C (minor allele homozygote) group was 4.5 children (sd= 2.65, n =4). The mean value of NCI in the A/A-C/A (major allele homozygote, heterozygote) group was 2.4 children (sd=1.52, n =1702). The mean value of NCI was significantly higher in the minor allele homozygote group than in the major allele homozygote + heterozygote group (Cohen’s d = 0.97, large effect size).

Table 25. Detailed rs7447815 association analysis with NCI subset for selected genetic models

	n	mean	se	dif	lower	upper	p-value	AIC
Dominant								
C/C	693	2.505	0.0612	0			0.0285	6290
G/C-G/G	1014	2.342	0.0460	-0.1648	-0.3120	-0.0175		
Recessive								
C/C-G/C	1486	2.445	0.0407	0			0.0122	6288
G/G	221	2.163	0.0809	-0.2755	-0.4908	-0.0602		
Overdominant								
C/C-G/G	914	2.422	0.0506	0			0.6375	6294
G/C	793	2.392	0.0543	-0.0349	-0.1801	0.1103		
log-Additive								
0,1,2				-0.1542	-0.2609	-0.0476	0.0046	6286

n- sample size, se – standard error, dif- means difference, lower/upper 95% Confidence Interval for dif, AIC- Akaike information criterion

The sixth RS in the NCI subset study analyzed that passed the nominal p.value threshold for the recessive model was rs7447815. The best fitted model, based on AIC values, was the recessive model **Table 25**. This SNP is located in *SLC6A18* gene. This SNP followed Hardy-Weinberf equilibrium (p= 0.87) The mean value of NCI in the G/G (minor allele homozygote) group was 2.163 children (sd= 1.2, n =221). The mean value of NCI in the C/C-G/C (major allele homozygote, heterozygote) group was 2.445 children (sd=1.563, n =1486). The mean value of NCI was slightly lower in the minor homozygote group than in the major allele homozygote + heterozygote group (Cohen's d= 0.2, very small effect size).

Table 26. Detailed rs1861050 assocaction analysis with NCI subset for selected genetic models

	n	mean	se	dif	lower	upper	p-value	AIC
Dominant								
C/C	420	2.236	0.0657	0			0.0035	4631
T/C-T/T	842	2.495	0.0548	0.2647	0.0876	0.4418		
Recessive								
C/C-T/C	856	2.343	0.0496	0			0.0257	4634
T/T	406	2.547	0.0816	0.2039	0.0250	0.3828		
Overdominant								
C/C-T/T	826	2.389	0.0524	0			0.4822	4639
T/C	436	2.447	0.0737	0.0633	-0.1131	0.2397		
log-Additive								
0,1,2				0.1576	0.0545	0.2607	0.0028	4631

n- sample size, se – standard error, dif- means difference, lower/upper 95% Confidence Interval for dif, AIC- Akaike information criterion

The first RS from the previous NCI study (from the Manhattan plot in **Figure 23)** that passed Bonferroni-corrected p.value threshold in the previous analysis was rs1861050. In the subset analysis the dominant model, with the lowest AIC score, passed the nominal p.value threshold **Table 26**. This SNP in the subset analysis didn't follow Hardy-Weinberg equilibrium ($p < 0.05$). The mean value of NCI in the C/C (major allele homozygote) group was 2.236 children (sd=1.34, n = 420). The mean value of NCI in the T/T-C/T (minor allele homozygote, heterozygote) group was 2.49 children (sd=1.57, n =842). The mean value of NCI was slightly higher in the T/T-C/T group than in the major allele homozygote group (Cohen's d = 0.178, very small effect size).

Table 27. Detailed rs4788587 association analysis with NCI subset for selected genetic models

	n	mean	se	dif	lower	upper	p-value	AIC
Dominant								
G/G	1163	2.447	0.0447	0			0.1800	6280
A/G-A/A	540	2.331	0.0661	-0.1067	-0.2626	0.0492		
Recessive								
G/G-A/G	1647	2.419	0.0380	0			0.2332	6280
A/A	56	2.161	0.1524	-0.2476	-0.6546	0.1593		
Overdominant								
G/G-A/A	1219	2.434	0.0433	0			0.3619	6281
A/G	484	2.351	0.0715	-0.0749	-0.2357	0.0860		
log-Additive								
0,1,2				-0.1057	-0.2397	0.0283	0.1223	6280

n- sample size, se – standard error, dif- means difference, lower/upper 95% Confidence Interval for dif, AIC- Akaike information criterion

The second RS that passed the Bonferroni-corrected p.value threshold in the previous (whole dataset with adjustments) analysis was rs4788587. In the subset analysis there were no statistically significant associations between selected RS and number of children born for any genetic models analyzed **Table 27**.

All of the raw p.value data used to generate **Figure 234** is available as the excel spreadsheet “NCI_ANALYSIS_SUB” on CD.

4.4 Association analysis- Body Mass Index as a continuous variable

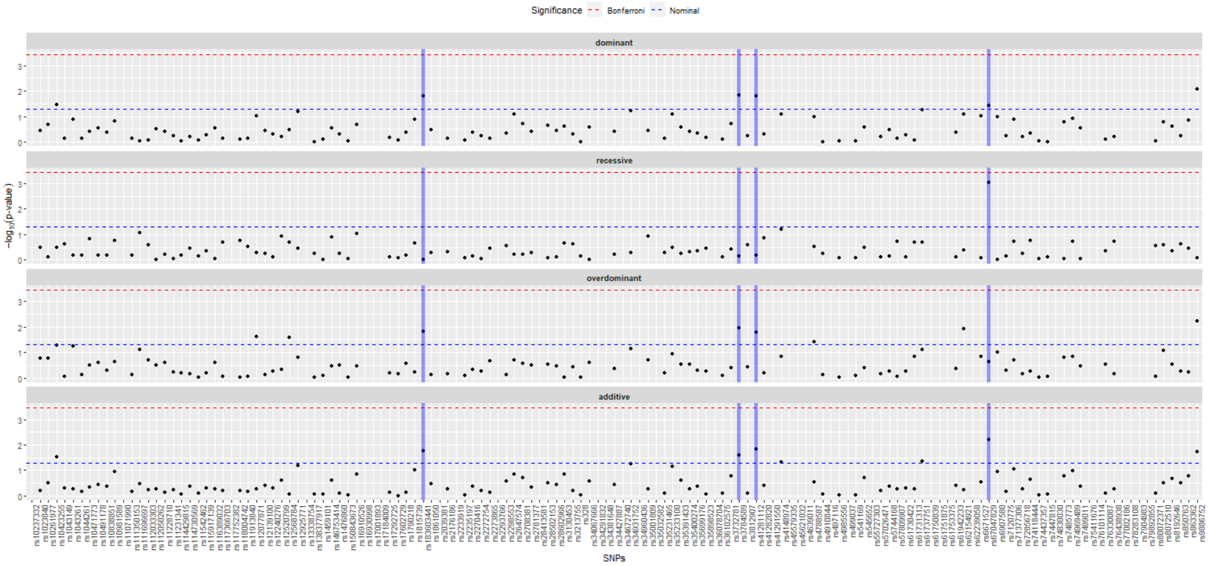

Figure 25. Manhattan-type plot of association p-values between selected PTCs with body mass index for different genetic models

The analysis presented in **Figure 25** shows that, for the continuous BMI study, 4 SNPs passed the nominal p.value threshold (p<0.05) for either recessive or dominant models (blue stripes). A more detailed analysis for these 4 RSs was conducted. A high-resolution graph is available as an electronic version on the attached cd (named Figure 25).

Table 28. Detailed rs183603441 association analysis with BMI for selected genetic models

	n	mean	se	dif	lower	upper	p-value	AIC
Dominant								
T/T	4931	25.35	0.0648	0			0.0153	28870
A/T-A/A	114	24.65	0.4261	-0.9724	-1.7580	-0.1871		
Recessive								
T/T-A/T	5044	25.33	0.0641	0			0.9901	28876
A/A	1	27.72	0.0000	0.0526	-8.2420	8.3467		
Overdominant								
T/T-A/A	4932	25.35	0.0648	0			0.0148	28870
A/T	113	24.62	0.4290	-0.9813	-1.7700	-0.1926		
log-Additive								
0,1,2				-0.9469	-1.7220	-0.1718	0.0167	28870

n- sample size, se – standard error, dif- means difference, lower/upper 95% Confidence Interval for dif, AIC- Akaike information criterion

The first RS in the continuous BMI study that passed the nominal p.value threshold for the dominant model was rs183603441. Based on AIC values the dominant model was one of the best models for this dataset **Table 28**. This SNP is located in *HYKK* gene on chromosome 15. This SNP followed Hardy-Weinberg equilibrium (p=0.49). The mean value of BMI in the T/T (major allele homozygote) group was 25.35 kg/m^2 (sd=4.55, n =4931). The mean value of BMI in the A/T-A/A (heterozygote, minor allele homozygote) group was 24.65 kg/m^2 (sd=4.47, n=114). The mean value of BMI was slightly lower in the heterozygote + minor allele group than in the major allele homozygote (Cohen's d = 0.155, very small effect size).

Table 29. Detailed rs3784589 association analysis with BMI for selected genetic models

	n	mean	se	dif	lower	upper	p-value	AIC
Dominant								
C/C	4233	25.27	0.0694	0			0.0151	28896
A/C-A/A	817	25.71	0.1640	0.3926	0.0760	0.7092		
Recessive								
C/C-A/C	5021	25.34	0.0643	0			0.7297	28902
A/A	29	24.68	0.6491	-0.2722	-1.8161	1.2718		
Overdominant								
C/C-A/A	4262	25.26	0.0691	0			0.0112	28895
A/C	788	25.75	0.1683	0.4161	0.0949	0.7374		
log-Additive								
0,1,2				0.3424	0.0423	0.6425	0.0254	28897

n- sample size, se – standard error, dif- means difference, lower/upper 95% Confidence Interval for dif, AIC- Akaike information criterion

The second RS in the continuous BMI study analyzed that passed the nominal p.value threshold for overdominant, log-additive and dominant models was rs3784589. Based on AIC values, the overdominant model was the best fitted model for this dataset

Table 29. This SNP is located in *TRPM1* gene on chromosome 15. This SNP followed Hardy-Weinberg equilibrium (p=0.23). The mean value of BMI in the A/C (heterozygote) group was 25.75 kg/m^2 (sd = 4.74, n=788). The mean value of BMI for the C/C-A/A (major allele homozygote, minor allele homozygote) group was 25.26 kg/m^2 (sd=4.51, n=4262). The mean value of BMI was slightly higher in the A/C (heterozygote) group than in the C/C-A/A group (Cohen’s d = 0.106, very small effect size).

Table 30.Detailed rs41281112 association analysis with BMI for selected genetic models

	n	mean	se	dif	lower	upper	p-value	AIC
Dominant								
C/C	4874	25.31	0.0651	0			0.0153	28857
T/C-T/T	169	26.02	0.3551	0.8022	0.1539	1.4510		
Recessive								
C/C-T/C	5041	25.34	0.0641	0			0.6476	28863
T/T	2	26.49	0.6670	1.3677	-4.4968	7.2320		
Overdominant								
C/C-T/T	4876	25.31	0.0651	0			0.0170	28857
T/C	167	26.02	0.3593	0.7946	0.1425	1.4470		
log-Additive								
0,1,2				0.7906	0.1536	1.4280	0.0150	28857

n- sample size, se – standard error, dif- means difference, lower/upper 95% Confidence Interval for dif, AIC- Akaike information criterion

The third RS in the continuous BMI study analyzed that passed the nominal p.value threshold for dominant and overdominant models was rs41281112. The best genetic model couldn't be determined with AIC values and therefore the dominant model will be described **Table 30**. This SNP is located in *CLYBL* gene on chromosome 13. This SNP followed Hardy-Weinberg equilibrium. The mean value of BMI in the C/C (major allele homozygote) group was 25.31 kg/m^2 (sd = 4.54, n =4874). The mean value of BMI in the T/C-T/T (heterozygote, minor allele homozygote) group was 26.02 kg/m^2 (sd=4.6, n =169). The mean value of BMI was slightly higher in the group with at least one minor allele (Cohen's d = 0.15, very small effect size).

Table 31. Detailed rs67047829 association analysis with BMI for selected genetic models

	n	mean	se	dif	lower	upper	p-value	AIC
Dominant								
G/G	4159	25.4	0.0710	0			0.0376	28898
A/G-A/A	891	25.03	0.1473	-0.3246	-0.6305	-0.0187		
Recessive								
G/G-A/G	4994	25.36	0.0644	0			0.0009	28891
A/A	56	23.35	0.4866	-1.8807	-2.9935	-0.7679		
Overdominant								
G/G-A/A	4215	25.38	0.0704	0			0.2300	28900
A/G	835	25.14	0.1530	-0.1923	-0.5063	0.1217		
log-Additive								
0,1,2				-0.3884	-0.6674	-0.1094	0.0064	28894

n- sample size, se – standard error, dif- means difference, lower/upper 95% Confidence Interval for dif, AIC- Akaike information criterion

The last RS in the continuous BMI study analyzed that passed the nominal p.value threshold for recessive and additive model was rs67047829. Based on AIC values, the recessive model was the best model for this dataset **Table 31**. This SNP is located in *ERV3-1* gene on chromosome 7. This SNP followed Hardy-Weinberg equilibrium (p=0.068). The mean value of BMI in the A/A (minor allele homozygote) group was 23.35 kg/m^2 (sd=3.64; n = 56). The mean value of BMI in the G/G-A/G group was 25.36 kg/m^2 (sd=4.55, n=4994). The mean value of BMI was slightly lower in the minor allele homozygote group (Cohen’s d 0.48, small effect size)

All of the raw p.value data used to generate **Figure 25** is available as the excel spreadsheet “BMI_ANALYSIS” on CD.

4.5 Post-hoc association analysis – BMI qualitative variable (BMI group)

Table 32. Association analysis between BMI group and PTC SNPs. Post-hoc analysis compares the normal weight group with the obesity group

RS name	Genotype	under weight	normal weight	overweight	obesity	Statistics	Post Hoc analysis	Quantitative analysis
rs114429815	T/T	67	611	470	166	Fisher test Dominant model p= 0.0497	OR= 1.17, 95% CI:[0.95, 1.44], p = 0.148	p.value dominant model = 0.95
	C/T - C/C	91	1230	826	391			
rs3732781	A/A	87	1289	797	366	Fisher test Dominant model p = 0.0097	OR= 1.12, 95% CI:[0.95, 1.32], p = 0.18	p.value dominant model = 0.20
	C/A-C/C	98	1182	898	376			
rs7447815	C/C	84	1029	644	319	Fisher test Dominant model p = 0.0226	OR= 0.94, 95% CI:[0.79, 1.12], p = 0.18	p.value dominant model = 0.97
	G/C-G/G	101	1442	1052	422			
rs6907580	G/G	167	2216	1532	686	Cochran Armitage Trend test Dominant model p=0.0489	OR= 0.71, 95% CI:[0.51, 0.96], p = 0.024	p.value dominant model = 0.10
	A/G-A/A	18	256	164	56			
rs67047829	G/G-A/G	181	2436	1682	740	Fisher test Recessive model p=0.0072 Cochran Armitage Trend test Recessive model p=0.0011	OR= 0.183, 95% CI:[0.021, 0.713], p = 0.0059	p.value recessive model = 0.00093
	A/A	4	36	14	2			
rs138377917	G/G	166	2228	1529	666	Fisher test Dominant model p = 0.00418	OR= 1.096, 95% CI:[0.61, 1.88], p = 0.78	p.value dominant model = 0.97
	A/G-A/A	13	58	33	19			

rs35898523	G/G-T/G	185	2464	1685	737	Cochran Armitage Trend test Recessive model p=0.0269	OR= 2.786, 95% CI:[0.67, 10.99], p = 0.141	p.value recessive model = 0.35
	T/T	0	6	11	5			
rs9886752	G/G	150	1927	1320	544	Fisher test Dominant model p=0.0284 Cochran Armitage Trend test Dominant model p=0.0112	OR= 1.29, 95% CI:[1.06, 1.565], p = 0.008	p.value dominant model = 0.82
	A/G-A/A	35	543	374	198			
rs35233100	C/C-T/C	184	2457	1695	739	Fisher test Dominant model p= 0.0178	OR= 0.66, 95% CI:[0.12, 2.36], p = 0.78	p.value dominant model = 0.085
	T/T	1	15	1	3			
rs1790218	A/A-G/A	137	1932	1380	567	Fisher test Recessive model p= 0.0072	OR= 1.19, 95% CI:[0.95, 1.48], p = 0.127	p.value recessive model = 0.68
	G/G	43	396	251	138			
rs16910526	A/A-C/A	181	2462	1690	739	Fisher test Recessive model p= 0.024	OR= 0.99, 95% CI:[0.17, 3.895], p = 1	p.value recessive model = 0.09
	C/C	4	10	5	3			
rs71377306	C/C-T/C	183	2459	1689	733	Cochran Armitage Trend test Recessive model p=0.0472	OR= 3.35, 95% CI:[0.77, 14.61], p = 0.057	p.value recessive model = 0.19
	T/T	0	5	4	5			
rs118004742	T/T-G/T	183	2455	1691	741	Cochran Armitage Trend test Recessive model p=0.0439	OR= 0, 95% CI:[0, 2.3], p = 0.36	p.value recessive model = 0.18
	G/G	2	7	4	0			
rs61737751	C/C	175	2284	1540	674	Cochran Armitage Trend test Dominant model p=0.0267	OR= 1.22, 95% CI:[0.90, 1.65], p = 0.189	p.value recessive model = 0.20
	T/C-T/T	10	188	156	68			

The study analyzed 141 SNPs for possible association with BMI and conducted both qualitative and quantitative analyses. The provided table summarizes the results of the qualitative analysis, which includes Fisher tests and Cochran-Armitage trend tests for recessive and dominant models, as well as post-hoc analysis comparing the normal weight group to the obesity group in order to assess the risk of obesity. **Table 33** represents only the statistically significant results.

Based on **Table 33**, the SNP rs114429815 showed a statistically significant association with BMI under a dominant model (p=0.0497), but the post hoc analysis did not show a significant difference between normal weight and obesity groups. Similarly, the SNPs rs3732781, rs7447815, rs138377917, rs35233100 showed a statistically significant association with BMI under a dominant model, and SNPs rs61737751, rs35898523, rs1790218, rs16910526, rs71377306, rs118004742, showed a statistically significant association with BMI under a recessive model, but the post hoc analysis did not reveal a significant difference between normal weight and obesity groups.

In contrast, the SNP rs6907580 showed a statistically significant association with BMI under a dominant model (p=0.0489), and a significant difference between normal weight and obesity groups in the post-hoc analysis (p=0.024), suggesting that this SNP may be a risk factor for obesity. Similarly for SNPs rs9886752 for the dominant model and rs67047829 for the recessive model. The last SNP also showed significant association analysis results for the continuous variable (p= 0.00093).

Overall, the table suggests that some SNPs may be associated with BMI and potentially play a role in obesity risk, particularly rs67047829 in the *ERV3-1* gene.

All of the raw p.value data used to generate **Table 33** is available as the excel spreadsheet “BMI_Qualitative_ANALYSIS” on CD.

5. Discussion

Recent advances in genetics have allowed for a deeper understanding of how our DNA affects various aspects of our health and well-being. One such area of interest is the association between specific genetic mutations and various health outcomes.

This present study in this thesis investigated the association between nonsense mutations and three key health outcomes: age (longevity), number of children born (fertility), and BMI (the risk of obesity). By analyzing a large dataset of genetic information and health outcomes, this might provide a better understanding of how nonsense mutations may impact these important health metrics.

The findings from this study could have important implications for both clinical and public health practice. By identifying specific genetic mutations that are associated with improved health outcomes, it might be possible to develop more personalized approaches to disease prevention and treatment. Additionally, a better understanding of the genetic factors that contribute to longevity, fertility, and obesity risk could lead to the development of more effective interventions and strategies to improve overall health and well-being.

5.1 Longevity

The first analyzed phenotype was longevity. According to a study published in the journal Nature Genetics, genetics play a significant role in determining human lifespan. The study identified several genes associated with longevity, including those involved in DNA repair, immune response, and cell signaling pathways. However, environmental and lifestyle factors can also affect the expression of these genes and thus influence lifespan [58]. Environmental factors such as exposure to pollutants, toxins, and radiation can also impact human longevity. A study published in the journal Aging Cell found that exposure to air pollution, specifically fine particulate matter, was associated with shorter telomere length, a marker of cellular aging, in older adults [59]. Other environmental factors that can affect lifespan include access to healthcare, nutrition, and social support. Lifestyle factors, such as diet, exercise, and smoking, also have a significant impact on human longevity. A study published in the journal Circulation found that maintaining a healthy lifestyle, including a healthy diet, regular exercise, not smoking, and moderate alcohol consumption, was associated with an increased lifespan of up to 7 years [60]. Overall, human longevity is a complex phenomenon influenced by a combination of genetic, environmental, and lifestyle factors.

This shows the first limitation of the present study, concerning the question of whether the longevity phenotype can be analyzed by association with the patient’s age alone. Age analysis without proper adjustments can only unequivocally demonstrate large genotype effects. Nevertheless, statistically significant results for smaller effects might potentially form the basis of a new study, which should include additional factors affecting longevity. Some studies have suggested that an association analysis with longevity should be performed between two groups: supercentenarians and a healthy, younger control group [61,62]. The original idea for the main purpose of producing the populous database was to study the obesity phenotype [36], and in this case therefore it wasn’t possible to create a proper supercentenarian group, which can be considered as a limitation of this study.

Even so, pretermination codons in the following genes might give some insight into longevity, as long as the results are treated with caution.

5.1.1 The *SULT1C3* gene

The *SULT1C3* gene encodes a member of the sulfotransferase family of enzymes, which catalyzes the transfer of a sulfate group from a donor molecule to a variety of acceptor molecules, including drugs, hormones, and neurotransmitters [63–65]. The physiological roles of SULT1C enzymes have not yet been fully elucidated, despite recent research showing that SULT1C2 and SULT1C4 are capable of catalyzing the sulphation of procarcinogenic hydroxyarylamines: N-hydroxy-2-acetylaminofluorene, resulting in the activation of their carcinogenic activity [66,67]. Furthermore, despite the fact that SULT1C2 and SULT1C4 have been partially cloned, produced, and described [67,68], the molecular identification of SULT1C3 is still unknown. It has been found that members of the SULT1C subfamily are more strongly expressed in fetal tissues than in adult tissues [69]. Freimuth et al(2004)'s computational analysis of the human genome was the first to predict the existence of *SULT1C3* in the human genome. The potential for alternative splicing in this location was highlighted by these researchers when they observed what appeared to be a duplication of two exons, which they termed exons 7 and 8 (the first designated exon in this work was exon 2) in the order exon 7a, exon 8a, exon 7b, and exon 8b. Theoretically, *SULT1C3* pre-mRNA might be alternatively spliced to produce four transcripts with exons 7a/8a, 7a/8b, 7b/8a, or 7b/8b. These transcripts could then be translated into the appropriate proteins, known as SULT1C3 isoforms a, b, c, and d, respectively [70]. Another study indicated that *SULT1C3* is expressed in intestinal tissues and cells [71]. Enzymatic analysis showed that SULT1C3d was able to sulphate a variety of substances, including bile acids, thyroid hormones, chloro phenols, and hydroxypyrenes, but *SULT1C3*a only showed mild sulphating activity toward chloro phenols[72].

The SNP rs112050262 acts as a pretermination codon in the *SULT1C3* gene and potentially reduces the length of the expressed protein from one of the transcripts by 88.2% from 305 to 36 amino acids.

According to the present study, one of the statistically significant PTCs (rs112050262) for the recessive model was found in this gene. However, 24.7% of patients had missing genotype data for this PTC. The patients with the minor allele homozygote had a slightly higher mean age value. Nevertheless, due to the high amount of missing genotype data, the ambiguous protein function, and the low effect size, this PTC is unlikely to have any clinical significance on human longevity. It is worth noting that missing genotype data can significantly impact the accuracy and reliability of genetic association studies. Therefore, the high amount of missing data for this PTC limits the study's ability to draw meaningful conclusions about its association with longevity. Furthermore, the fact that the protein function of the gene is ambiguous adds to the uncertainty surrounding the clinical significance of this PTC. Without a clear understanding of how the gene and its associated PTC influence the aging process, it is difficult to draw any conclusions about their potential clinical relevance. Finally, the study's finding of a very low effect size for this PTC further diminishes its clinical significance. While statistically significant, the small effect size suggests that this PTC is unlikely to have a significant impact on human longevity.

5.1.2 The *KIAA1755* gene

By using cDNA sequencing, the Kazusa project discovered *KIAA* genes. The HUGE (Human Unidentified Gene-Encoded large protein database) database contains the described sequences of human large cDNAs longer than 4 kb that promote the production of huge proteins (>50 kDa) [73,74]. The PTC rs1205434 in the *KIAA1755* gene was previously identified to be associated with the incidence of breast cancer in the Chinese population [75], yet the full functionality of the protein coded by this gene hasn't been described in the literature.

The SNP rs41282820 acts as a pretermination codon in the *KIAA1755* gene and potentially reduces the length of the expressed protein by 57.5 % from 1201 to 510 amino acids.

The PTC rs41282820 gave a statistically significant model in the present study. There were no missing genotype data of this RS. Given the limited sample size and ambiguous protein function of the *KIAA1755* gene, it is difficult to draw strong conclusions about the potential association between the rs41282820 PTC and longevity. While this PTC did pass the nominal p-value threshold for several models, including dominant and overdominant, the AIC values for all models were the same, indicating that it is difficult to determine which model fits the data best. Additionally, the effect size of the PTC was very small, with a Cohen's d of 0.157. These factors, combined with the small number of subjects with minor allele homozygosity, suggest that this PTC may not have a significant impact on human longevity. Therefore, while it is possible that there may be some association between this PTC and longevity, the current evidence suggests that the association is likely weak at best, and may not be clinically significant.

5.1.3 The OAS3 gene

OAS3 dsRNA-activated antiviral enzyme induced by interferon that is essential for cellular innate immunity to viruses. Moreover, it can be involved in several biological processes as apoptosis, cell development, differentiation, and gene regulation. Produces preferentially dimers of 2'-5'-oligoadenylates (2-5A) from ATP, which upon binding to the monomeric form of inactive ribonuclease L (RNase L), dimerizes and activates the enzyme. When RNase L is activated, both cellular and viral RNA are degraded. This inhibits the production of proteins, stopping viral replication. It can either use a different antiviral pathway not dependent on RNase L to produce the antiviral action or the traditional RNase L-dependent pathway [76]. The enzyme demonstrates antiviral action against the viruses Dengue, Sindbis, Chikungunya, and Semliki Forest (SFV) [77,78].

The SNP rs61942233 acts as a pretermination codon in this gene and potentially reduces the length of the expressed protein by 22.4 % from 1088 to 844 amino acids.

The OAS3 gene PTC rs61942233 was found to be statistically significant for the dominant and log-additive models in this study. However, the minor allele homozygote group had only one subject, which limits the statistical power of the analysis. The mean age of subjects with at least one minor allele was slightly higher than that of major allele homozygotes, but the effect size was small. Furthermore, the study has several limitations, such as the lack of a supercentenarian group and the small sample size of the minor allele homozygote group. Additionally, the functional role of the OAS3 protein is complex and not fully understood, and it is involved in various biological processes, such as apoptosis, cell development, differentiation, and gene regulation. While the enzyme produced by OAS3 is known to have antiviral action against several viruses, including Dengue, Sindbis, Chikungunya, and Semliki Forest, it is unclear whether the PTC identified in this study has any direct effect on the antiviral activity of OAS3. It is also worth noting that the viruses mentioned in the gene summary are not normally present in Poland, where the study was conducted (although in the past these might have affected an ancestral population). Taken together, while the finding of a statistically significant association between the OAS3 PTC and longevity phenotype is interesting, the serious limitations of the study with respect to longevity and the ambiguous protein function of OAS3 suggest that further research is needed to confirm and clarify the potential clinical significance of this association.

5.1.4 *The TAAR2* gene

The name of the gene *TAAR2* refers to trace amine associated receptor 2, earlier called GPR58 (G-protein coupled receptor). G protein-coupled receptors (GPCRs, or GPRs) include seven transmembrane domains and use heterotrimeric G proteins to transmit extracellular signals. The sequences of a human cerebellum cDNA encoding phBL5, also known as GPR58, and a rabbit smooth muscle cDNA encoding GPR58 were taken from the patent literature by Lee et al. in 2000. Using genomic DNA, they extracted the whole human GPR58 coding region [79]. Lindemann et al. (2005) discovered a long isoform of *TAAR2* by screening the genomic sequence using a nonredundant list of all vertebrate G protein-coupled receptors as queries [80]. The functionality of the protein coded by this gene, has not yet been covered by any study, although predicted functionality indicates that the TAAR2 protein enables trace-amine receptor activity, is involved in G protein-coupled receptor signaling pathway and is located in plasma membrane [https://www.alliancegenome.org].

The SNP rs8192646 acts as a pretermination codon in this gene and potentially reduces the length of the expressed protein by 52.3 % from 352 to 168 amino acids.

Based on the results of this study, the *TAAR2* gene may be associated with human longevity. The PTC rs8192646 located in the *TAAR2* gene on chromosome 6 was found to be statistically significant in the dominant model, with a large effect size (Cohen's d=0.94). The minor allele homozygote (T/T) group had a significantly higher mean age compared to the major allele homozygote (C/C) and heterozygote (T/C) group. The fact that the genotype missing data was only 0.1% is a positive aspect of this study, as it indicates that the results are likely reliable. Given the lack of information on the gene function and the small sample size, we should be cautious in interpreting the results. Nonetheless, the large effect size of this PTC could suggest a potential clinical significance for further investigation.

5.1.5 The *GBGT1* gene

The *GBGT1* gene is responsible for encoding a glycosyltransferase involved in synthesizing the Forssman glycolipid (FG), which is a member of the globoseries glycolipid family. FG, and other glycolipids like it, are attachment sites for pathogens to bind to cells, and the expression of this protein may determine host tropism to microorganisms. The full name of the gene is globoside alpha-1,3-N-acetylgalactosaminyltransferase 1 (FORS blood group), and mutations in this gene may result in the loss of the ability to synthesize the Forssman glycolipid antigen (FORS1/FG) [81].

The SNP rs35898523 acts as a pretermination codon in this gene and potentially reduces the length of the expressed protein by 65.2 % from 348 to 121 amino acids.

The PTC rs35898523 in the *GBGT1* gene showed a statistically significant association with age in a recessive model analysis. The effect size was medium, indicating a potential clinical significance. However, it should be noted that the sample size in the minor allele homozygote group was small, with only 22 subjects. Additionally, this study did not include a supercentenarian group, which limits the conclusions that can be drawn regarding the role of this PTC in human longevity. Despite these limitations, the findings suggest that the *GBGT1* gene may play a role in human aging and disease susceptibility, and further research is warranted to explore this potential association.

5.2 Fertility

Human fertility, the ability to conceive and reproduce offspring, can be influenced by various factors. These factors can be broadly classified into biological, environmental, and lifestyle factors.

Fertility decreases with age, especially for women. According to a study published in the journal Obstetrics and Gynecology, the probability of conceiving in any given menstrual cycle declines from 25% for women in their early 20s to 5% for women in their 40s [82]. Certain health conditions, such as polycystic ovary syndrome (PCOS), endometriosis, and thyroid disorders, can affect fertility. According to a study published in the Journal Fertility and Sterility, PCOS is the most common endocrine disorder affecting women of reproductive age and is a leading cause of infertility [83]. Lifestyle factors such as smoking, excessive alcohol consumption, and obesity can also impact fertility [84,85]. According to a study published in the journal Fertility and Sterility and study published in Human Reproduction update, smoking can decrease both male and female fertility, and quitting smoking can improve fertility outcomes [84,86]. Another example, a diet rich in fruits, vegetables, and whole grains can improve fertility, while a high intake of processed and fast foods can decrease fertility [87]. High levels of stress can impact fertility, although the exact mechanism is not fully understood [88]. Exposure to environmental toxins and pollutants can also impact fertility. For example, a study published in the journal Human Reproduction found that exposure to bisphenol A (BPA) can reduce female fertility [89], and another study indicates exposure on certain pesticides can affect male fertility [90]. Genetic factors can also play a role in fertility. Certain genetic mutations can affect male fertility, such as chromosomal abnormalities or mutations in genes involved in reproductive function [91].

The following genes might be associated with fertility and have therefore been analyzed.

5.2.1 The *MROH2B* gene

There is limited information available about the *MROH2B* (human) gene, and its exact function is not well understood. *MROH2B* is a member of the *MROH* (Maestro Heat-like repeat containing protein) family of genes. The predicted protein product of *MROH2B* contains two Maestro heat-like repeats and a C-terminal transmembrane domain, suggesting that it may be a membrane-bound protein [https://www.alliancegenome.org]. There is currently no known association between mutations in the *MROH2B* gene and any specific disease or disorder.

The SNP rs1023840 acts as a pretermination codon in the *MROH2B* gene, which is located on Chromosome 5, and potentially reduces the length of the expressed protein by 88 % from 1586 to 191 amino acids.

Rs1023840 passed the nominal p.value threshold for the recessive model. While the mean value of the number of children born was slightly lower in the minor homozygote group than in the major homozygote and heterozygote group, the effect size was very small (Cohen's d= 0.12). It is also important to note that the analysis was done with age, sex and district adjustments, which may not be enough for the association analysis of such a complex phenotype like fertility. Given the limited information available about the *MROH2B* gene, the small effect size observed in the association analysis, and that the findings weren't confirmed in the subset analysis, it is unlikely that the rs1023840 PTC in the *MROH2B* gene has a significant impact on human fertility.

5.2.2 The *KRT83* gene

The *KRT83* gene encodes a type II hair keratin, which is a protein component of hair fibers. Specifically, *KRT83* encodes the keratin KRT83B, which is expressed in the cuticle layer of the hair shaft and plays a crucial role in hair formation and maintenance. Studies have shown that mutations in the *KRT83* gene can lead to hair disorders, such as monilethrix, which is characterized by hair fragility and abnormal hair growth patterns [92,93]. Furthermore, *KRT83* has been identified as a potential biomarker for hair regeneration. In a study by Sennett et al. (2015), *KRT83* was found to be one of the most highly upregulated genes during hair regeneration in mice, suggesting a potential role for this gene in promoting hair growth and maintenance [94].

The SNP rs146753414 acts as a pretermination codon in this gene and potentially reduces the length of the expressed protein by 59.3 % from 494 to 201 amino acids.

While the *KRT83* gene has been shown to play an important role in hair formation and maintenance, there is currently no known association between mutations in this gene and fertility. Furthermore, it is important to note that nonsense mutations rarely follow a dominant genetic model. Therefore, the finding that rs146753414 in *KRT83* follows a dominant genetic model may not be biologically meaningful. Additionally, the effect size of this PTC on the number of children born was very small (Cohen's d = 0.27)- for the whole group, and small (Cohen's d=0.42)- for the subset data, which suggests that it is unlikely to have a significant impact on human fertility. Finally, it is important to consider that the analysis was done with adjustments for age, sex, and district, which may not be sufficient for a complex phenotype like fertility. Therefore, more research is needed to fully understand the potential impact of the *KRT83* gene on human fertility.

5.2.3 The *MADD* gene

The *MADD* gene (also known as *IG20* or MAP-kinase activating death domain) encodes a protein that plays a role in various cellular processes, including cell proliferation, apoptosis, and differentiation. The *MADD* protein is a multifunctional adaptor protein that contains several functional domains, including a death domain, a proline-rich domain, and a C-terminal SH3 domain [95]. One of the primary functions of the *MADD* protein is to regulate the mitogen-activated protein kinase (MAPK) signaling pathway, which is involved in the regulation of various cellular processes, including cell growth, differentiation, and apoptosis. The *MADD* protein can activate MAPKs by binding to and activating the upstream kinases MEKK1 and MEKK4 [95]. Additionally, the *MADD* protein has been implicated in the regulation of insulin signaling and glucose metabolism. Studies have shown that *MADD* can interact with the insulin receptor and insulin receptor substrate-1 (IRS-1) and regulate insulin-mediated glucose uptake in adipocytes [96].

The SNP rs35233100 acts as a pretermination codon in this gene and potentially reduces the length of the expressed protein by 53.5 % from 1648 to 766 amino acids.

The results from this study suggest that the rs35233100 PTC in the *MADD* gene may be associated with a slightly higher mean value of number of children born in minor allele homozygotes than in major homozygotes and heterozygotes, based on a very small effect size (Cohen’s d= 0.162). However, these results were not confirmed in the subset analysis, and the potential role of this gene on human fertility is most likely not clinically significant.

5.2.4 The *CC2D2A* gene

The *CC2D2A* gene encodes a protein called Coiled-coil and C2 domain-containing protein 2A. This protein is involved in the development and function of cilia, which are hair-like structures on the surface of cells that play important roles in sensing the environment and transmitting signals.

Mutations in the *CC2D2A* gene have been linked to several ciliopathies, which are genetic disorders caused by defects in cilia structure or function. These include Joubert syndrome, a rare neurological disorder characterized by developmental delay, abnormal eye movements, and breathing abnormalities, as well as Meckel-Gruber syndrome, a lethal condition characterized by brain malformations, kidney cysts, and other abnormalities [97–99]. The exact role of the CC2D2A protein in cilia function is not fully understood, but studies suggest that it may be involved in regulating the formation and maintenance of the ciliary membrane and the transport of proteins and other molecules into and out of the cilium [100].

The SNP rs1861050 acts as a pretermination codon in the *CC2D2A* gene on chromosome 4 and potentially reduces the length of the expressed protein by 28.5 % from 123 to 88 amino acids.

In the whole group analysis, the PTC rs1861050 p.value was statistically significant after Bonferroni correction for the dominant model. The mean value of NCI was slightly higher in the T/C-T/T group than in the C/C group, with a very small effect size. However, this PTC did not follow Hardy-Weinberg equilibrium ($p < 0.05$), and there is a high rate of missing genotype data. In the subset analysis, rs1861050 also passed nominal p.value threshold for the dominant model, but it also did not follow Hardy-Weinberg equilibrium ($p < 0.05$). The mean value of NCI was slightly higher in the T/T-C/T group than in the C/C group, with a very small effect size. It is important to note that there is a high rate of missing genotype data, over 25%, and the frequency of the T allele in the study is 50%, which is higher than the frequency of the T allele in the world population (5-10%, according to NCBI). In conclusion, the association analysis for rs1861050 in the *CC2D2A* gene showed a slightly higher mean value of NCI in the T/C-T/T or T/T-C/T groups compared to the C/C group, with very small effect sizes in both the whole group and subset analyses. However, the high rate of missing genotype data and departure from Hardy-Weinberg equilibrium raise limitations in the interpretation of the results. Furthermore, the high frequency of the T allele in the study population may suggest a possible founder effect or genetic drift, and caution should be taken when generalizing the findings to other populations.

5.2.5 The *PKD1L3* gene

The *PKD1L3* gene is a gene that encodes for a protein called polycystin 1 like 3. This protein interacts with transient receptor potential (TRP) ion channel proteins, which are involved in taste transduction. Ishimaru et al. hypothesized that TRP ion channel proteins other than TRPM5 might be expressed in taste cells, and using in situ hybridization, they detected expression of *PKD1L3* in mouse circumvallate taste cells. Confocal microscopy and coimmunoprecipitation analysis revealed that *PKD1L3* is coexpressed with PKD2L1, but not TRPM5, in the apical ends of taste cells in circumvallate and foliate papillae. Coexpression of *PKD1L3* and *PKD2L1* was necessary for inducing changes in intracellular calcium concentration in response to acid solutions, suggesting that *PKD1L3*/PKD2L1 heteromers function as sour taste receptors. PKD2L1 was also expressed in taste cells in other areas of the tongue and palate [101,102].

The SNP rs4788587 acts as a pretermination codon in this gene and potentially reduces the length of the expressed protein by 54.5 % from 1733 to 789 amino acids.

However, despite the interesting findings on *PKD1L3*'s role in taste transduction, the genetic association analysis performed here found no significant associations between selected pretermination codons (PTCs) from single nucleotide polymorphisms (SNPs) in the *PKD1L3* gene and the number of children per individual. While one PTC (rs4788587) passed the Bonferroni correction p-value threshold for the dominant model, the effect size was very small (Cohen's d=0.13) and the mean value of NCI was only slightly lower in the A/G-A/A (heterozygote, minor allele homozygote) group compared to the G/G (major allele homozygote) group. Furthermore, in subset analysis, no statistically significant associations were found between the selected PTCs and number of children per individual for any genetic models analyzed. These limitations suggest that the *PKD1L3* gene may not have a significant role in fertility or reproductive success.

5.2.6 The *PKD1L2* gene

PKD1L2 is a gene that encodes for a transmembrane protein which belongs to the polycystin protein family. These proteins play important roles in the development of cilia and maintaining calcium homeostasis in renal tubular cells. *PKD1L2* interacts with GNAS and GNAI1 and is thought to function as a G-protein-coupled component or regulator of cation channel pores [103]. The long isoform of this protein contains 11 transmembrane domains, a GPS domain, and a PLAT domain. Knockdown of *PKD1L2* inhibits HIV-1 replication in HeLa-derived TZM-bl cells [104]. Additionally, a copy number variation in *PKD1L2* has been linked to an increased predisposition to colorectal cancer in the Korean population [105].

The SNP rs12925771 acts as a pretermination codon in this gene and potentially reduces the length of the expressed protein by 49.5 % from 806 to 407 amino acids.

There are some limitations to consider in the interpretation of the findings related to the *PKD1L2* gene. Firstly, while the PTC rs12925771 in the *PKD1L2* gene was found to be associated with differences in NCI, its effect size was very small, additionally the result was only present in the subset study.

Therefore, caution should be exercised when interpreting the clinical significance of this finding, as the overall impact of this variant on the regulation of NCI is likely to be minimal.

5.2.7 The *ZNF883* gene

Research on *ZNF883* is still in its early stages, and there is limited information available on its specific functions and roles in human biology. The *ZNF883* protein is predicted to have DNA-binding and transcription factor activities specific to RNA polymerase II and cis-regulatory DNA sequences, and is also predicted to play a role in regulating transcription by RNA polymerase II in the nucleus [https://www.alliancegenome.org].

The SNP rs10981589 acts as a pretermination codon in this gene and potentially reduces the length of the expressed protein by 10.3 % from 380 to 341 amino acids.

The limitations of this study include the fact that the research on *ZNF883* is still in its early stages and its specific functions and roles in human biology are not yet well understood. Additionally, the sample size of the minor allele homozygote group in this study was very small (only 5 subjects), which may limit the generalizability of the results. Furthermore, the NCI phenotype measured in this study may not be directly related to the function of *ZNF883*, and additional research is needed to establish a clear connection between this gene and reproductive outcomes. Finally, although the p-value for the association between rs10981589 and NCI passed the nominal threshold, it did not meet the Bonferroni correction threshold for multiple testing, indicating the possibility of false positive results. Therefore, further studies are needed to confirm these findings and establish the clinical significance of this gene in relation to fertility.

5.2.8 The *CLEC7A* gene

The *CLEC7A* gene encodes the protein Dectin-1, which belongs to the C-type lectin family and is involved in the recognition of fungal pathogens. Dectin-1 is expressed on the surface of various immune cells, including dendritic cells, macrophages, and neutrophils, and plays a critical role in the initiation of the innate immune response to fungal infections [106]. Several studies have investigated the structure, function, and regulation of the *CLEC7A* gene and its protein product, Dectin-1. For example, a study by Brown et al. (2003) identified the *CLEC7A* gene and demonstrated that it encodes a type II transmembrane protein with a single extracellular C-type lectin-like domain. In addition, the study showed that Dectin-1 binds to β-glucans, a component of the fungal cell wall, and triggers downstream signaling pathways that activate innate immune responses [106]. Another study by Goodridge et al. (2011) investigated the regulation of *CLEC7A* gene expression in human dendritic cells. The study showed that the expression of *CLEC7A* and Dectin-1 is induced by the fungal pathogen

Candida albicans through the activation of the transcription factor NF-κB. The study also showed that the expression of *CLEC7A* and Dectin-1 is regulated by multiple microRNAs, which play a role in the modulation of the innate immune response [107]. However, there is no direct evidence linking the known function of this gene to fertility.

The SNP rs16910526 acts as a pretermination codon in this gene and potentially reduces the length of the expressed protein by 5 % from 202 to 192 amino acids.

The present study in this thesis analyzed the association between the PTC rs16910526 in the *CLEC7A* gene and fertility. The mean NCI was significantly higher in the minor allele homozygote group than in the major allele homozygote + heterozygote group, with a large effect size (Cohen's d = 0.97). However, it's important to note that the p-value for this association did not pass Bonferroni correction, and caution should be exercised when drawing conclusions. The nominal p-value threshold was used to identify the association, and this could result in false positives. Also, the study has some limitations, such as a relatively small sample size in the minor allele homozygote group, and a lack of enough adjustment for potential confounding factors. Therefore, further research is needed to validate these findings and investigate the possible clinical significance of the association between PTC rs16910526 in the *CLEC7A* gene and fertility.

5.2.9 The *SLC6A18* gene

The *SLC6A18* gene encodes a member of the solute carrier family 6, which is a group of sodium- and chloride-dependent neurotransmitter transporters. The SLC6A18 transporter, also known as Xtrp2, act as specific transporter of amino acids, neurotransmitters, and osmolytes like betaine, taurine, and creatine across the plasma membrane of epithelial cells in the intestine and kidney [108]. Mutations in the *SLC6A18* gene have been linked to iminoglycinuria and hyperglycinuria phenotypes [109].

Another statistically significant association result in the NCI subset study was the PTC rs7447815, found in the *SLC6A18* gene. The SNP rs7447815 acts as a pretermination codon in this gene and potentially reduces the length of the expressed protein by 49.3 % from 629 to 319 amino acids.

The mean value of NCI in the minor allele homozygote group was slightly lower than in the major allele homozygote and heterozygote group, with a very small effect size of Cohen's d = 0.2. This small effect size, combined with other limitations such as the failure to pass the Bonferroni correction and the absence of a proven connection between *SLC6A18* gene function and fertility, makes it difficult to draw any firm conclusions from this study regarding the relationship between this PTC and fertility.

5.3 Obesity

Obesity is a complex condition that results from a combination of genetic, environmental, and behavioral factors. A person's genetic makeup can increase their susceptibility to obesity. According to a study published in the New England Journal of Medicine, the heritability of obesity is estimated to be 40-70% [110]. Mutations in the *FTO, MC4R, TMEM18, SH2B1, PCSK1* genes have been associated with increased risk of obesity [111–115]. Environmental factors such as access to high-calorie foods, sedentary lifestyle, and lack of physical activity can contribute to obesity. A diet high in calories, sugar, and saturated and trans fats can increase the risk of obesity. According to a study published in the American Journal of Clinical Nutrition, a high intake of sugar-sweetened beverages is associated with an increased risk of obesity [116]. Lack of physical activity is a major risk factor for obesity. According to a study published in the International Journal of Epidemiology, physical inactivity is responsible for approximately 30% of obesity cases [117]. Certain medical conditions such as hypothyroidism, Cushing's syndrome, and polycystic ovary syndrome can increase the risk of obesity. According to a study published in the Journal of Clinical Endocrinology and Metabolism, approximately 10% of obesity cases are due to an underlying medical condition [118]. Certain medications such as corticosteroids, antidepressants, and antipsychotics can cause weight gain and increase the risk of obesity. According to a study published in the Journal of Clinical Psychopharmacology, approximately 20-30% of individuals taking antipsychotic medications experience significant weight gain [119].

The following genes have a possible association with obesity.

5.3.1 The *HYKK* gene

The *HYKK* gene encodes for a protein that enables the activity of hydroxylysine kinase. This protein is predicted to be involved in the catabolic process of lysine and is expected to be located in the mitochondrial matrix [https://www.alliancegenome.org].

The SNP rs183603441 acts as a pretermination codon in this gene and potentially reduces the length of the protein by 61.2 % from 374 to 145 amino acids.

While the PTC rs183603441 in the *HYKK* gene passed the nominal p-value threshold for a dominant model and showed a very small effect size (Cohen's d = 0.155), the difference in mean BMI between the major allele homozygote group and the heterozygote + minor allele group was also very small. Additionally, there is currently no known association between mutations in the *HYKK* gene and any specific disease or disorder. Moreover, qualitative analysis did not confirm the result. Therefore, it is unlikely that these results have any practical clinical implications at this time.

5.3.2 The *TRPM1* gene

The *TRPM1* gene is a member of the transient receptor potential (TRP) family of ion channels. It is expressed in the retina, where it plays a crucial role in the development and function of certain cells called ON bipolar cells [120]. These cells are involved in the first step of visual processing and are responsible for transmitting visual signals from the photoreceptor cells to the ganglion cells in the retina. *TRPM1* is thought to regulate the release of neurotransmitters from ON bipolar cells, which is essential for normal visual function [121]. Mutations in the *TRPM1* gene have been associated with a rare form of congenital stationary night blindness (CSNB), a condition that affects the ability to see in low light conditions [122].

The SNP rs3784589 acts as a pretermination codon in this gene and potentially reduces the length of the protein by 14.3 % from 1604 to 1375 amino acids.

In this study, the PTC (rs3784589) p.value passed the nominal p.value threshold for overdominant, log-additive and dominant models. However, the best-fitted model was the overdominant model. The mean value of BMI was slightly higher in the A/C (heterozygote) group than in the C/C-A/A group, with a very small effect size (Cohen’s d = 0.106). Despite these findings, the limitations of this study should be taken into account when interpreting the results. First, there is no connection between the functionality of the *TRPM1* gene and obesity. Second, the overdominant model is hard to interpret for a nonsense mutation as these mutations usually follow a recessive pattern. Therefore, the overdominant model may not be the best fit for this PTC, and the results should be interpreted with caution. Overall, the findings related to this gene are not likely to be clinically significant due to the low effect size and limitations mentioned above.

5.3.3 The *CLYBL* gene

The *CLYBL* gene codes for the enzyme cysteine lyase beta-lyase (CLYBL), which has been shown to enable (S)-citramalyl-CoA lyase activity, magnesium ion binding activity, and malate synthase activity. It has also been implicated in protein homotrimerization and the regulation of cobalamin metabolic processes. The protein is predicted to be located in the mitochondrion and may also be an integral component of the membrane [123]. The *CLYBL* gene is broadly expressed in various tissues, including the kidney (RPKM 2.9) and liver (RPKM 2.6). Recent studies have suggested that the CLYBL enzyme plays an important role in the metabolism of vitamin B12 and is involved in the conversion of vitamin B12 into a form that can be used by the body to produce energy. In fact, the human knockout gene for CLYBL has been shown to connect itaconate to vitamin B12 [124].

The SNP rs41281112 acts as a pretermination codon in this gene and potentially reduces the length of the expressed protein by 24 % from 341 to 259 amino acids.

Based on the genetic association analysis in the present thesis, rs41281112 passed the nominal p.value threshold for dominant and overdominant models. However, a better genetic model could not be determined with AIC value, and therefore the dominant model was used for analysis. This PTC is located in the *CLYBL* gene on and follows Hardy-Weinberg equilibrium. The mean value of BMI was slightly higher in the group with at least one minor allele, with a Cohen's d of 0.15, indicating a very small effect size. It is important to note that the genetic model for this RS is hard to interpret as it follows a dominant pattern, while nonsense mutations typically follow a recessive pattern. Furthermore, the effect size observed in the study is very small, indicating that the findings related to this gene may not have any clinical significance in regards to obesity.

5.3.4 The *ERV3-1* gene

The *ERV3-1* gene, also known as endogenous retrovirus group 3 member 1, codes for a protein with sequence derived from an endogenous retrovirus, which is why it is similar to multiple other loci in the human genome. The transcripts at this locus encode a conserved protein with a predicted signal peptide and similarity to the Env polyprotein, suggesting a possible role in viral entry and fusion. In addition to its role in regulating the immune response and the maternal-fetal interface, the ERV3-1 protein is also overexpressed in certain types of cancer, including colorectal cancer and other cancers [125,126]. It is expressed at high levels in the adrenal gland and fat tissue [127].

The SNP rs67047829 acts as a pretermination codon in this gene and potentially reduces the length of the expressed protein by 73.1 % from 605 to 223 amino acids.

rs67047829, on *ERV3-1,* showed a potential clinical significance in its association with BMI. The rs67047829 PTC passed the nominal p-value threshold for recessive and additive models, with the recessive model being the best fit for this dataset based on AIC. This PTC is located in the *ERV3-1* gene on chromosome 7. The minor allele homozygote group (A/A) had a significantly lower mean BMI compared to the major allele homozygote group (G/G-A/G) (Cohen's d = 0.48, small effect size). In the A/A group, 64.3% of the subjects were of normal weight, while in the G/G-A/G group, only 48.3% of the subjects were of normal weight. Furthermore, the post hoc analysis revealed a significant difference in the proportion of subjects with normal weight versus obesity between the two groups (OR= 0.183, 95% CI:[0.021, 0.713], p = 0.0059). These findings suggest that the minor allele homozygote of rs67047829 may be associated with a lower risk of obesity.

5.3.5 The *GPRC6A* gene

The *GPRC6A* gene encodes a member of the G protein-coupled receptor family, which is expressed in a variety of tissues, including bone, kidney, testis, and brain. The protein product of this gene is called the GPRC6A receptor. The GPRC6A receptor is activated by a variety of ligands, including amino acids, calcium ions, and osteocalcin, a hormone secreted by bone cells. Activation of the receptor has been shown to play a role in regulating insulin secretion, glucose homeostasis, and bone metabolism [128].

The SNP rs6907580 acts as a pretermination codon in this gene and potentially reduces the length of the protein by 93.9 % from 927 to 57 amino acids.

Rs6907580 passed the nominal p-value threshold. The quantitative analysis found no significant difference in BMI mean values between the G/G (major allele homozygote) and A/G-A/A (heterozygote, minor allele homozygote) groups. Moreover, the post-hoc analysis (normal weight vs obesity) indicated a small effect size (OR=0.71, 95% CI:[0.51, 0.96], p=0.024) between groups. Therefore, the likelihood of clinical significance of rs6907580 in relation to obesity risk is considered low or close to none (upper CI = 0.96).

5.3.6 The *LCN10* gene

The *LCN10* gene, also known as Lipocalin 10, is a member of the lipocalin protein family. Lipocalins are small extracellular proteins that are involved in a variety of biological processes, including transport of small molecules, regulation of inflammation, and modulation of cell signaling [129]. The exact function of the LCN10 protein is not yet fully understood.

The SNP rs9886752 acts as a pretermination codon in this gene and potentially reduces the length of the protein by 19.9 % from 201 to 161 amino acids.

Rs9886752 passed the nominal p-value threshold for qualitative analysis. The results showed that the minor allele of rs9886752 was associated with an increased risk of obesity. In the A/G-A/A group, 17.2% were obese compared to 13.8% in the G/G group, and post hoc analysis showed significant differences in proportions (OR=1.29, 95% CI:[1.06, 1.565], p=0.008) between the two groups. However, the quantitative analysis did not confirm this result. It is important to note that although there was a significant difference in proportions, the effect size was relatively small, and the clinical significance of this finding is not clear at this time. Further studies are needed to investigate the potential role of *LCN10* in the development of obesity.

5.4 Summary

The present study aimed to investigate the possible associations between pretermination codons (PTCs) and three different phenotypes: longevity, fertility, and obesity risk. The study analyzed 141 PTC single nucleotide polymorphisms (SNPs) (from 5095 patients) and found that 21 of them showed statistically significant associations with at least one of the three phenotypes. However, caution is advised when interpreting these results, as complex phenotypes such as longevity, fertility, and obesity risk require careful study design to identify true genetic associations.

One promising candidate for further analysis is the *ERV3-1* gene, as a SNP on this gene (rs67047829) showed potential clinical significance in its association with BMI. The minor allele homozygote group (A/A) had a significantly lower mean BMI compared to the major allele homozygote group (G/G-A/G), and post-hoc analysis revealed a significant difference in the proportion of subjects with normal weight versus obesity between the two groups. These findings suggest that the minor allele homozygote of rs67047829 may be associated with a lower risk of obesity.

Despite the interesting findings, the study has several limitations. Firstly, research on some of the PTC SNPs is still in its early stages, and the specific functions and roles of these SNPs in human biology are not yet well understood. Secondly, the sample size of the minor allele homozygote group for some of the SNPs in this study was very small, which may limit the generalizability of the results. Thirdly, the phenotypes measured in this study may not be directly related to the function of some of the PTC SNPs, and additional research is needed to establish clear connections between these genes and outcomes. Finally, although the p-values for the associations between some of the PTC SNPs and the three phenotypes passed a nominal threshold, they did not meet the Bonferroni correction threshold for multiple testing, indicating the possibility of false positive results. Therefore, further studies are needed to confirm these findings and establish the clinical significance of these genes in relation to longevity, fertility, and obesity risk.

5.4.1 Limitations

- Longevity: Some studies have suggested that an association analysis with longevity should be performed between two groups: supercentenarians and a healthy, younger control group. However, since the original purpose of database used in this study was an association analysis using the obesity phenotype, it wasn't possible to create a proper supercentenarian group. Thus, the study's findings regarding longevity should be interpreted with caution. Additionally, the present study analyzed single-gene associations with the complex phenotype of longevity, which is influenced by a combination of genetic, environmental, and lifestyle factors. This complexity presents a limitation in the study, as age analysis without proper adjustments can only unequivocally demonstrate large genotypic effects.
- A large amount of missing genotype data for some of the SNPs: The study found statistically significant PTCs for multiple genes, but a significant proportion of patients had missing genotype data for some of these PTCs. This limits the study's ability to draw meaningful conclusions about the association of these PTCs with longevity.
- Ambiguous protein function: The protein function of some of the genes under investigation is unclear. This adds to the uncertainty surrounding the clinical significance of these PTCs. Without a clear understanding of how these genes and their associated PTCs influence the analyzed phenotype, it is difficult to draw any conclusions about their potential clinical relevance.
- Sample size: the sample size for some groups (minor allele homozygotes) for some SNPs in this study was very small, which may limit the statistical power of the analysis and the generalizability of the results

- Low effect size: The study found statistically significant PTCs for multiple genes, but the effect size for some of these PTCs was very low. This suggests that these PTCs are unlikely to have a significant impact on human longevity. While statistically significant, the small effect size diminishes the clinical significance of these PTCs.
- Impact of missing genotype data on study accuracy and reliability: The high amount of missing genotype data for some of the PTCs in the study can significantly impact the accuracy and reliability of genetic association studies. Without complete and accurate genotype data, it is difficult to draw valid conclusions about the association of PTCs with phenotype.

- P.value: Although some associations between genetic variants and the phenotype passed a nominal threshold, they did not meet the Bonferroni correction threshold for multiple testing, indicating the possibility of false positive results

6. Abstract

Mutations play a crucial role in adaptation to new environmental conditions, including changes in protein functionality that are necessary for adaptation in biochemical pathways. However, mutations can also lead to diseases such as cystic fibrosis, Duchenne muscular dystrophy, β-thalassemia, and cancer. The occurrence of premature stop codons is a common cause of such diseases, and these can arise from germline or somatic DNA mutations, inaccurate pre-mRNA splicing, or lack of optimization of RNA editing. Natural selection can result in new alleles with high frequencies of derived alleles and varying levels of population diversity. In the present study, 141 SNPs which lead to premature stop codons were studied, most of which have high enough frequencies (>5%) to be regarded as being subject to near-neutral selection. These PTCs were selected from a study by Fujikura et al, and some of the genes harboring the mutation have been ontologically categorized as being involved in metabolism, drug metabolism, the immune system, zinc fingers, and keratin. Based on this ontology, the phenotypes analyzed in the present study were: obesity, overweight, fertility, and life expectancy.

The specific aims of the present study were the analysis of possible associations between pretermination codons and age (life span), analysis of possible associations between pretermination codons and the number of children (fertility) and analysis of possible associations between pretermination codons and body mass index (obesity and overweight).

Association analysis was performed on a database obtained through an agreement with the University of Lodz, which contained 5,600 samples from healthy people in Poland, with 500,000 SNPs from each subject, including 141 PTC SNPs. All statistical association analyses for this thesis were performed using the R statistical platform. The standardized effect sizes used in this thesis were: Cohen's d, Spearman's r and odds ratio. For association analysis with quantitative variables, linear and logistic regression models, implemented in the R SNPassoc package, were used. All statistical tests were two-tailed and two statistical significance p.value thresholds were set, the nominal p < 0.05 and after Bonferroni correction < (0.05/141 SNPs analyzed) = 3.55 x 10-4. For the qualitative variable of BMI categories versus SNP for dominant and recessive models, the Fisher exact test and Cochran Armitage trend test were used.

The study aimed to investigate possible associations between PTCs and three different phenotypes: longevity, fertility, and obesity risk. The study analyzed 141 PTC SNPs (from 5095 patients) and found that 21 of them showed statistically significant associations with one of the three phenotypes. One promising candidate for further analysis is the *ERV3-1* gene, as a SNP on this gene (rs67047829) showed a potential clinical significance in its association with BMI. Despite the interesting findings, the study has several limitations, and these results should be treated with caution.

6.1 Streszczenie

Mutacje odgrywają kluczową rolę w adaptacji do nowych warunków środowiskowych, w tym w zmianach funkcjonalności białek, które są niezbędne do adaptacji w szlakach biochemicznych. Jednak mutacje mogą także prowadzić do chorób, takich jak mukowiscydoza, dystrofia mięśniowa Duchenne'a, β-talasemia i nowotwory. Występowanie przedwczesnych kodonów stop jest powszechną przyczyną takich chorób, a mogą one wynikać z mutacji germlinej lub somatycznej DNA, niedokładnego splicingu pre-mRNA lub braku optymalizacji edycji RNA. Naturalna selekcja może prowadzić do nowych alleli z wysoką częstością pochodnych alleli i zróżnicowanym poziomem zróżnicowania populacji. W niniejszym badaniu przeanalizowano 141 SNPs prowadzących do przedwczesnych kodonów stop, z których większość ma wystarczająco wysokie występowanie (> 5%) by być uważane za podlegające niemal neutralnej selekcji. PTCs zostały wybrane z badania Fujikury i wsp., a niektóre z genów zawierających mutację zostały ontologicznie sklasyfikowane jako związane z metabolizmem, metabolizmem leków, układem odpornościowym, palcami cynkowymi i keratyną. Na podstawie tej ontologii w niniejszym badaniu analizowane były następujące fenotypy: otyłość, nadwaga, płodność i długość życia.

Konkretne cele niniejszego badania to analiza możliwych związków między kodonami przedterminacyjnymi a wiekiem (długość życia), analiza możliwych związków między kodonami przedterminacyjnymi a liczbą dzieci (płodność) oraz analiza możliwych związków między kodonami przedterminacyjnymi a wskaźnikiem masy ciała (otyłość i nadwaga).

Analiza skojarzeń została przeprowadzona na bazie danych uzyskanej na podstawie umowy z Uniwersytetem Łódzkim, która zawierała 5 600 próbek zdrowych ludzi w Polsce, z 500 000 SNP od każdego badanego, w tym 141 SNP PTC. Wszystkie analizy skojarzeń statystycznych w tej pracy doktorskiej zostały przeprowadzone za pomocą platformy statystycznej R. Standaryzowane wielkości efektów używane w tej pracy to: d Cohena, r Spearmana i współczynnik szans. Do analizy skojarzeń z zmiennymi ilościowymi wykorzystano modele regresji liniowej i logistycznej, zaimplementowane w pakiecie R SNPassoc.

Wszystkie testy statystyczne były dwustronne i ustalono dwa progi istotności statystycznej: nominalny p <0,05 oraz skorygowany Bonferronim p < (0,05 / 141 SNP analizowanych) = 3,55 x 10^{-4}. Dla zmiennej jakościowej kategorii BMI względem SNP dla modeli dominujących i recesywnych użyto testu exact dokładnego Fishera oraz testu trendy Cochran-Armitage'a.

Celem badania było zbadanie możliwych związków między PTC a trzema różnymi fenotypami: długowiecznością, płodnością i ryzykiem otyłości. W badaniu przeanalizowano 141 SNP PTC (z 5095

pacjentów) i stwierdzono, że 21 z nich wykazywało statystycznie istotne związki z jednym z trzech fenotypów. Jednym obiecującym kandydatem do dalszej analizy jest gen ERV3-1, ponieważ SNP na tym genie (rs67047829) wykazał potencjalne znaczenie kliniczne w związku z BMI. Mimo ciekawych wyników, badanie ma kilka ograniczeń, a te wyniki należy traktować z rozwagą.

CITATIONS

[1] Loewe L, Hill WG. The population genetics of mutations: good, bad and indifferent. Philos Trans R Soc Lond B Biol Sci 2010;365:1153–67. https://doi.org/10.1098/rstb.2009.0317.

[2] Keeling KM, Du M, Bedwell DM. Therapies of Nonsense-Associated Diseases. Landes Bioscience; 2013.

[3] Lachance J, Vernot B, Elbers CC, Ferwerda B, Froment A, Bodo J-M, et al. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. Cell 2012;150:457–69. https://doi.org/10.1016/j.cell.2012.07.009.

[4] Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, et al. Classic selective sweeps were rare in recent human evolution. Science 2011;331:920–4. https://doi.org/10.1126/science.1198878.

[5] Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, et al. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 2001;293:455–62. https://doi.org/10.1126/science.1061573.

[6] Hamblin MT, Thompson EE, Di Rienzo A. Complex signatures of natural selection at the Duffy blood group locus. Am J Hum Genet 2002;70:369–83. https://doi.org/10.1086/338628.

[7] Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, et al. Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. Am J Hum Genet 2004;74:1111–20.

[8] Campbell MC, Ranciaro A, Froment A, Hirbo J, Omar S, Bodo J-M, et al. Evolution of functionally diverse alleles associated with PTC bitter taste sensitivity in Africa. Mol Biol Evol 2012;29:1141–53. https://doi.org/10.1093/molbev/msr293.

[9] Menashe I, Man O, Lancet D, Gilad Y. Different noses for different people. Nat Genet 2003;34:143–4. https://doi.org/10.1038/ng1160.

[10] Scriver CR. The principles of human biochemical genetics. By Harry Harris. North-Holland, Amsterdam; American Elsevier, New York. 328 pp. 1970. vol. 5. 1972.

[11] Mühlemann O, Eberle AB, Stalder L, Zamudio Orozco R. Recognition and elimination of nonsense mRNA. Biochim Biophys Acta 2008;1779:538–49. https://doi.org/10.1016/j.bbagrm.2008.06.012.

[12] Morais P, Adachi H, Yu Y-T. Suppression of Nonsense Mutations by New Emerging Technologies. International Journal of Molecular Sciences 2020;21:4394. https://doi.org/10.3390/ijms21124394.

[13] Jennings MT, Riekert KA, Boyle MP. Update on key emerging challenges in cystic fibrosis. Med Princ Pract 2014;23:393–402. https://doi.org/10.1159/000357646.

[14] Chowdhury HM, Siddiqui MA, Kanneganti S, Sharmin N, Chowdhury MW, Nasim MT. Aminoglycoside-mediated promotion of translation readthrough occurs through a non-stochastic mechanism that competes with translation termination. Hum Mol Genet 2018;27:373–84. https://doi.org/10.1093/hmg/ddx409.

[15] Howard M, Frizzell RA, Bedwell DM. Aminoglycoside antibiotics restore CFTR function by overcoming premature stop mutations. Nat Med 1996;2:467–9. https://doi.org/10.1038/nm0496-467.

[16] Nussbaum RL. Thompson & Thompson Genetics in Medicine. 8th ed. London: Elsevier Health Sciences; 2015.

[17] Acuna-Hidalgo R, Veltman JA, Hoischen A. New insights into the generation and role of de novo mutations in health and disease. Genome Biology 2016;17:241. https://doi.org/10.1186/s13059-016-1110-1.

[18] Erickson RP. Somatic gene mutation and human disease other than cancer: An update. Mutation Research/Reviews in Mutation Research 2010;705:96–106. https://doi.org/10.1016/j.mrrev.2010.04.002.

[19] Antonarakis SE, Cooper DN. Human gene mutation in inherited disease: Molecular mechanisms and clinical consequences. In: Rimoin DL, Pyeritz RE, Korf B, editors., Elsevier; 2013, p. 1–48. https://doi.org/10.1016/B978-0-12-383834-6.00007-0.

[20] Tardif S, Cormier N. Role of zonadhesin during sperm-egg interaction: a species-specific acrosomal molecule with multiple functions. Mol Hum Reprod 2011;17:661–8. https://doi.org/10.1093/molehr/gar039.

[21] Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Mol Cell Proteomics 2014;13:397–406. https://doi.org/10.1074/mcp.M113.035600.

[22] Fujikura K. Premature termination codons in modern human genomes. Sci Rep 2016;6:22468. https://doi.org/10.1038/srep22468.

[23] David P. Clark;, Nanette J. Pazdernik;, Michelle R. McGehee. Mutations and Repair. Molecular Biology. 3rd ed., 2019, p. 832–79.

[24] Chu D, Wei L. Nonsynonymous, synonymous and nonsense mutations in human cancer-related genes undergo stronger purifying selections than expectation. BMC Cancer 2019;19:359. https://doi.org/10.1186/s12885-019-5572-x.

[25] Camps M, Herman A, Loh E, Loeb LA. Genetic Constraints on Protein Evolution. Crit Rev Biochem Mol Biol 2007;42:10.1080/10409230701597642. https://doi.org/10.1080/10409230701597642.

[26] Gregory TR. Understanding Natural Selection: Essential Concepts and Common Misconceptions. Evo Edu Outreach 2009;2:156–75. https://doi.org/10.1007/s12052-009-0128-1.

[27] Kim VN, Kataoka N, Dreyfuss G. Role of the nonsense-mediated decay factor hUpf3 in the splicing-dependent exon-exon junction complex. Science 2001;293:1832–6. https://doi.org/10.1126/science.1062829.

[28] Brogna S, Wen J. Nonsense-mediated mRNA decay (NMD) mechanisms. Nat Struct Mol Biol 2009;16:107–13. https://doi.org/10.1038/nsmb.1550.

[29] Hug N, Longman D, Cáceres JF. Mechanism and regulation of the nonsense-mediated decay pathway. Nucleic Acids Res 2016;44:1483–95. https://doi.org/10.1093/nar/gkw010.

[30] Okada-Katsuhata Y, Yamashita A, Kutsuzawa K, Izumi N, Hirahara F, Ohno S. N- and C-terminal Upf1 phosphorylations create binding platforms for SMG-6 and SMG-5:SMG-7 during NMD. Nucleic Acids Res 2012;40:1251–66. https://doi.org/10.1093/nar/gkr791.

[31] Sulkowska A, Wawer I. Sens nonsensu czyli na straży jakości mRNA. Postępy Biochemii 2017;63:185–9.

[32] Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. Nat Rev Genet 2009;10:681–90. https://doi.org/10.1038/nrg2615.

[33] Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nat Rev Methods Primers 2021;1:1–21. https://doi.org/10.1038/s43586-021-00056-9.

[34] Palsson R, Indridason OS, Edvardsson VO, Oddsson A. Genetics of common complex kidney stone disease: insights from genome-wide association studies. Urolithiasis 2019;47:11–21. https://doi.org/10.1007/s00240-018-1094-2.

[35] Zeng P, Zhao Y, Qian C, Zhang L, Zhang R, Gou J, et al. Statistical analysis for genome-wide association study. J Biomed Res 2015;29:285–97. https://doi.org/10.7555/JBR.29.20140007.

[36] Sobalska-Kwapis M, Suchanecka A, Słomka M, Siewierska-Górska A, Kępka E, Strapagiel D. Genetic association of FTO/IRX region with obesity and overweight in the Polish population. PLOS ONE 2017;12:e0180295. https://doi.org/10.1371/journal.pone.0180295.

[37] Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 2019;47:D1005–12. https://doi.org/10.1093/nar/gky1120.

[38] Thierry van de Wetering. Association studies between selected chromosomal regions 1q21.3, 5q21.3, 14q21.2 and 17q21.31 with the number of offspring in Poles - analysis of data from the POPULOUS database. 2019.

[39] Barban N, Jansen R, de Vlaming R, Vaez A, Mandemakers JJ, Tropf FC, et al. Genome-wide analysis identifies 12 loci influencing human reproductive behavior. Nat Genet 2016;48:1462–72. https://doi.org/10.1038/ng.3698.

[40] Sullivan GM, Feinn R. Using Effect Size—or Why the P Value Is Not Enough. Journal of Graduate Medical Education 2012;4:279–82. https://doi.org/10.4300/JGME-D-12-00156.1.

[41] Sawilowsky S. New Effect Size Rules of Thumb. Journal of Modern Applied Statistical Methods 2009;8:26. https://doi.org/10.22237/jmasm/1257035100.

[42] Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale, N.J.: L. Erlbaum Associates; 1988.

[43] Valentine JC, Pigott TD, Rothstein HR. How many studies do you need? A primer on statistical power for meta-analysis. Journal of Educational and Behavioral Statistics 2010;35:215–47. https://doi.org/10.3102/1076998609346961.

[44] Bland JM, Altman DG. The odds ratio. BMJ 2000;320:1468.

[45] Min SH, Zhou J. smplot: An R Package for Easy and Elegant Data Visualization. Frontiers in Genetics 2021;12.

[46] R-Core-Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2017.

[47] González JR, Armengol L, Solé X, Guinó E, Mercader JM, Estivill X, et al. SNPassoc: an R package to perform whole genome association studies. Bioinformatics 2007;23:654–5. https://doi.org/10.1093/bioinformatics/btm025.

[48] Ehret GB, Ferreira T, Chasman DI, Jackson AU, Schmidt EM, Johnson T, et al. The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals. Nat Genet 2016;48:1171–84. https://doi.org/10.1038/ng.3667.

[49] Hoffmann TJ, Ehret GB, Nandakumar P, Ranatunga D, Schaefer C, Kwok P-Y, et al. Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. Nat Genet 2017;49:54–64. https://doi.org/10.1038/ng.3715.

[50] Liang J, Le TH, Edwards DRV, Tayo BO, Gaulton KJ, Smith JA, et al. Single-trait and multi-trait genome-wide association analyses identify novel loci for blood pressure in African-ancestry populations. PLoS Genet 2017;13:e1006728. https://doi.org/10.1371/journal.pgen.1006728.

[51] Lane JM, Liang J, Vlasac I, Anderson SG, Bechtold DA, Bowden J, et al. Genome-wide association analyses of sleep disturbance traits identify new loci and highlight shared genetics with neuropsychiatric and metabolic traits. Nat Genet 2017;49:274–81. https://doi.org/10.1038/ng.3749.

[52] Burnham KP, Anderson DR, editors. Information and Likelihood Theory: A Basis for Model Selection and Inference. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, New York, NY: Springer; 2002, p. 49–97. https://doi.org/10.1007/978-0-387-22456-5_2.

[53] Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control 1974;19:716–23. https://doi.org/10.1109/TAC.1974.1100705.

[54] Willing MC, Deschenes SP, Slayton RL, Roberts EJ. Premature chain termination is a unifying mechanism for COL1A1 null alleles in osteogenesis imperfecta type I cell strains. Am J Hum Genet 1996;59:799–809.

[55] Thein SL. Molecular basis of β thalassemia and potential therapeutic targets. Blood Cells Mol Dis 2018;70:54–65. https://doi.org/10.1016/j.bcmd.2017.06.001.

[56] Duan D, Goemans N, Takeda S, Mercuri E, Aartsma-Rus A. Duchenne muscular dystrophy. Nat Rev Dis Primers 2021;7:1–19. https://doi.org/10.1038/s41572-021-00248-3.

[57] Naehrig* S, Chao* C-M, Naehrlich L. Cystic Fibrosis. Dtsch Arztebl Int 2017;114:564–74. https://doi.org/10.3238/arztebl.2017.0564.

[58] Deelen J, Evans DS, Arking DE, Tesi N, Nygaard M, Liu X, et al. A meta-analysis of genome-wide association studies identifies multiple longevity genes. Nat Commun 2019;10:3669. https://doi.org/10.1038/s41467-019-11558-2.

[59] Wang H, Liu H, Guo F, Li J, Li P, Guan T, et al. Association Between Ambient Fine Particulate Matter and Physical Functioning in Middle-Aged and Older Chinese Adults: A Nationwide Longitudinal Study. J Gerontol A Biol Sci Med Sci 2022;77:986–93. https://doi.org/10.1093/gerona/glab370.

[60] Li Y, Schoufour J, Wang DD, Dhana K, Pan A, Liu X, et al. Healthy lifestyle and life expectancy free of cancer, cardiovascular disease, and type 2 diabetes: prospective cohort study. BMJ 2020;368:l6669. https://doi.org/10.1136/bmj.l6669.

[61] Kojima T, Kamei H, Aizu T, Arai Y, Takayama M, Nakazawa S, et al. Association analysis between longevity in the Japanese population and polymorphic variants of genes involved in insulin and insulin-like growth factor 1 signaling pathways. Experimental Gerontology 2004;39:1595–8. https://doi.org/10.1016/j.exger.2004.05.007.

[62] Garagnani P, Marquis J, Delledonne M, Pirazzini C, Marasco E, Kwiatkowska KM, et al. Whole-genome sequencing analysis of semi-supercentenarians. ELife 2021;10:e57849. https://doi.org/10.7554/eLife.57849.

[63] Klaassen CD, Boles JW. The importance of 3′-phosphoadenosine 5′-phosphosulfate (PAPS) in the regulation of sulfation. The FASEB Journal 1997;11:404–18. https://doi.org/10.1096/fasebj.11.6.9194521.

[64] Meinl W, Donath C, Schneider H, Sommer Y, Glatt H. SULT1C3, an orphan sequence of the human genome, encodes an enzyme activating various promutagens. Food Chem Toxicol 2008;46:1249–56. https://doi.org/10.1016/j.fct.2007.08.040.

[65] Kurogi K, Rasool MI, Alherz FA, El Daibani AA, Bairam AF, Abunnaja M, et al. SULT genetic polymorphisms: physiological, pharmacological and clinical implications. Expert Opin Drug Metab Toxicol 2021;17:767–84. https://doi.org/10.1080/17425255.2021.1940952.

[66] Guidry AL, Tibbs ZE, Runge-Morris M, Falany CN. EXPRESSION, PURIFICATION, AND CHARACTERIZATION OF HUMAN CYTOSOLIC SULFOTRANSFERASE (SULT) 1C4. Horm Mol Biol Clin Investig 2017;29:27–36. https://doi.org/10.1515/hmbci-2016-0053.

[67] Her C, Kaur GP, Athwal RS, Weinshilboum RM. Human sulfotransferase SULT1C1: cDNA cloning, tissue-specific expression, and chromosomal localization. Genomics 1997;41:467–70. https://doi.org/10.1006/geno.1997.4683.

[68] Sakakibara Y, Yanagisawa K, Katafuchi J, Ringer DP, Takami Y, Nakayama T, et al. Molecular cloning, expression, and characterization of novel human SULT1C sulfotransferases that catalyze the sulfonation of N-hydroxy-2-acetylaminofluorene. J Biol Chem 1998;273:33929–35. https://doi.org/10.1074/jbc.273.51.33929.

[69] Runge-Morris M, Kocarek TA. Expression of the sulfotransferase 1C family: implications for xenobiotic toxicity. Drug Metab Rev 2013;45:450–9. https://doi.org/10.3109/03602532.2013.835634.

[70] Freimuth RR, Wiepert M, Chute CG, Wieben ED, Weinshilboum RM. Human cytosolic sulfotransferase database mining: identification of seven novel genes and pseudogenes. Pharmacogenomics J 2004;4:54–65. https://doi.org/10.1038/sj.tpj.6500223.

[71] Duniec-Dmuchowski Z, Rondini EA, Tibbs ZE, Falany CN, Runge-Morris M, Kocarek TA. Expression of the Orphan Cytosolic Sulfotransferase SULT1C3 in Human Intestine: Characterization of the Transcript Variant and Implications for Function. Drug Metab Dispos 2014;42:352–60. https://doi.org/10.1124/dmd.113.055665.

[72] Kurogi K, Shimohira T, Kouriki-Nagatomo H, Zhang G, Miller ER, Sakakibara Y, et al. Human Cytosolic Sulphotransferase SULT1C3: genomic analysis and functional characterization of splice variant SULT1C3a and SULT1C3d. J Biochem 2017;162:403–14. https://doi.org/10.1093/jb/mvx044.

[73] Suyama M, Nagase T, Ohara O. HUGE: a database for human large proteins identified by Kazusa cDNA sequencing project. Nucleic Acids Res 1999;27:338–9. https://doi.org/10.1093/nar/27.1.338.

[74] Kikuno R, Nagase T, Nakayama M, Koga H, Okazaki N, Nakajima D, et al. HUGE: a database for human KIAA proteins, a 2004 update integrating HUGEppi and ROUGE. Nucleic Acids Res 2004;32:D502-504. https://doi.org/10.1093/nar/gkh035.

[75] Zhou J, Chen C, Liu S, Zhou W, Du J, Jiang Y, et al. Potential functional variants of KIAA genes are associated with breast cancer risk in a case control study. Ann Transl Med 2021;9:549. https://doi.org/10.21037/atm-20-6108.

[76] Rebouillat D, Hovnanian A, Marié I, Hovanessian AG. The 100-kDa 2′,5′-oligoadenylate synthetase catalyzing preferentially the synthesis of dimeric pppA2′p5′A molecules is composed of three homologous domains. J Biol Chem 1999;274:1557–65. https://doi.org/10.1074/jbc.274.3.1557.

[77] Lin R-J, Yu H-P, Chang B-L, Tang W-C, Liao C-L, Lin Y-L. Distinct antiviral roles for human 2′,5′-oligoadenylate synthetase family members against dengue virus infection. J Immunol 2009;183:8035–43. https://doi.org/10.4049/jimmunol.0902728.

[78] Bréhin A-C, Casadémont I, Frenkiel M-P, Julier C, Sakuntabhai A, Desprès P. The large form of human 2′,5′-Oligoadenylate Synthetase (OAS3) exerts antiviral effect against Chikungunya virus. Virology 2009;384:216–22. https://doi.org/10.1016/j.virol.2008.10.021.

[79] Lee DK, Lynch KR, Nguyen T, Im D-S, Cheng R, Saldivia VR, et al. Cloning and characterization of additional members of the G protein-coupled receptor family. Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression 2000;1490:311–23. https://doi.org/10.1016/S0167-4781(99)00241-9.

[80] Lindemann L, Ebeling M, Kratochwil NA, Bunzow JR, Grandy DK, Hoener MC. Trace amine-associated receptors form structurally and functionally distinct subfamilies of novel G protein-coupled receptors. Genomics 2005;85:372–85. https://doi.org/10.1016/j.ygeno.2004.11.010.

[81] Xu H, Storch T, Yu M, Elliott SP, Haslam DB. Characterization of the human Forssman synthetase gene. An evolving association between glycolipid synthesis and host-microbial interactions. J Biol Chem 1999;274:29390–8. https://doi.org/10.1074/jbc.274.41.29390.

[82] Dunson DB, Baird DD, Colombo B. Increased infertility with age in men and women. Obstet Gynecol 2004;103:51–6. https://doi.org/10.1097/01.AOG.0000100153.24061.45.

[83] Cooney LG, Dokras A. Beyond fertility: polycystic ovary syndrome and long-term health. Fertility and Sterility 2018;110:794–809. https://doi.org/10.1016/j.fertnstert.2018.08.021.

[84] Li Y, Lin H, Li Y, Cao J. Association between socio-psycho-behavioral factors and male semen quality: systematic review and meta-analyses. Fertility and Sterility 2011;95:116–23. https://doi.org/10.1016/j.fertnstert.2010.06.031.

[85] Du Plessis SS, Cabler S, McAlister DA, Sabanegh E, Agarwal A. The effect of obesity on sperm disorders and male infertility. Nat Rev Urol 2010;7:153–61. https://doi.org/10.1038/nrurol.2010.6.

[86] Homan GF, Davies M, Norman R. The impact of lifestyle factors on reproductive performance in the general population and those undergoing infertility treatment: a review. Human Reproduction Update 2007;13:209–23. https://doi.org/10.1093/humupd/dml056.

[87] Gaskins AJ, Chavarro JE. Diet and fertility: a review. Am J Obstet Gynecol 2018;218:379–89. https://doi.org/10.1016/j.ajog.2017.08.010.

[88] Hajela S, Prasad S, Kumaran A, Kumar Y. Stress and infertility: a review. International Journal of Reproduction, Contraception, Obstetrics and Gynecology 2016;5:940–3. https://doi.org/10.18203/2320-1770.ijrcog20160846.

[89] Ehrlich S, Williams PL, Missmer SA, Flaws JA, Ye X, Calafat AM, et al. Urinary bisphenol A concentrations and early reproductive health outcomes among women undergoing IVF. Hum Reprod 2012;27:3583–92. https://doi.org/10.1093/humrep/des328.

[90] Giulioni C, Maurizi V, Castellani D, Scarcella S, Skrami E, Balercia G, et al. The environmental and occupational influence of pesticides on male fertility: A systematic review of human studies. Andrology 2022;10:1250–71. https://doi.org/10.1111/andr.13228.

[91] Angell RR, Aitken RJ, van Look PF, Lumsden MA, Templeton AA. Chromosome abnormalities in human embryos after in vitro fertilization. Nature 1983;303:336–8. https://doi.org/10.1038/303336a0.

[92] Winter H, Rogers MA, Gebhardt M, Wollina U, Boxall L, Chitayat D, et al. A new mutation in the type II hair cortex keratin hHb1 involved in the inherited hair disorder monilethrix. Hum Genet 1997;101:165–9. https://doi.org/10.1007/s004390050607.

[93] Wu J, Lin Y, Xu W, Li Z, Fan W. A mutation in the type II hair keratin KRT86 gene in a Han family with monilethrix. Journal of Biomedical Research 2011;25:49. https://doi.org/10.1016/S1674-8301(11)60006-7.

[94] Sennett R, Wang Z, Rezza A, Grisanti L, Roitershtein N, Sicchio C, et al. An Integrated Transcriptome Atlas of Embryonic Hair Follicle Progenitors, Their Niche, and the Developing Skin. Dev Cell 2015;34:577–91. https://doi.org/10.1016/j.devcel.2015.06.023.

[95] Kurada BRVVSN, Li LC, Mulherkar N, Subramanian M, Prasad KV, Prabhakar BS. MADD, a Splice Variant of IG20, Is Indispensable for MAPK Activation and Protection against Apoptosis upon Tumor Necrosis Factor-α Treatment. J Biol Chem 2009;284:13533–41. https://doi.org/10.1074/jbc.M808554200.

[96] Schievella AR, Chen JH, Graham JR, Lin LL. MADD, a novel death domain protein that interacts with the type 1 tumor necrosis factor receptor and activates mitogen-activated protein kinase. J Biol Chem 1997;272:12069–75. https://doi.org/10.1074/jbc.272.18.12069.

[97] Bachmann-Gagescu R, Dempsey JC, Phelps IG, O'Roak BJ, Knutzen DM, Rue TC, et al. Joubert syndrome: a model for untangling recessive disorders with extreme genetic heterogeneity. Journal of Medical Genetics 2015;52:514–22. https://doi.org/10.1136/jmedgenet-2015-103087.

[98] Parisi MA. The molecular genetics of Joubert syndrome and related ciliopathies: The challenges of genetic and phenotypic heterogeneity. Transl Sci Rare Dis n.d.;4:25–49. https://doi.org/10.3233/TRD-190041.

[99] Wheway G, Parry DA, Johnson CA. The role of primary cilia in the development and disease of the retina. Organogenesis 2014;10:69–85. https://doi.org/10.4161/org.26710.

[100] Higginbotham HR, Gleeson JG. The centrosome in neuronal development. Trends Neurosci 2007;30:276–83. https://doi.org/10.1016/j.tins.2007.04.001.

[101] Ishimaru Y, Inada H, Kubota M, Zhuang H, Tominaga M, Matsunami H. Transient receptor potential family members PKD1L3 and PKD2L1 form a candidate sour taste receptor. Proc Natl Acad Sci U S A 2006;103:12569–74. https://doi.org/10.1073/pnas.0602702103.

[102] Li A, Tian X, Sung S-W, Somlo S. Identification of two novel polycystic kidney disease-1-like genes in human and mouse genomes. Genomics 2003;81:596–608. https://doi.org/10.1016/s0888-7543(03)00048-x.

[103] Yuasa T, Takakura A, Denker BM, Venugopal B, Zhou J. Polycystin-1L2 is a novel G-protein-binding protein. Genomics 2004;84:126–38. https://doi.org/10.1016/j.ygeno.2004.02.008.

[104] Brass AL, Dykxhoorn DM, Benita Y, Yan N, Engelman A, Xavier RJ, et al. Identification of host proteins required for HIV infection through a functional genomic screen. Science 2008;319:921–6. https://doi.org/10.1126/science.1152725.

[105] Park C, Kim J-I, Hong SN, Jung HM, Kim TJ, Lee S, et al. A copy number variation in PKD1L2 is associated with colorectal cancer predisposition in korean population. Int J Cancer 2017;140:86–94. https://doi.org/10.1002/ijc.30421.

[106] Brown GD, Taylor PR, Reid DM, Willment JA, Williams DL, Martinez-Pomares L, et al. Dectin-1 Is A Major β-Glucan Receptor On Macrophages. J Exp Med 2002;196:407–12. https://doi.org/10.1084/jem.20020470.

[107] Goodridge HS, Reyes CN, Becker CA, Katsumoto TR, Ma J, Wolf AJ, et al. Activation of the innate immune receptor Dectin-1 upon formation of a 'phagocytic synapse.' Nature 2011;472:471–5. https://doi.org/10.1038/nature10071.

[108] Höglund PJ, Adzic D, Scicluna SJ, Lindblom J, Fredriksson R. The repertoire of solute carriers of family 6: identification of new human and rodent genes. Biochem Biophys Res Commun 2005;336:175–89. https://doi.org/10.1016/j.bbrc.2005.08.048.

[109] Bröer S, Bailey CG, Kowalczuk S, Ng C, Vanslambrouck JM, Rodgers H, et al. Iminoglycinuria and hyperglycinuria are discrete human phenotypes resulting from complex mutations in proline and glycine transporters. J Clin Invest 2008;118:3881–92. https://doi.org/10.1172/JCI36625.

[110] Maes HH, Neale MC, Eaves LJ. Genetic and environmental factors in relative body weight and human adiposity. Behav Genet 1997;27:325–51. https://doi.org/10.1023/a:1025635913927.

[111] Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature 2015;518:197–206. https://doi.org/10.1038/nature14177.

[112] Farooqi S, O'Rahilly S. Genetics of obesity in humans. Endocr Rev 2006;27:710–8. https://doi.org/10.1210/er.2006-0040.

[113] Loos RJF, Yeo GSH. The bigger picture of FTO: the first GWAS-identified obesity gene. Nat Rev Endocrinol 2014;10:51–61. https://doi.org/10.1038/nrendo.2013.227.

[114] Yeo GSH. Genetics of obesity: can an old dog teach us new tricks? Diabetologia 2017;60:778–83. https://doi.org/10.1007/s00125-016-4187-x.

[115] Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet 2010;42:937–48. https://doi.org/10.1038/ng.686.

[116] Malik VS, Hu FB. The role of sugar-sweetened beverages in the global epidemics of obesity and chronic diseases. Nat Rev Endocrinol 2022;18:205–18. https://doi.org/10.1038/s41574-021-00627-6.

[117] Ekelund U, Ward HA, Norat T, Luan J, May AM, Weiderpass E, et al. Physical activity and all-cause mortality across levels of overall and abdominal adiposity in European men and women: the European Prospective Investigation into Cancer and Nutrition Study (EPIC). Am J Clin Nutr 2015;101:613–21. https://doi.org/10.3945/ajcn.114.100065.

[118] Rosenbaum M, Hirsch J, Murphy E, Leibel RL. Effects of changes in body weight on carbohydrate metabolism, catecholamine excretion, and thyroid function. Am J Clin Nutr 2000;71:1421–32. https://doi.org/10.1093/ajcn/71.6.1421.

[119] Henderson DC, Vincenzi B, Andrea NV, Ulloa M, Copeland PM. Pathophysiological mechanisms of increased cardiometabolic risk in people with schizophrenia and other severe mental illnesses. The Lancet Psychiatry 2015;2:452–64. https://doi.org/10.1016/S2215-0366(15)00115-7.

[120] Morgans CW, Zhang J, Jeffrey BG, Nelson SM, Burke NS, Duvoisin RM, et al. TRPM1 is required for the depolarizing light response in retinal ON-bipolar cells. Proc Natl Acad Sci U S A 2009;106:19174–8. https://doi.org/10.1073/pnas.0908711106.

[121] Kozuka T, Chaya T, Tamalu F, Shimada M, Fujimaki-Aoba K, Kuwahara R, et al. The TRPM1 Channel Is Required for Development of the Rod ON Bipolar Cell-AII Amacrine Cell Pathway in the Retinal Circuit. J Neurosci 2017;37:9889–900. https://doi.org/10.1523/JNEUROSCI.0824-17.2017.

[122] van Genderen MM, Bijveld MMC, Claassen YB, Florijn RJ, Pearring JN, Meire FM, et al. Mutations in TRPM1 Are a Common Cause of Complete Congenital Stationary Night Blindness. Am J Hum Genet 2009;85:730–6. https://doi.org/10.1016/j.ajhg.2009.10.012.

[123] Strittmatter L, Li Y, Nakatsuka NJ, Calvo SE, Grabarek Z, Mootha VK. CLYBL is a polymorphic human enzyme with malate synthase and β-methylmalate synthase activity. Hum Mol Genet 2014;23:2313–23. https://doi.org/10.1093/hmg/ddt624.

[124] Shen H, Campanello GC, Flicker D, Grabarek Z, Hu J, Luo C, et al. The Human Knockout Gene CLYBL Connects Itaconate to Vitamin B12. Cell 2017;171:771-782.e11. https://doi.org/10.1016/j.cell.2017.09.051.

[125] Lee S-H, Kang Y-J, Jo J-O, Ock MS, Baek K-W, Eo J, et al. Elevation of human ERV3-1 env protein expression in colorectal cancer. J Clin Pathol 2014;67:840–4. https://doi.org/10.1136/jclinpath-2013-202089.

[126] Nakagawa S, Kawashima M, Miyatake Y, Kudo K, Kotaki R, Ando K, et al. Expression of ERV3-1 in leukocytes of acute myelogenous leukemia patients. Gene 2021;773:145363. https://doi.org/10.1016/j.gene.2020.145363.

[127] Kang Y-J, Jo J-O, Ock MS, Chang H-K, Baek K-W, Lee J-R, et al. Human ERV3-1 env protein expression in various human tissues and tumours. J Clin Pathol 2014;67:86–90. https://doi.org/10.1136/jclinpath-2013-201841.

[128] Pi M, Quarles LD. Multiligand specificity and wide tissue expression of GPRC6A reveals new endocrine networks. Endocrinology 2012;153:2062–9. https://doi.org/10.1210/en.2011-2117.

[129] Campanacci V, Nurizzo D, Spinelli S, Valencia C, Tegoni M, Cambillau C. The crystal structure of the Escherichia coli lipocalin Blc suggests a possible role in phospholipid binding. FEBS Lett 2004;562:183–8. https://doi.org/10.1016/S0014-5793(04)00199-1.