ESCUELA TÉCNICA
SUPERIOR INGENIEROS
INDUSTRIALES VALENCIA

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

MASTER'S THESIS IN BIOMEDICAL ENGINEERING

EXPLORING THE BRAME GENOMICS THROUGH VARIANTS-BASED DATA ANALYSIS

AUTOR: WERONIKA BRYJAK

SUPERVISOR: OSCAR PASTOR LÓPEZ

Academic year: 2019-20

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

The future is already here — it’s just not very evenly distributed.

William Gibson

ABSTRACT

The dynamic development of genomics technologies such as Next-Generation Sequencing (NGS) in recent years has contributed to the collection of enormous amounts of data related to the human genome. It provides a more complete picture of the genetic architecture and etiology of diseases enabling greater treatment perspective but also introduces new challenges in terms of efficient management of genomic data.

The massive and constantly expanding amount of information is ready to be used, but only part of it is sufficiently relevant to clinical practice application. It creates the need for a quality management system that allows data to be extracted from the "Big Data" to "Smart Data" perspective.

On this basis, the aim of the master's thesis is to identify genomic variants related to Autism Spectrum Disorder (ASD), Attention Deficit Hyperactivity Disorder (ADHD) and Schizophrenia which are the major psychiatric disorders affecting people.

The study was performed using the SILE method which provides appropriate tools for high quality data management.

As a result, of all the initial data, 7.33% of variants identified for ASD and 1.16% for Schizophrenia are relevant. Additionally, 0.44% of ASD variants, 3.03% of ADHD and 1.16% of Schizophrenia are considered promising.

The development and improvement of the management system for the rapid processing of raw genomic data into high quality information for clinical use gives strong perspectives of its application in precision medicine. Improving the process of selecting clinically relevant data and identifying variants associated with mental illness could increase the possibilities of today's healthcare industry by enabling personalized therapy for people suffering disease.

Keywords: variations, SILE, genomics, Attention Deficit Hyperactivity Disorder, Autism Spectrum Disorder, Schizophrenia, precision medicine

RESUMEN

El desarrollo dinámico de tecnologías genómicas, como Next-Generation Sequencing (NGS) en los últimos años, ha contribuido a la recopilación de enormes cantidades de datos relacionados con el genoma humano. Esto proporciona un visión más completa de la arquitectura genética y la etiología de las enfermedades, lo que ofrece una mayor perspectiva de tratamiento, pero además introduce nuevos desafíos en la gestión eficiente de datos genómicos.

La cantidad masiva y en constante de información está disponible para ser explotada, pero sólo una parte de ella es lo suficientemente relevante para la aplicación en la práctica clínica. Esto genera la necesidad de un sistema de gestión de calidad que permita extraer datos desde la perspectiva del "BIG DATA" y "SMART DATA".

A partir de esta base, el objetivo de la tesis de maestría es identificar las variantes genómicas relacionadas con el Trastorno del Espectro Autista (TEA), el Trastorno de Déficit de Atención e Hiperactividad (TDAH) y la Esquizofrenia, que son los principales trastornos psiquiátricos que afectan a las personas.

El estudio se realizó utilizando el método SILE, que proporciona herramientas apropiadas para la gestión de datos de alta calidad.

Como resultado, de todos los datos iniciales, el 7,33% de las variantes identificadas para el TEA y el 1,16% para la Esquizofrenia son relevantes. Además, el 0,44% de las variantes del TEA, el 3,03% del TDAH y el 1,16% de la Esquizofrenia se consideran prometedores.

El desarrollo y la mejora del sistema de gestión para el procesamiento rápido de los datos genómicos primarios en información de alta calidad para uso clínico ofrece sólidas perspectivas de su aplicación en la medicina de precisión. La optimización del proceso de selección de datos clínicamente pertinentes y la identificación de variantes asociadas a las enfermedades mentales podría aumentar las posibilidades de la industria de la salud actual al permitir una terapia personalizada para las personas que sufren enfermedades.

Palabras clave: variaciones, SILE, genómica, Trastorno de Déficit de Atención e Hiperactividad, Trastorno del Espectro de Autismo, Esquizofrenia, medicina de precisión

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ACMG	American College of Medical Genetics and Genomics
ADHD	Attention Deficit/Hyperactivity Disorder
AMA	American Medical Association
APA	American Psychiatric Association
API	Application Programming Interface
ASD	American Psychiatric Association
CLIA	Clinical Laboratory Improvement Amendments
CM	Conceptual Schema
CSHG	Conceptual Schema of the Human Genome
DNA	Deoxyribonucleic Acid
DQ	Data Quality
DSM	Diagnostic Statical Manual of Mental Disorders
ENCODE	Encyclopedia of DNA Elements
GLP	Good laboratory practice
HG	Human Genome
HGVS	Human Genome Variation Society
ICD	International Statistical Classification of Diseases and Related Health Problems
ISO	International Organization for Standardization
NAR	Nucleic Acids Research
NIH	National Institutes of Health
NSGC	National Society of Genetic Counselors
OMIM	Online Mendelian Inheritance in Man
SILE	Search Identification Load Exploitation

CHAPTER 1. INTRODUCTION

1.1 Motivation

In recent years, science has taken a huge step towards understanding the magnificent creation of nature, the human genome. This has created a completely new approach to understanding diseases, paving the way for medical breakthroughs.

The major goal of deciphering human code and creating an ever wider picture of genetic architecture is to understand the link between genomic variation, phenotype and disease.

Modern sequencing technologies, such as Next generation sequencing (NGS), have combined many new genes with rare diseases, and have generated a number of variants that cannot yet be interpreted. [1]

A huge amount of genomic data is generated every day and is stored in a wide range of heterogeneous databases. These databases cover various parts of human biology, from genetic sequencing to pharmacotherapy and differ in the structure, query languages, data models and quality of the information stored. This inconsistent mosaic of repositories makes accessing and aggregating relevant data very difficult and generates the need for adequate tools to efficiently manage information. [2]

Numerous research institutes, investment centers, as well as bioinformatics companies focus on developing a framework for genomic data management with the aim of transforming big amount of raw data into useful knowledge and obtaining valuable information for practical use in medicine. One of the institutions exploring the field is The Research Center on Software Production Methods (PROS) at the Universitat Politècnica de València. [3] PROS develops bioinformatic solutions applied to Human Genomics to maintain complex biological mechanisms and its relation to specific diseases. One of the institute's projects is the SILE methodology, a systematic approach for efficient management of genomic data based on conceptual modeling techniques and the data quality principles. The first two steps of the SILE methodology: Search and Identification are applied in the master’s thesis.

As a target of the project, three common mental disorders are taken into consideration: Autism Spectrum Disorder, Attention Deficit Hyperactivity Disorder and Schizophrenia. These disorders are characterized by extremely high morbidity, mortality and personal/social costs. Each of them is known by high heritability and its pathogenesis is the result of interactions between genetic and environmental factors. [4] However, the association between psychiatric disorders and genetics is very complex and difficult to estimate.

Improving the process of selecting clinically relevant data and identifying variants associated with mental illness will increase the possibilities of today's healthcare industry by enabling personalized therapy for people suffering disease.

This master’s thesis refers to the genomic data analysis branch of bioinformatics targeting precision medicine.

1.2 Aims and objectives

The general objective of the master's thesis is to identify a set of genomic variants associated with mental disorders: Attention Deficit Hyperactivity Disorder, Autism Spectrum Disorder, Schizophrenia using a method developed in the PROS Research Center of the Universidad Politècnica de Valencia.

The fundamental objectives of this thesis are the following:

1. To familiarize with the significance and challenges facing genomic data analysis in the mental health branch.
2. To select the genomic databases that correspond to the human genome and provide relevant information.
3. To use the SILE methodology to search and identify data that can later be loaded into the database.
4. To analyze the relationship between the selected variants of Attention Deficit Hyperactivity Disorder, Autism Spectrum Disorder and Schizophrenia, if any.

1.3 Structure of the document

Once the main objectives have been set, the structure of the master's thesis is arranged and consists of 6 chapters:

- CHAPTER 1. Presents the main aims and objectives of the work and the structure of the document.
- CHAPTER 2. Introduces the fundaments of molecular biology, genomic development and principal concepts associated with operating the genomic data.
- CHAPTER 3. Introduces phenotypes of Autism Spectrum Disorder, Attention Deficit Hyperactivity Disorder and Schizophrenia being a target of the research.
- CHAPTER 4. Describes the data management SILE method.
- CHAPTER 5. Evaluates the process of applying the SILE method to selected phenotypes.
- CHAPTER 6. Collects additional results and analysis based on the performance of the previous chapter including genetic relation of the results.
- CHAPTER 7. Presents final conclusions and future prospects associated with the study.

CHAPTER 2. STATE OF THE ART

2.1 Precision medicine

According to the Precision Medicine Initiative, precision medicine is "an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person." [5]

The risk of suffering a disease and the probability of response to medical treatments are conditioned by our genes and the environment we are exposed. The integration of genomic data and other omics sciences with the clinical data set of the patient, allows a clinical practice tailored to the individual characteristics of each patient.

The advancement in precision medicine is only possible through the prior development of the omics sciences and bioinformatics, as it is by expanding knowledge of the molecular and genetic bases of diseases and identifying a large number of biomarkers that allow the creation of more accurate and patient-specific protocols for diagnosis and treatment.

The following figure describes a simplified scheme of the genomic information flow that influences the development of precision medicine.

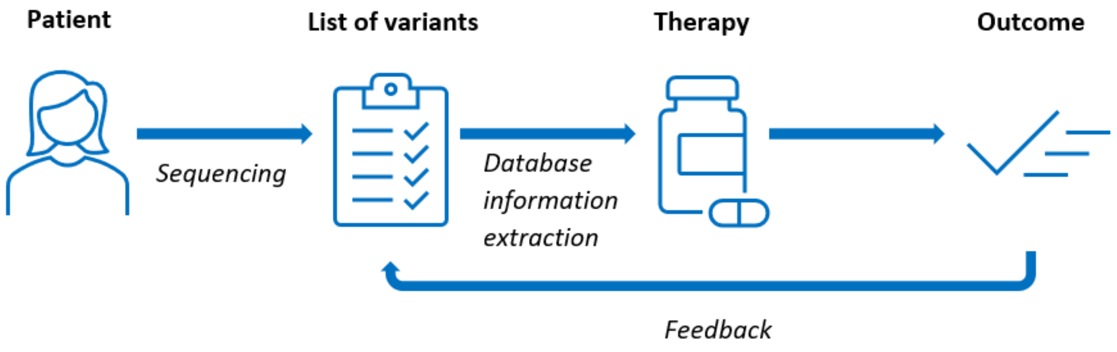


Figure 2.1 Simplified process diagram for precise genomic medicine

Precision medicine develops through three main segments:

- Therapeutics: cardiovascular, oncology, infectious diseases, central nervous system diseases and **psychiatric disorders**
- Technological: gene sequencing, pharmacogenomics, **big data analytics**, **bioinformatics**
- Ecosystem: diagnostic, pharmaceutical and biotechnological companies and clinical laboratories

2.2 Genomic initiatives

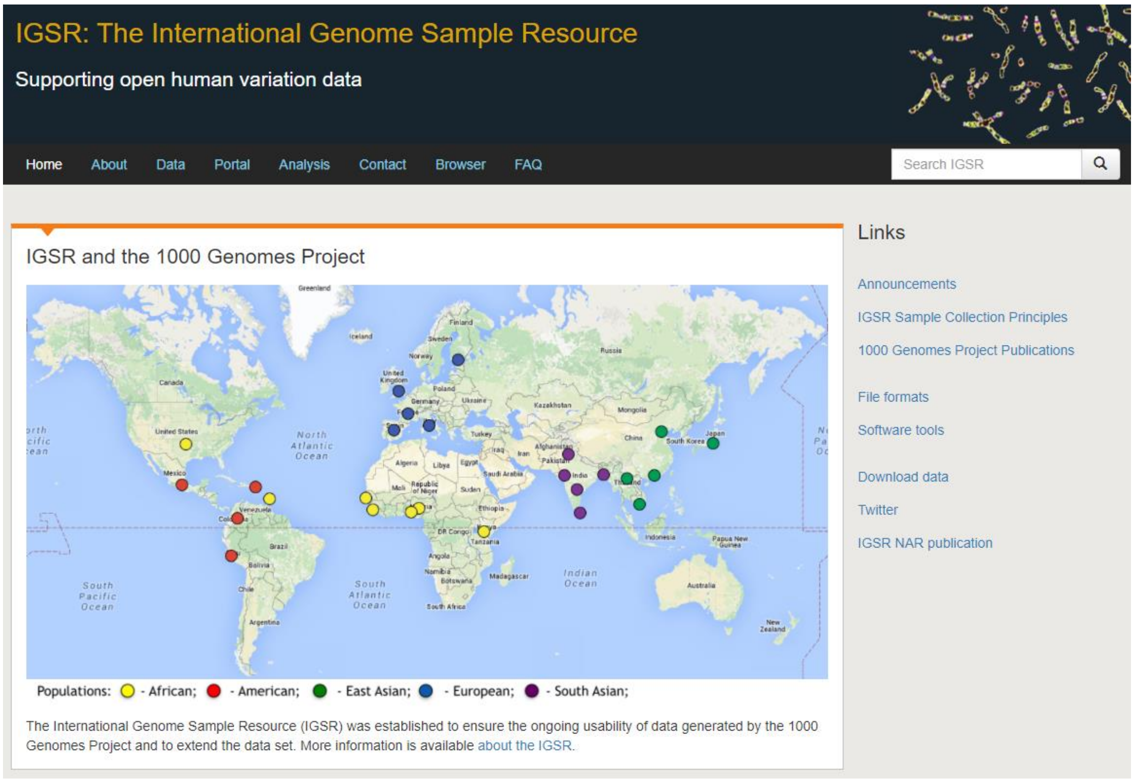
As mentioned earlier, the dynamic development of genomics and bioinformatics has been influenced by many research projects which now serve as a treasure trove of information aiding the development of gene-specific therapies.

Without the essential base of knowledge, they provided, it would not have been possible to perform research on the subject of this thesis.

The most significant of these are described in this section.

2.2.1 The 1000 Genomes Project

The 1000 Genomes Project, ran between 2008-2015, aimed to provide a comprehensive resource of widespread human genetic variability by applying whole genome sequencing to a diverse collection of individuals from multiple populations [6]. The project benefited from the fast development of sequencing technology, which significantly reduced sequencing costs, and resulted in the creation of the largest public catalogue of human variations and genotype data. According to one of the main The 1000 Genomes Project publications released in the journal Nature [7], the last project release called "Phase 3" carried out the genome sequencing of 2,504 individuals from 26 different populations around the world, what can be seen in the map in figure 2.2, and describes a catalogue of 84.7 million variants increasing the number of known variants of the human genome by 40%. Despite this high genetic diversity, the majority of variants (85% of these 84.7 million) are restricted to individuals from one continent only, especially among the sub-Saharan populations of the African continent.



IGSR: The International Genome Sample Resource

Supporting open human variation data

IGSR and the 1000 Genomes Project

Populations: ○ - African; ● - American; ● - East Asian; ● - European; ● - South Asian;

The International Genome Sample Resource (IGSR) was established to ensure the ongoing usability of data generated by the 1000 Genomes Project and to extend the data set. More information is available about the IGSR.

Links

- Announcements
- IGSR Sample Collection Principles
- 1000 Genomes Project Publications
- File formats
- Software tools
- Download data
- Twitter
- IGSR NAR publication

Figure 2.2 The 1000 Genomes official website [8]

The data from the 1000 Genomes Project are freely accessible and can be seen in a genomic context through genomic browsers. The Ensembl database provides a genomic search engine, where the 1000 Genomes Project data can be viewed along with a wide variety of additional data sources. [9,10] The 1000 genomes browser is also hosted on the NCBI and UCSC Genome Browser websites.

2.2.2 The HapMap Project

The International HapMap Project originated in October 2002 with a collaboration between academic research centers, nonprofit biomedical research groups, and private companies from Canada, China, Japan, Nigeria, the United Kingdom, and the United States [11]

The main objective of the HapMap is to determine the patterns of common genetic diversity in the human genome and develop a haplotype map of these patterns by characterizing sequence variants, their frequencies and correlations between them. [12]

By identifying haplotypes and mapping their chromosomal locations, scientists are able to associate genetic variants with specific diseases and disorders. The project is used to investigate the influence of genes on disease and may enhance our ability to choose targets for therapeutic intervention in the future.

It should be mentioned that the 1000 genomes project is displacing the HapMap and it is now considered as a main research standard for population genetics and genomics.[13]

2.2.3 The Human Genome Project

Figure 2.3 Official Logo of the Human Genome Project [14]

The Human Genome Project (HGP) was the international, collaborative research program aimed at determining the human genome sequence and identifying all the genes it contains. It was coordinated between October 1990 and April 2003 by the National Institutes of Health, the U.S. Department of Energy. Additional associates came from Germany, France, United Kingdom, Japan and China, and many technical universities across the United States. [15]

Scientists managed to extract the knowledge they were looking for from the sea of the information and met the main project's objectives:

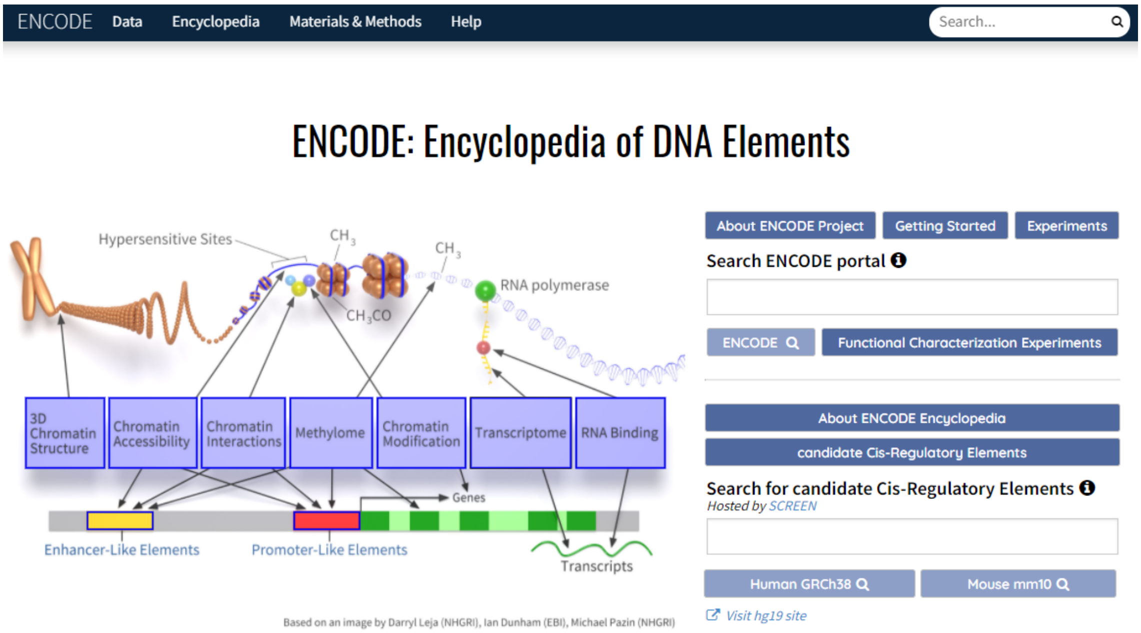
- To identify approximately 20,500 genes encoding proteins
- To establish the sequences of 3 billion chemical base pairs that form human DNA
- To create a database and store all the information from the HGP
- To improve tools for working with data
- To provide associated technology to the private sector
- To handle the ethical, legal and social (ELSI) aspects of the project

There is no doubt that the research was one of the greatest genomics projects and completely revolutionized this field. The HGP gave us a source of detailed information about the complete set of human genes, its structure, organization and function. Thanks to the project, it is known that around 4 000 genetic disorders are linked to a mutation of a particular gene. [15]

The project made scientists aware of the need for qualitative and efficient operation on a huge amount of data. Although the HGP is completed, the analysis of the data will be continued for many years. The collection of genomic data alone is insufficient, because it needs to be properly analyzed and selected, which is the mission of bioinformatics.

2.2.4 The Encyclopedia of DNA Elements

The Encyclopedia of DNA Elements (ENCODE) is a public research project launched by the US National Human Genome Research Institute (NHGRI) in September 2003. The main objective of the ENCODE project is to identify all functional elements in the human genome and thus it is a follow-up to the Human Genome Project. [16]



ENCODE: Encyclopedia of DNA Elements

Search ENCODE portal

Search for candidate Cis-Regulatory Elements

Hosted by SCREEN

Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

Figure 2.4 the ENCODE official website [17]

ENCODE found that 80% of DNA contains elements linked to biochemical functions, dismissing the idea that much of the DNA is simply evolutionary redundant and forcing a redefinition of the gene concept as a minimal inherited unit. Until now science had focused on 2% of the genome, which was thought to have a utility.

What the ENCODE project shows is that DNA and the biochemical regulation of the cell is much more complex than ever imagined, since a large part of non-coding DNA, which is not expressed in proteins, has regulatory functions, so it can be related to diseases and become therapeutic targets. [16]

2.3 Biological background

In order to perform the analysis and moderation of biological data, it is necessary to familiarize with the fundamentals of the field from the molecular biology and genetics perspective. Describing the human genome as a domain in an information system requires the integration of all important concepts into a conceptual schema and this is an essential step in analyzing genetic variation. The domain should be described under principles of the biological structure.

2.3.1 Human genome

A genome represents all the genetic material contained in the cells of a particular organism including both the genes and non-coding sequences. In eukaryotic species, the genome refers to the DNA located in the nucleus and cells' mitochondria (mtDNA) arranged into chromosomes.

The human genome (HG) includes 46 chromosomes, grouped into 23 pairs. These are inherited from both parents, where 22 of them are autosomes and the last pair determines the sex of the individual: XX position in females and XY in males. Each set of 23 chromosomes includes about 3.1 billion DNA sequence bases. [18]

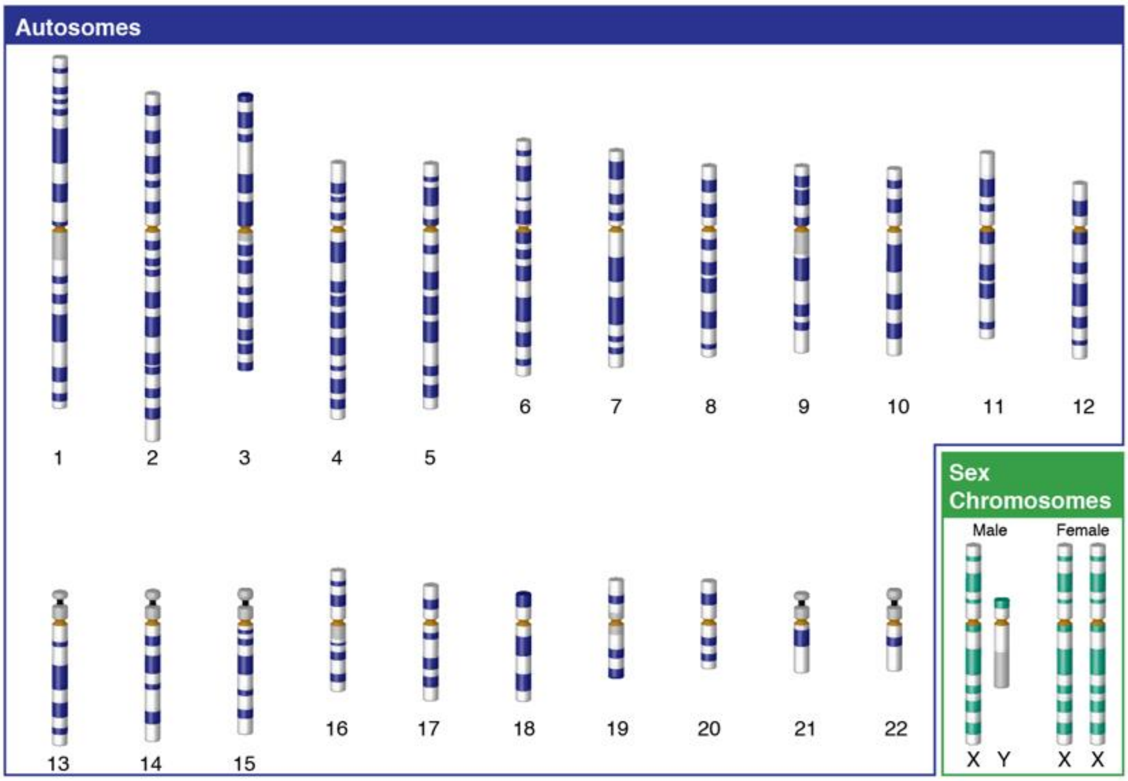


Figure 2.5 Chromosomes of the human genome [18]

HG is composed of approximately 25,000 to 30,000 different genes divided into structural, regulatory or protein-coding. Each of these genes contains the information necessary to form one or more proteins.

Genome holds all the information needed to structure and maintain the organism.

2.3.2 DNA

2.3.2.1 Historical background

DNA was first observed in 1869 by Frederich Miescher, a German biochemist. For many years, scientists were unaware of the great importance of this molecule. It was not until 1953 that James Watson, Francis Crick, Maurice Wilkins and Rosalind Franklin discovered the structure of DNA. Watson, Crick and Wilkins received the Nobel Prize in Medicine in 1962 "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material." [19]

2.3.2.2 Structure

DNA (deoxyribonucleic acid) is a linear, unbranched biopolymer composed of deoxyribonucleotide monomers. The nucleotide, the basic unit of DNA organization, consists of three functional groups [20]:

- A sugar composed of five carbon atoms
- A phosphate group, comprising from one to three linked phosphate units
- A nitrogenous base

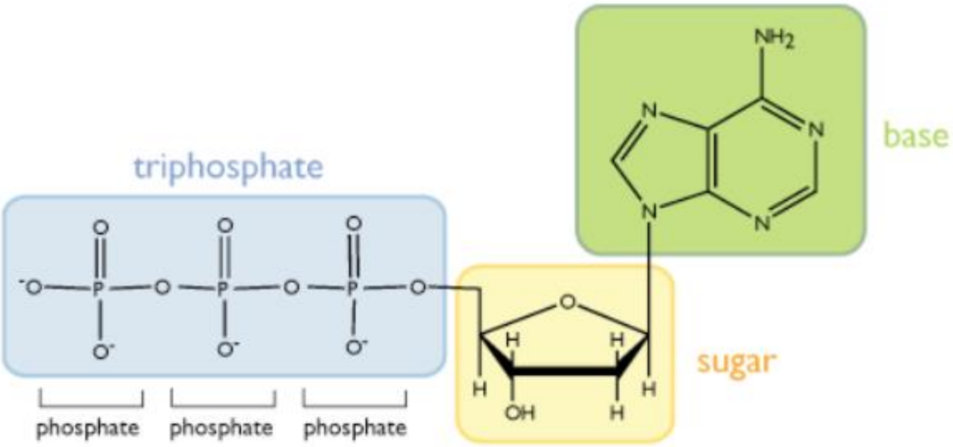


Figure 2.6 Nucleotide structure [19]

The nitrogen-containing bases are the derivatives of purine: adenine (A) and guanine (G) or pyrimidine: cytosine (C) and thymine (T). The DNA-building sugar is called deoxyribose, which, when combined with the nucleobase, forms a nucleoside. The nucleoside bound to the phosphate group creates one of the four types of nucleotides, depending on the attached base: A, G, C, T. [22] Nucleotides therefore differ only in the type of nitrogenous base.

The way in which the nucleotide subunits are bound together gives the DNA strand chemical polarity. Nucleotides, covalently linked in a chain by sugars and phosphates, form an "external skeleton". The phosphate group is attached to the third carbon atom of the first sugar (3') and to the fifth carbon atom of the second sugar (5'), which determines the orientation of the chains from the 5' end with a free phosphate to the 3’ end with a free OH group. [22]

The DNA has a two-row structure in the form of a double helix. According to the DNA structure model proposed by Watson and Crick, the DNA molecule is made up of two long polynucleotide chains, lying opposite each other, coiled around one axis.

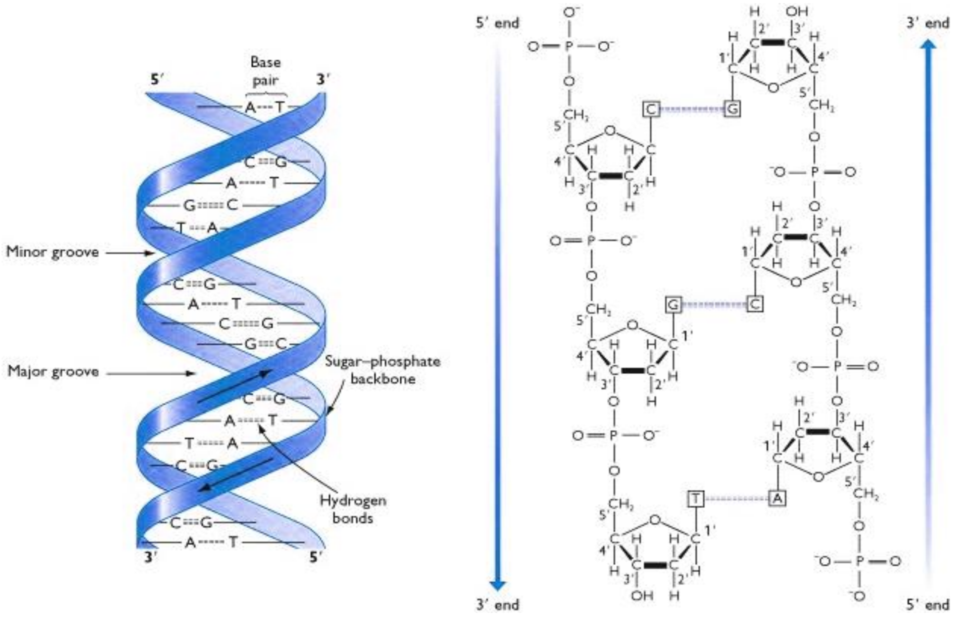


Figure 2.7 DNA structure [20]

DNA is the chemical medium of genetic information in a cell. It serves as a matrix for transcription of information and its replication. The genetic information is encoded by a specific base sequence in each chain, while the sugar and phosphate groups play a structural role. The coded information directs the biosynthesis of enzymes and other proteins and provides information inherited by the progeny cells, conditioning the nature of every living organism. [20]

2.3.3 Gene

The term gene was first introduced in 1909 by Wilhelm Johannsen, Danish botanist, plant physiologist and geneticist.

A gene is a basic unit of heredity that encodes a gene product synthesis, either protein or RNA. From a molecular point of view, it is a nucleotide sequence located in the DNA (or RNA in the case of viruses) which contains the information necessary for synthesis of macromolecule manifesting specific cellular function. [22]

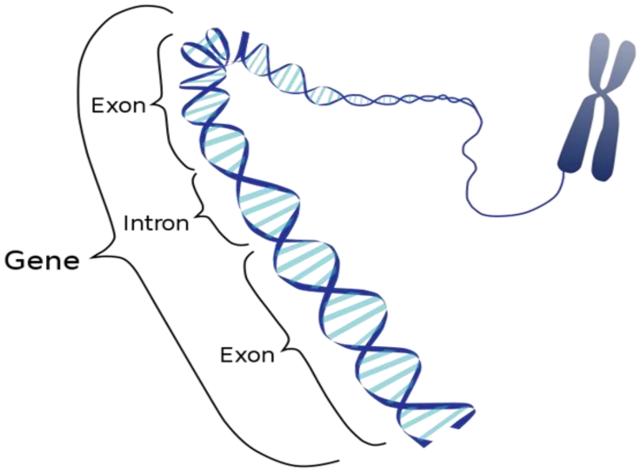


Figure 2.8 Gene with division into exon and intron part. [23]

As it is seen on the figure 2.8, genes and gene segments in sequence are discontinuous. The biological information is divided into coded DNA regions known as exons and non-coded regions called introns. [24] Each of them occupies a specific position in the chromosome called a locus.

A gene is not a structure that can be seen, but rather it is defined on a functional level. In order to know a gene, it is necessary to determine the DNA sequence and the number of nucleotides that forms it. Individual collection of genes forming a particular DNA sequence is called genotype. Genotype together with environmental and developmental factors determine the phenotype which definition is described in the next sections.

2.3.3.1 The function of human genes

Although the functions of all existing genes are not known, they were defined for about half of the 30000-40000 human genes. The vast majority encode proteins, and this is the most intensively studied group of genes. The division of functions for encoding genes is shown in the figure 2.9 based on [25].

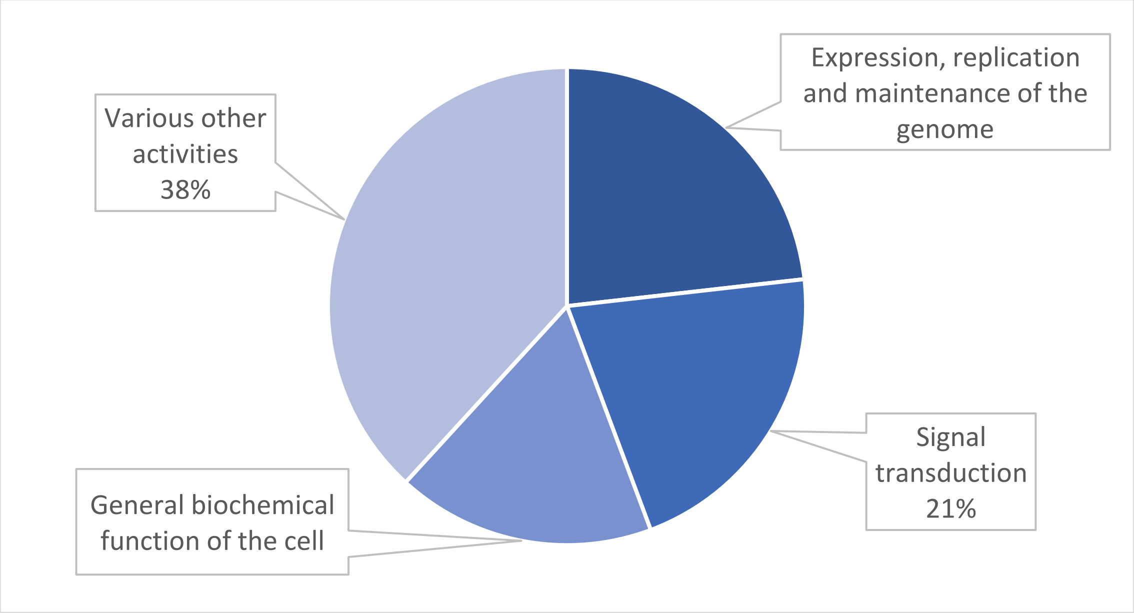


Figure 2.9 Categorization of the identified human protein-coding genes

2.3.3.2 Genomic nomenclature

In the era of widespread genomic information there is need for an efficient and universal scientific language and it is essential for nomenclature to be standardized. A commonly defined and applicable gene nomenclature was established by the HUGO Gene Nomenclature, a Committee (HGNC) of the Human Genome Organization. According to the HGNC guidelines, a unique symbol, HGNC ID and descriptive name are assigned to each gene. [26]

2.3.4 Proteins

Proteins are large, complex macromolecules that perform a vast array of critical roles in the organism, including DNA replication, providing structure and support for cells, transporting molecules, transmitting signals to coordinate biological processes, responding to stimuli and catalyzing metabolic reactions. [27]

Proteins are made up of one or more long chains of amino acid residues build up from hundreds or thousands of smaller units called amino acids. There are 20 different amino acid types that can be combined to form a particular protein. The sequence of amino acids in a chain determines a three-dimensional structure and its specific function to each protein.

Most genes in a DNA sequence carry information to encode protein molecules. The protein production is multilayered process completed through two major phases: transcription and translation.

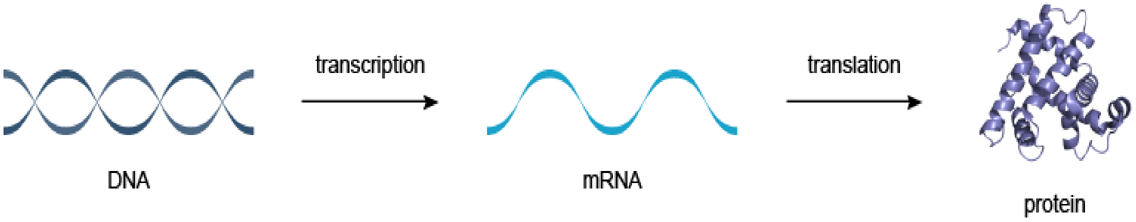


Figure 2.10 The schema of transcription and translation process [28]

Transcription is the process of copying the fragment of DNA code taking place in nucleus. The transcribed DNA message, called also RNA transcript or mRNA (Messenger Ribonucleic Acid), is then transported to the site where Translation process takes place. In case of eukaryotic cells, there are different RNA polymerase molecules types used to transcribe the DNA. Depending on the gene types: RNA polymerase I transcribes genes coding for ribosomal RNAs, RNA polymerase II is used for genes coding proteins and RNA polymerase III for genes that code for transfer RNAs.

Translation is a conversion of the message coded in mRNA to a protein. This takes place in the cytoplasm within a complex called ribosome. The RNA is read in three-letter combinations of nucleotides known as codon. Each codon determines which amino acid is placed to form the protein and encodes a start or stop signal of translation. A physical link between the mRNA and the amino acid sequence of proteins is provided by a tRNA (Transfer Ribonucleic Acid). The process of translation occurs in three main stages: initiation, elongation and termination. [29]

2.3.5 Alleles

An allele is each of the alternative forms of the same gene that differ in some part of the nucleotide sequence and that may be manifested as a specific modification of the gene's functions. Human being inherits two alleles for each gene, one from biological mother and one from biological father. Each pair of alleles is located in the same place within the chromosome.

There are dominant or recessive alleles that determine the manifestation of the characteristic for a certain function or feature. Alleles that express themselves are known as the dominant while the ones unable to express are recessive. If the alleles determining the trait are the same for the particular genotype it is called homozygous. If they differ, the genotype is defined as heterozygous. The summary of a given allele manifestation is presented in the table below. [25]

Table 2.1 Concept of the dominant and recessive inheritance

Condition	Allele 1.	Allele 2.	Phenotype
Homozygous	Dominant	Dominant	Dominant
Heterozygous	Dominant	Recessive	Dominant
Homozygous	Recessive	Recessive	Recessive

Dominant and recessive inheritance are essential concepts in terms of predicting the probability of an individual inheriting certain phenotypes. This is especially important in field of genetic disorders as some diseases are caused by mutated alleles called also genetic variations. [25]

From the information above it can be concluded that one copy of the dominant mutated allele results in a mutated phenotype. For instance, the recessive allele can only cause the disease in individuals that are homozygous to the mutated gene, i.e. both alleles must carry the mutation.

2.3.6 Haplotype

A haplotype is defined as the combination of alleles that are inherited as a unit from a single parent. The term also refers to variations at single positions in the DNA sequence among individuals known as single nucleotide polymorphism (SNPs). SNP is the most common type of genetic variation among people. [30]

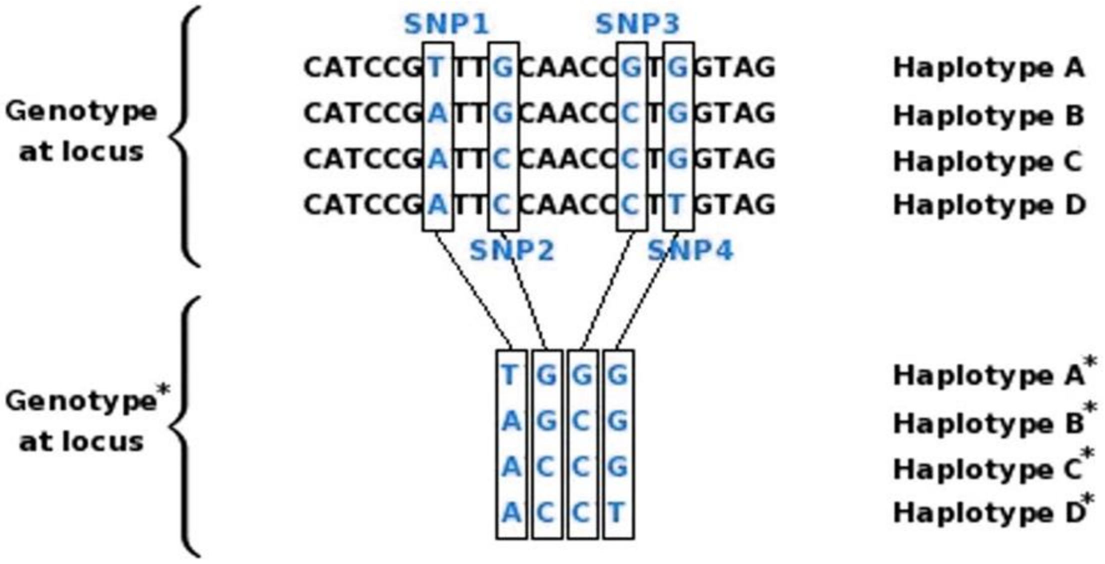


Figure 2.11 Haplotype structure [31]

This concept has been widely used in population-based genetic studies due to the association between SNP and the occurrence of certain diseases. Based on haplotypes, scientists can identify patterns of genetic variation and determine the genes responsible for disease onset.

There are approximately 4 to 5 million SNPs in a person's genome. These variations may be unique or appear in many individuals. More than 100 million SNPs have been found in populations around the world, leading to the conclusion that SNP alleles common in one ethnic group may be less frequent in another. [30]

For a given population, the SNP is assigned a value of minor allele frequency (MAF) which is defined as the second most common allele frequency at a locus that occurs in a particular population. MAF provides information to distinguish the common and rare variants in the population and therefore it is often used in population genetics studies as a criterion for high-quality selection of genetic variants. For example, this criterion is included in the recommendations for the assessment of pathogenicity of variants in clinical testing established by the American College of Medical Genetics and Genomics and the Association of Molecular Pathologists (ACMG/AMP). [32] ACMG/AMP is considered in the methodological part of the work.

2.3.7 Mutations

In accordance with [33] the mutation is "a permanent, heritable change in the nucleotide sequence of a chromosome, usually in a single gene; commonly leads to a change in or loss of the normal function of the gene product".

Mutations may include small changes in a single DNA building block or nucleotide base. However, larger mutations can affect various genes on the chromosome [34]. Based on the effect on the gene structure, mutations are divided into two main types:

1. Small scale mutations define mutations with a single nucleotide base change, insertion or deletion. Usually take place during DNA replication. According to the categories created by the German molecular biologist Ernst Freese point mutation are divided into:
    a) Transitions – replacement of a purine base (i.e., adenine [A] or guanine [G]) with a different purine base or a pyrimidine base (i.e., thymine [T] or cytosine [C]) with different pyrimidine base
    b) Transversions – replacement of a purine base with a pyrimidine base or pyrimidine base with a purine base

Point mutations are also categorized according to their functionality:

    a) Substitution:
        - A nonsense mutation - occurs with a substitution of one nucleotide leading to the formation of a stop codon instead of coding codon.
        - A missense mutation - occurs with a substitution of one nucleotide and formation of a different codon
        - Silent mutation - codes for the same amino acid and does not affect the protein function.
    b) Insertion and deletion
        - Frameshift mutation - occurs by inserting or deleting one or more DNA bases and result in change of the reading frame. Both of the types can cause a drastic loss of function which is why insertion and deletion are grouped together.

2. Large scale mutation refers to changes in the structure of the whole chromosome or its part and can result in major phenotypic consequences. The large-scale mutations can be divided into:
    a) Inversion - a segment of a chromosome is reversed end to end
    b) Deletion - a chromosome segment is deleted from the sequence
    c) Translocation - segments of non-homologous chromosomes are rearranged

d) Insertion - a chromosome segment is added to the sequence

When discussing mutations, the term Single nucleotide polymorphisms (SNPs) should be explained. SNP is described as the phenomenon of DNA sequence variability leading to change of a single nucleotide (A, T, C or G) between individuals of a given species.

SNPs represent about 90% of the total variability found in the human genome and are considered to be the most extensive source of human genetic variability. Polymorphism is understood as a variant with a frequency above 1%. [34]

According to the recommendation of American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) the terms “mutation” and “polymorphism” are replaced by the term “variant” with the following modifiers: pathogenic, likely pathogenic, uncertain significance, likely benign, or benign. The aim is to standardize terminology and to prevent incorrect assumptions about pathogenicity which facilitates variant data analysis. [32]

2.3.8 Phenotype

Phenotype is a set of characteristics, including not only morphology, but also e.g. physiological properties, fertility, behavior, ecology, life cycle, biological changes, environmental impact on the body. It describes complex observable traits or patterns of an organism. [35]

2.4 From Big Data to Smart Data

An important consequence of the dynamic development of sequencing technologies is a tremendous amount of data produced each day.

The development has led to the "omic age", an era in which a global vision of biological processes is based on the analysis of a large volume of data, and therefore bioinformatics support is needed to distinguish and interpret the results.

According to the report published by META group [36] Big Data is described by three Vs:

- Volume - indicates the size of the data.
- Velocity - informs about the data generation speed.
- Variety - informs about different types and sources of data.

However, based only on these three characteristics, it is not possible to obtain high quality results suitable to be used in a clinical practice and therefore two additional characteristics should be introduced:

- Veracity - informs about the reliability of the data obtained.
- Value - informs about utility of data.

Data characterized by all these factors is defined as Smart Data and an essential step to achieve it is to apply conceptual modelling and data quality management techniques. [37]

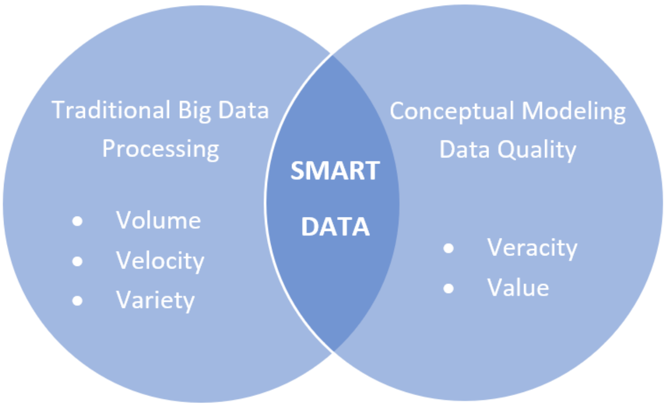

Traditional Big Data Processing
Volume
Velocity
Variety
SMART DATA
Conceptual Modeling Data Quality
Veracity
Value

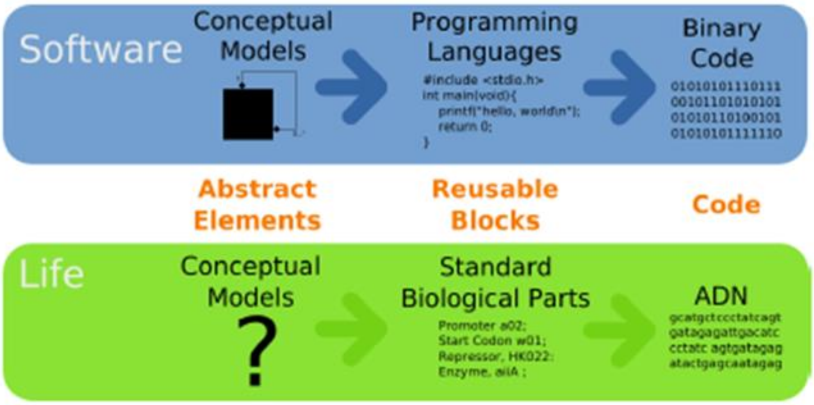
Figure 2.12 The 5 Vs of the Smart Data Perspective

2.5 The Conceptual Schema of the Human Genome

Conceptual modelling refers to "activity that elicits and describes the general knowledge an Information System needs to know". [38] The CM can represent a system by using concepts and ideas to help understand and manage complex domains. This representation of system is called a conceptual schema.

Conceptual modelling is used in the different areas, from socioeconomics to software development.

The figure below shows the application of the conceptual model in the IT field for software development. In addition, it confronts the software with the biological model, indicating similarities between the programming code and the DNA code.



```
#include <stdio.h>
int main(void){
    printf("hello, world\n");
    return 0;
}
```

```
01010101110111
00101101010101
01010110100101
01010101111110
```

```
Promoter a02;
Start Codon w01;
Repressor, HK022:
Enzyme, aiiA ;
```

```
gcatgctccctatcagt
gatagagattgacatc
cctatc agtgatagag
atactgagcaatagag
```

Figure 2.13 Relation between Software and Life perspective [39]

This approach is also used to define a conceptual model that represents the characteristics and behavior of the human genome.

The Conceptual Schema of the Human Genome (CSHG) has been designed by the PROS Research Centre at Universitat Politècnica de València (UPV) in response to the lack of a united conceptual perspective on the genomic data representation.

Managing a large collection of diverse data sources with a large amount of data representing knowledge related to continuous evolution is a challenge that bioinformatics faces. In a context such as genomics, in which knowledge is constantly changing, the CMGH perspective appears to be fundamental for data integration and subsequent analysis.

The current developed version of the CSHG is organized in five main views [40]:

- The Structural View - describes the structure of the genome, it is composed by the basic elements of the DNA sequence.
- The Transcription View - shows the components and concepts related to protein synthesis.
- The Variation View - characterizes the changes in the reference sequence.
- The Pathway View - provides information about the metabolic pathways.
- The Bibliography and Data Bank View - evaluates every information source in order to establish from where each data comes from.

The model is essential during the process of identifying variants related to a given disease. It establishes the framework for the information search and therefore sets the conditions that repositories providing data must meet.

The following figure shows a simplified CSHG model necessary to identify the relationship between genetic variant and the disease.

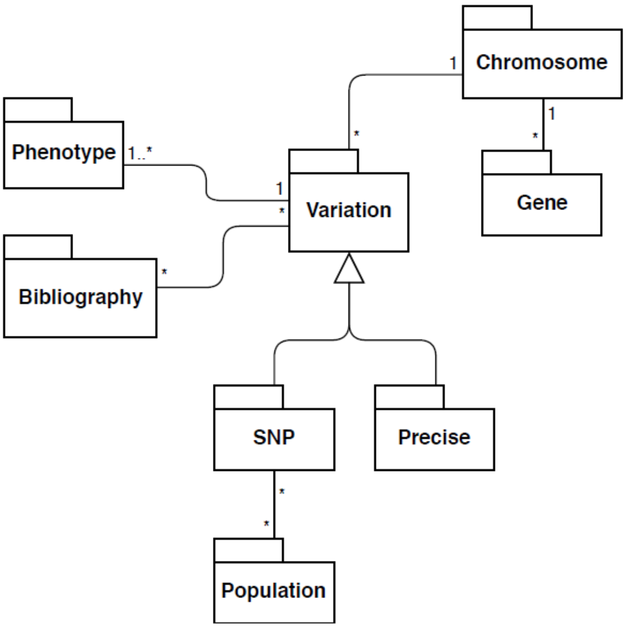


Figure 2.14 Simplified view of the CSHG [41]

CHAPTER 3. PHENOTYPES USED

Most neuropsychiatric disorders are moderately to highly heritable and the probability of their occurrence depends on genetic factors. Researchers are intensively studying the degree to which genetic variation is unique to individual disorders or shared across more than one.

Although information is still incomplete, there are studies supporting the existence of common genetic factors for the three disorders being the subject of this work: autistic spectrum disorder (ASD), attention deficit hyperactivity disorder (ADHD) and schizophrenia.

Investigations indicate that about 20-80% of children with ASD show also symptoms of ADHD. [42] A significant number of clinically diagnosed children with ADHD meet the criteria for ASD. [43] Besides ADHD possible manifestations in adulthood [44], the severity of ASD seems to be consistent with the co-existence of ADHD symptoms. In addition, the relationship between the occurrence of autistic symptoms and schizotypal traits is observed within individual, which means that patients with diagnosed ASD show features of the schizophrenia spectrum in adolescence. [45]

Disorders share common factors in terms of neurocognitive and clinical deficits. People suffering from ASD, ADHD and schizophrenia may have similar deficits in cognitive function, language skills and social cognition. [46,47] It can be observed that individuals with ASD and ADHD exhibit mutual emotional and performance dysfunction [48], while those with ASD and schizophrenia have somatomotor deficits and share similar patterns of social retardation and poor eye contact. [49]

From a neurobiological point of view, there are common phenotypic traits that intersect in diagnoses, with converging findings of cerebral connectivity abnormalities across mentioned disorders. [50-53]

It is fundamental that each of these diseases originates from the same nervous system-related category as shown in the diagram below.

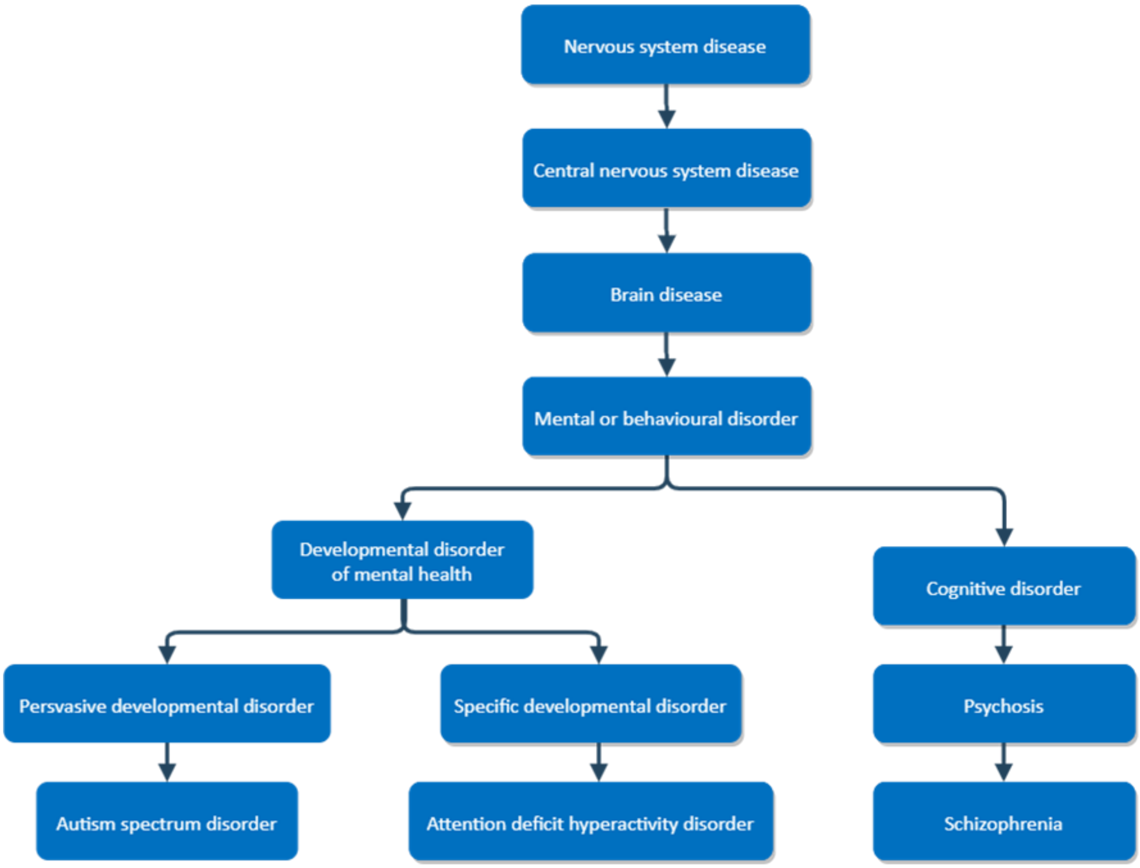

Nervous system disease
Central nervous system disease
Brain disease
Mental or behavioural disorder
Developmental disorder of mental health
Cognitive disorder
Persvasive developmental disorder
Specific developmental disorder
Psychosis
Autism spectrum disorder
Attention deficit hyperactivity disorder
Schizophrenia

Figure 3.1 The branch of the analyzed diseases.

3.1 Autism spectrum disorder

3.1.1 Evolution of the diagnostic concept

What we define today as Autism Spectrum Disorder was first described in 1943 at Johns Hopkins University School of Medicine in Baltimore by an Austrian-American psychiatrist, Leo Kanner. [50] Among children with various developmental disorders, psychoses, mental retardation, he noticed several cases with a common communication deficit and unique patterns of behavior. Kanner summarized diagnosis as “children’s inability to relate themselves in the normal way to people and situations from the beginning of life” [55] and named the syndrome "early infantile autism". Kanner's observations were very detailed and many of them remain valid. Children who show the most typical symptoms of autism: avoidance of social relations and emotional connections are classified to have classic or Kanner’s autism. [56]

Just one year later, pediatrician Hans Asperger at the University Children’s hospital in Vienna published an article describing a group of children with the symptoms, partially mentioned before by Kanner, of what he called “autistic psychopathy” [57].

Despite the enormous contribution to the science of ASD and the description of the first study of the syndrome later named after his name, Asperger's article written in German went almost unnoticed whereas Kanner's paper became highly cited. It slowed down the development of knowledge about the disease and after almost 40 years, in 1981, child psychiatrist Lorna Wing at the UK's Institute of Psychiatry in London published an article that brought Asperger’s syndrome to the world's attention. [58]

Throughout the years, there have been many changes in the way ASD is perceived and it has undergone many updates to the definition and description. Until 1980, the American Psychiatric Association, the main professional association of psychiatrists and psychiatric students in the United States and the largest of its kind in the world, characterized ASD as a form of childhood schizophrenia and only then introduced the term "total developmental disorder" separating these two diseases.

Further research in various fields, such as psychiatry, neurology and genetics, has progressively improved the perception of ASD.

The most recent version of the diagnostic criteria for classifying ASD and other disorders was published in the fifth edition of The Diagnostic and Statistical Manual of Mental Disorders (DSM-V).

3.1.2 Definition

Autism spectrum disorder is one of the most prevalent group of neurodevelopmental disorders and refers to a broad range of conditions characterized by the presence of three core symptoms with varying degree of severity [59]:

- Impairment in social interactions
- Abnormal development and linguistic abilities
- Repetitive patterns of behavior, interests and activities

The prevalence of ASD in the general population is estimated to be around 1-2% with the average male to female ratio 4-5:1. [60]

ASD symptoms are present from early childhood and limit or impair daily functioning. Communication and social impairment of patients with ASD may lead to delayed or absent speech, unusual vocabulary use, aggressive or disruptive behavior and persistent echolalia that hinder human interaction. Occurrence of interest impairments is manifested by inability to cope with change, unusual response to sensory stimuli and repetitive use of objects. Intellectual and language impairment causes slow talking and motor deficits. [49]

Individuals affected by ASD are prone to anxiety and depression.

Spectrum of autistic disorders encompasses disorders that were previously characterized as atypical autism, childhood autism, infantile autism, Kanner's autism, high-functioning autism, pervasive developmental disorder not otherwise specified (PDD-NOS), childhood disintegration disorders and Asperger's disorders. [49]

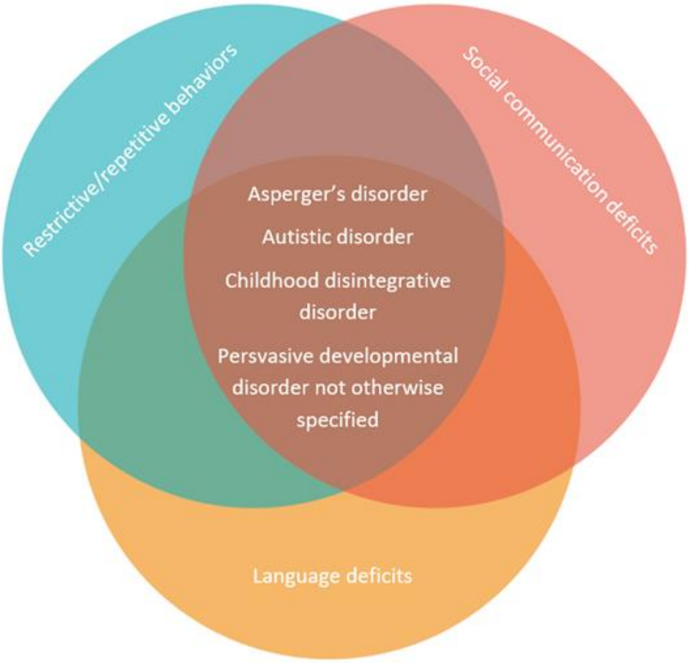

Restrictive/repetitive behaviors
Social communication deficits
Asperger's disorder
Autistic disorder
Childhood disintegrative disorder
Persvasive developmental disorder not otherwise specified
Language deficits

Figure 3.2 Autism Spectrum Disorder

The complexity of the criteria for classifying ASD allows some other diseases to meet certain conditions. It should be remembered that if only part of the criteria is met, such diseases cannot be classified into the spectrum, but may occur simultaneously. For instance, in case of Rett Syndrome autism spectrum disorder should be considered only when all diagnostic criteria are met. According to the DSM-5 [49], the same applies to Attention-deficit/hyperactivity disorder and Schizophrenia.

In accordance with Online Mendelian Inheritance in Man (OMIM) there are 31 ASD subcategories and each one has the appropriate acronym and alternative names (see Appendix A). In the context of searching for information about the disease in databases, such information is very important.

3.1.3 Risk factors and genetics

The risk factors for ASD are both environmental and genetic. Determining which factor influences the development of the disease is challenging and the evidence vary between studies. Accumulating evidence suggests advanced parental age, low birth weight and fetal exposure to valproate to have contribution in the etiology of ASD.

ASD is a complex disorder with strong genetic factors. According to twin and population-based studies, the heredity of ASD may vary between 50-95% [60-62]. Over the last ten years of ASD genetic research, many different approaches have been used, including whole genome sequencing (WGS) [63], whole-exome sequencing (WES) [64], genome-wide association study (GWAS) [65] and copy number variations (CNVs) analysis [66]. The set of these investigations provided a picture of ASD genetic etiology [67].

ASD can be caused by a mutation of one of many genes. The number of reported genetic changes associated with ASD is very high, reaching 1000 genes. However, there are many unconfirmed associations and not identified variants. The impact of a genetic variant on the disease development is uncertain due to co-occurring environmental factors.

3.2 Attention deficit hyperactivity disorder

3.2.1 Evolution of the diagnostic concept

The earliest description of symptoms and behavior similar to what is currently recognized as part of ADHD was made by the German physician Melkior Adam Weikard in 1775. Weikard reported the symptoms of inattention and distractibility – Attentio Volubilis, found surprisingly in adults, similar to the inattentive ADHD type [68].

In 1798, the Scottish physician Alexander Crichton published a textbook An inquiry into the nature and origin of mental derangement : comprehending a concise system of the physiology and pathology of the human mind, and a history of the passions and their effects. [69] He described alterations of attention and “mental restlessness” that corresponds to some of the symptom descriptions of ADHD in the DSM-IV.

Regardless of these early descriptions of symptoms similar to ADHD, most researchers consider that the scientific starting point of ADHD recognition is the report of the British pediatrician Sir George Frederic Still made in 1902. He provided a clinical panorama of children with a “defect of moral control” showing problems with sustained attention. [70]

In 1937 Charles Bradley described the first successful treatment of children's hyperactivity. Children with behavioral difficulties were treated alongside children suffering from neurological diseases due to a growing belief in the relationship between hyperactivity and possible brain damage. [71] 10 years later the term "minimal brain damage" was introduced. The concept referred to the impairment of brain function that affects perception and behavior.

In 1980, this concept was introduced into the diagnostic entity "attention deficit disorder" (ADD) in DSM-III. [72] The term “Attention Deficit Hyperactivity Disorder” was firstly introduced in the revised version of DSM-III from 1987. [73]

3.2.2 Definition

Attention-deficit/hyperactivity disorder is one of the most common neurodevelopmental disorders under characterized by age-inappropriate levels of inattention and/or hyperactivity/impulsivity. [49,74] ADHD is considered to manifest mostly in children, with prevalence of 5.3% [75]. Nevertheless, up to 15% of individuals diagnosed in childhood meet the clinical criteria of ADHD also in adulthood, with approximately 65% showing symptoms of ADHD, that do not meet diagnosis criteria [76].

Based on the Classification of the American Psychiatric Association [49], ADHD is divided into three different types, which are shown in the graphics below:

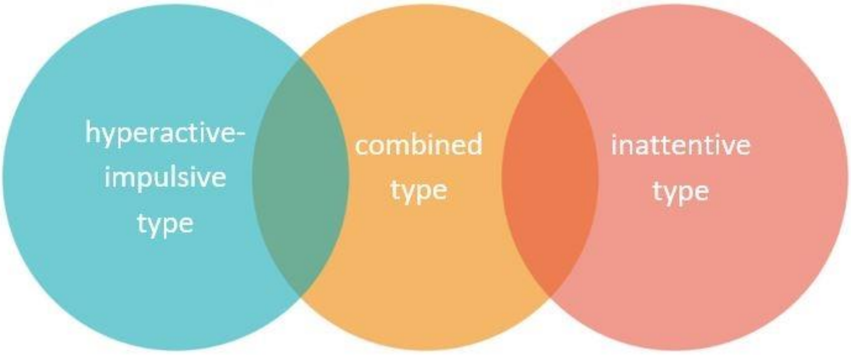


Figure 3.3 Three types of ADHD.

- Hyperactive/impulsive type (314.01 F90.1) - the least common type characterized by impulsive and hyperactive behaviors without inattention and distractibility.
- Combined type (314.01 F90.2) - the most common type characterized by impulsive and hyperactive behaviors as well as inattention and distractibility.
- Inattentive type (314.00 F90.0) - characterized by inattention and distractibility.

Most people with ADHD experience co-existing conditions such as mood or anxiety disorders, insomnia, learning disabilities or substance use disorders. Individuals affected by the disease may also suffer from Autism Spectrum Disorder. [75]

In accordance with OMIM there are 8 ADHD subcategories (see Appendix A).

3.2.3 Risk factors and genetics

ADHD risk factors include temperamental, environmental and genetic factors.

The first group includes reduced behavioral inhibition effortful control, negative emotionality and elevated novelty seeking. Environmental refers to low birth weight and smoking during pregnancy.

Molecular genetic research reveals the complexity of ADHD genetic architecture.

GWAS along with other studies show that the genes regulating directed neurite outgrowth have a strong link to ADHD etiology.[77]

Numerous investigations around the world show that symptoms of ADHD are highly heritable. From further GWAS analyses it can be concluded that a large part of ADHD inheritance is due to the polygenic effects of many common variants, which individually have a small impact. [78]

Dysfunctions can be shared in the family due to the same environment as well as genes. Adoption studies show that within the family there are important genetic determinants influencing ADHD, coexisting with environmental factors. Biological relatives of children with ADHD are more likely to have ADHD than adopted ones. [79] There are also studies on individual genes indicating a consistent pattern of preferential transmission of risk alleles by the father to children with ADHD. [80]

It is well established that the rate of ADHD is much higher in males than females.

3.3 Schizophrenia

3.3.1 Evolution of the diagnostic concept

The term "schizophrenia" is less than 100 years old. But even before that time, cases were described which today would be attributed to this disease. [81]

Schizophrenia was first described by German psychiatrist, Dr Emil Krapelin in 1887. For the individuals with the symptoms of what we called schizophrenia, he used the term “dementia praecox”.

First person who used and coined the term “schizophrenia” was Swiss psychiatrist, Eugen Bleuler, in 1911. He also divided symptoms of the disease to "positive" or "negative." [82]

The first drug for treating schizophrenia was identified in 1952 by Henri Laborit, a surgeon from France.

The significance and classification of the disease has changed over the years. DSM-III specified 5 subtypes of schizophrenia: disorganized, catatonic, paranoid, residual, and undifferentiated. [72] However, the subtypes have not proved useful and valuable in determining the condition and no longer appear in DSM-V. [49]

3.3.2 Definition

Schizophrenia is a common type of psychotic disorders that occurs in approximately 1% of general population. Schizophrenia is normally diagnosed between the age of the late teen and the early thirties. [83]

Is defined by abnormalities in one or more of the following five domains:

- Delusions
- Hallucinations (involving hearing voices)
- Disorganized thinking (speech)
- Grossly disorganized or abnormal motor behavior (including catatonia)
- Negative symptoms (lack of motivation, social withdrawal)

Schizophrenia often occurs with other psychotic disorders such as mood disorders including depression and bipolar disorder. Individuals with both mood disorders and schizophrenia are often given the diagnosis of schizoaffective disorder. [49]

Schizophrenia not only affects mental health, but statistics show that individuals suffering from the disease die 12-15 years earlier than the average population. About 5-6% of patients suffering from schizophrenia die from suicide. [84]

Patients with schizophrenia show deficits on executive functions being responsible for controlling behavior, attention, cognitive and information processing deficits and memory problems. All of them leads to functional consequences including educational difficulties, problems with handling easy tasks, lack of motivation, irritability, depressed mood and limited social relations. Some features co-occur with other diseases, for example, disorganized speech also appears in ASD and disorganized behavior in ADHD. [49]

In accordance with OMIM there are 20 Schizophrenia subcategories (see Appendix A).

3.3.3 Risk factors and genetics

As with previous mental disorders, the development of schizophrenia is influenced by the interaction of genetic and environmental risk factors.

Schizophrenia development probability can increase due to pregnancy and birth complications (low birth weight, premature labor), paternal age, autoimmune diseases, substance abuse, stress, urbanization, poverty, and social exclusion. [49]

Studies indicate that the heredity of schizophrenia is estimated between 70-80%. [85] The highest single risk factor for developing schizophrenia is having a first-degree relative with the disease. The risk can be almost 50% in the case of both affected parents. If one parent is affected the risk for their offspring is approximately 13%. [86]

CHAPTER 4. SILE METHODOLOGY

The successful identification of new genes and variants associated with the risk of suffering from a particular mental illness is highly dependent on our ability to collect and combine all relevant data. The various genomic repositories provide the great amount of ready-to-examine data, which differs significantly in terms of content, resources, infrastructure, information quality and relevance to clinical practice.

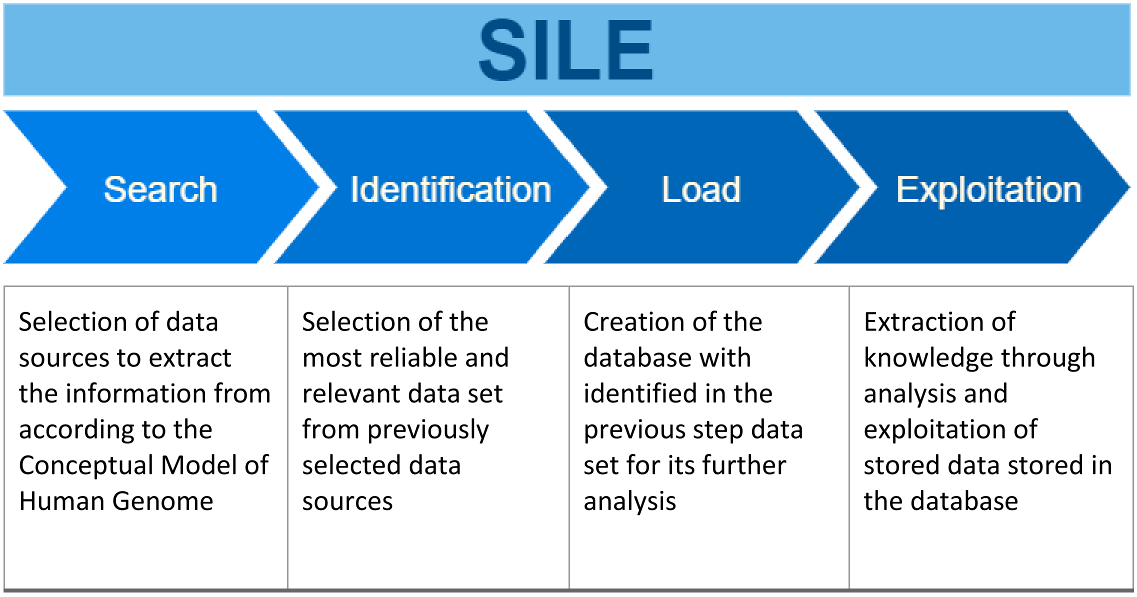
The bioinformatics domain requires strict control in the manipulation of data and its correct interpretation to ensure high accuracy of the results. That is why it is crucial to use a methodology based on well-established principles of data quality management, providing the selection of appropriate repositories and the most valuable data from the entire available data set.

Considering the importance of the quality of selection and analysis of genomic information, the Search-Identification-Load-Exploitation (SILE) methodology has been implemented in the current work.

The SILE method was developed by Óscar Pastor López within the research institute, mentioned already in the thesis, The Research Center of Software Production Methods (PROS) at the Universitat Politècnica de València to improve the process of loading genes and variants of the

Human Genome Database (HGDB). [87] Its creation was a response to the lack of an existing and tested method of managing genomic databases in order to select the most appropriate variants. Its most comprehensive and elaborate version has recently been presented in Ana León's Doctoral Thesis. [41]

The method is based on the idea of a conceptual model and the foundations of data quality management adapted to the genome domain. SILE consists of four pillars S - Search, I - Identification, L - Load and E - Exploitation, which main purpose is to systematize the process of identifying genetic variants associated to a disease.



Selection of data sources to extract the information from according to the Conceptual Model of Human Genome	Selection of the most reliable and relevant data set from previously selected data sources	Creation of the database with identified in the previous step data set for its further analysis	Extraction of knowledge through analysis and exploitation of stored data stored in the database

Figure 4.1 SILE methodology schema

Each of the pillars is described in more detail in the next part of the work.

SILE has already been applied in the study of genetic information related to different human diseases such as congenital cataract [87], alcohol sensitivity [88] and neuroblastoma [89]. Currently, it is also being used to identify genetic variations of Crohn’s Disease, migraine, epilepsy, and breast cancer. [37]

4.1 Search

This stage of the SILE method is designed for the selection of genetic repositories providing the information from which it is possible to obtain quality data that can be used in the clinical environment.

When searching for information, it is necessary to select and analyze databases in order to access the most suitable ones to identify the phenotype-related variants for classification. There are certain libraries that facilitate the search for relevant databases like Human Genome Variation Society (HGVS), or Nucleic Acids Research (NAR).

- Nucleic Acid Research is a publicly available, reviewed scientific journal published by Oxford University Press. The NAR “*publishes the results of leading-edge research into physical, chemical, biochemical and biological aspects of nucleic acids and proteins involved in nucleic acid metabolism and/or interactions*”. [90] In addition, the institution collects the information about biological databases and classifies them into 15 different categories which give a total of 1,656 repositories. The section relevant to my thesis is defined as *Human Genes and Diseases: General polymorphism databases.*

- Human Genome Variation Society is intended to encourage the discovery of new genomic variations and manages a large catalogue of databases with the aim of promoting the collection, documentation and free distribution of information on genomic variants. The biological bases are classified in 12 sections of which those of interest are *Disease Centered Central Mutation Databases* and *Central Mutation & SNP Databases.* Currently, the society disposes a collection of links to 1,750 databases. [91]

The process of selecting the repositories is based on Data Quality Methodology proposed in [37]. The Methodology is divided into 5 phases described in the table below.

Table 4.1 Data Quality Methodology

Phase no.	Name	Description
Phase I	Dimention Description	Describes the dimensions of interest to be measured.
Phase II	Metric Description	Describes the metric associated to each dimension.
Phase III	Variable Selection	Select the attributes based on the HGCM
Phase IV	Minimum DQ criteria	Establishes the minimum requirements that the variables must meet.
Phase V	DQ Assessment	Compares the information from the databases with the minimum DQ requirements

Based on the methodology presented, a set of relevant and interesting DQ dimensions should be identified as a starting point for the verification of the repositories. The publication [37] specifies 8 most important metrics to be validated.

Table 4.2 Description of the metrics used for the selection of the databases.

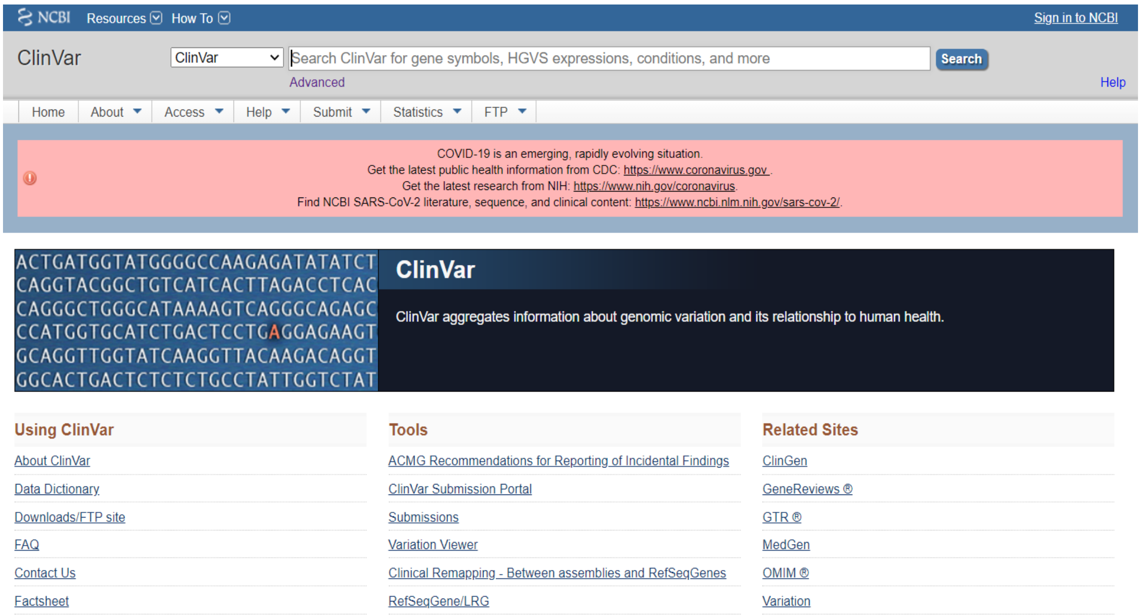
Believability	M1	The stored information must be manually curated or revised by experts.
	M2	The quality controls must be performed to ensure the accuracy of the information presented.
Relevancy	M3	The database must provide sufficient and useful information to determine the necessary data in accordance with the attributes of the CSHG.
Reputation	M4	The database must be maintained by research centers, institutions or associations of an international and well-known level.
Currency	M5	The database has to be regularly maintained and updated.
	M6	The data must be available to the public and freely accessible.

Accessibility	M7	The database must include download mechanisms for the requested information.
	M8	It is significantly recommended that the database enables programmatic access to the stored information through appropriate tools.

Clinvar

Clinvar is freely available, public archive of human genetic variants and interpretations of their significance to human health and disease. [92]

The first public release of the ClinVar took place in April 2013 at the National Center for Biotechnology Information at the National Institutes of Health (NIH) to provide a centralized, public open-access database with the aim of better availability, management, maintenance and distribution of variation-phenotype relation data. [93]


NCBI
Resources
How To
Sign in to NCBI
ClinVar
ClinVar
Search ClinVar for gene symbols, HGVS expressions, conditions, and more
Search
Advanced
Help
Home
About
Access
Help
Submit
Statistics
FTP
ACTGATGGTATGGGGCCAAGAGATATATCT
CAGGTACGGCTGTCATCACTTAGACCTCAC
CAGGGCTGGGCATAAAAGTCAGGGCAGAGC
CCATGGTGCATCTGACTCCTGAGGAGAAGT
GCAGGTTGGTATCAAGGTTACAAGACAGGT
GGCACTGACTCTCTCTGCCTATTGGTCTAT

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/.

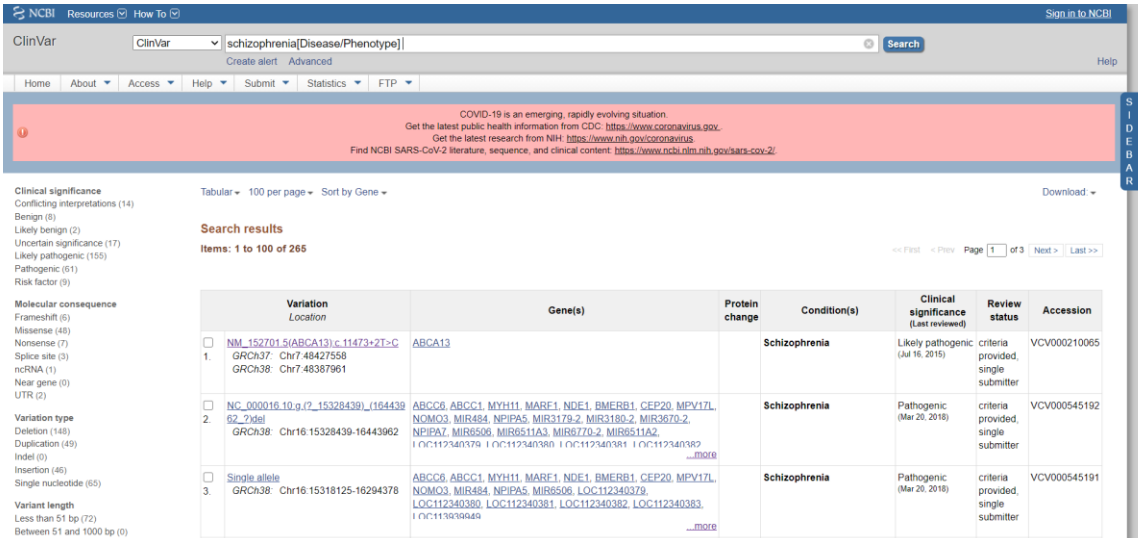
ClinVar

ClinVar aggregates information about genomic variation and its relationship to human health.

Using ClinVar	Tools	Related Sites
About ClinVar	ACMG Recommendations for Reporting of Incidental Findings	ClinGen
Data Dictionary	ClinVar Submission Portal	GeneReviews ®
Downloads/FTP site	Submissions	GTR ®
FAQ	Variation Viewer	MedGen
Contact Us	Clinical Remapping - Between assemblies and RefSeqGenes	OMIM ®
Factsheet	RefSeqGene/LRG	Variation

Figure 4.2 Clinvar homepage [94]

The information in this repository can be obtained by using the browser and a special query. To look for variants of a particular disease or phenotype, for example schizophrenia, after the name of the disease there should be [Disease/Phenotype] command posted. The data can be downloaded in XML file. The search provides the information for each variant divided in columns: Variation, Gene(s), Protein Change, Conditions(s), Clinical Significance, Review, and Accession.



schizophrenia[Disease/Phenotype]

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/.

Clinical significance
Conflicting interpretations (14)
Benign (8)
Likely benign (2)
Uncertain significance (17)
Likely pathogenic (155)
Pathogenic (61)
Risk factor (9)

Molecular consequence
Frameshift (6)
Missense (48)
Nonsense (7)
Splice site (3)
ncRNA (1)
Near gene (0)
UTR (2)

Variation type
Deletion (148)
Duplication (49)
Indel (0)
Insertion (46)
Single nucleotide (65)

Variant length
Less than 51 bp (72)
Between 51 and 1000 bp (0)

Search results

Items: 1 to 100 of 265

Page 1 of 3

	Variation *Location*	Gene(s)	Protein change	Condition(s)	Clinical significance (Last reviewed)	Review status	Accession
□ 1.	NM_152701.5(ABCA13):c.11473+2T>C *GRCh37:* Chr7:48427558 *GRCh38:* Chr7:48387961	ABCA13		**Schizophrenia**	Likely pathogenic (Jul 16, 2015)	criteria provided, single submitter	VCV000210065
□ 2.	NC_000016.10:g.(?_15328439)_(16443962_?)del *GRCh38:* Chr16:15328439-16443962	ABCC6, ABCC1, MYH11, MARF1, NDE1, BMERB1, CEP20, MPV17L, NOMO3, MIR484, NPIPA5, MIR3179-2, MIR3180-2, MIR3670-2, NPIPA7, MIR6506, MIR6511A3, MIR6770-2, MIR6511A2, LOC112340379, LOC112340380, LOC112340381, LOC112340382 ...more		**Schizophrenia**	Pathogenic (Mar 20, 2018)	criteria provided, single submitter	VCV000545192
□ 3.	Single allele *GRCh38:* Chr16:15318125-16294378	ABCC6, ABCC1, MYH11, MARF1, NDE1, BMERB1, CEP20, MPV17L, NOMO3, MIR484, NPIPA5, MIR6506, LOC112340379, LOC112340380, LOC112340381, LOC112340382, LOC112340383, LOC113939949 ...more		**Schizophrenia**	Pathogenic (Mar 20, 2018)	criteria provided, single submitter	VCV000545191

Figure 4.3 An example of searching for variants in Clinvar

Genomic variations and references to protein sequences are represented according to the HGVS standard by indicating the chromosome and the position of the variation. For example: NM_152701.5(ABCA13):c.11473+2T>C. The HGVS nomenclature is used as an international standard for reporting and exchanging information regarding variants found in DNA, RNA, and protein sequences. [95] The database also provides the dbSNP identifier, for example rs797045205. Both the rs and HGVS identifiers are connected, but not always dbSNP ID is provided.

Clinvar integrates four essential domains of information necessary in process of selecting genomic variations. All descriptions of variants, conditions and terms for clinical significance are standardized.

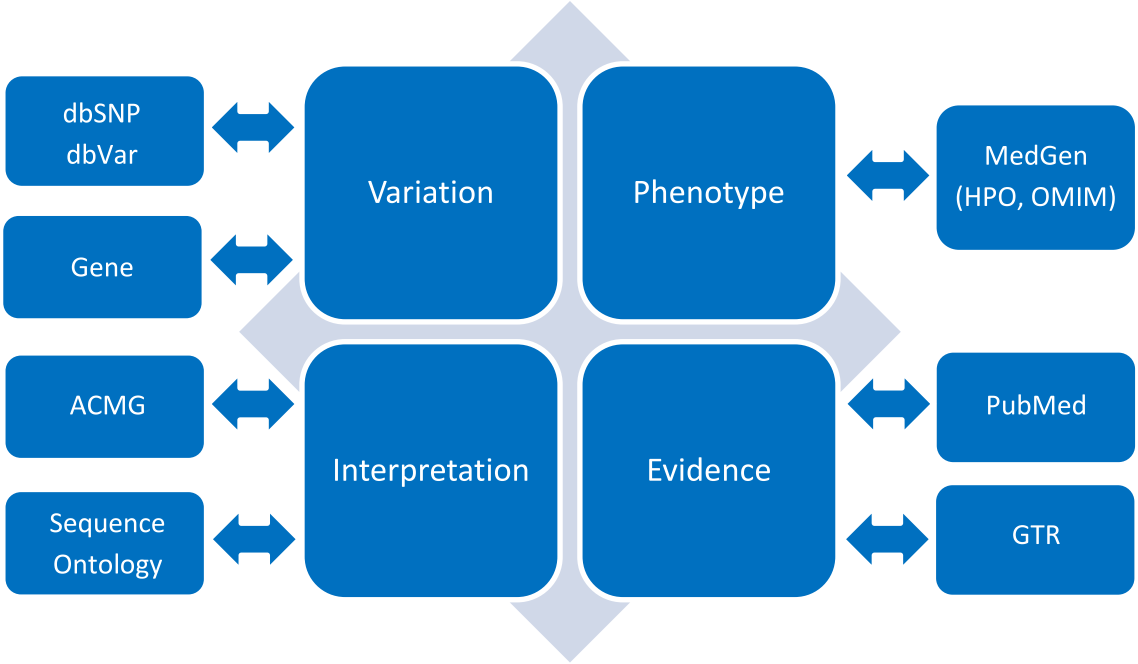


Figure 4.4 Clinvar integration of fours domains of information. [96]

Conditions are mapped to a medical genetics' portal, MedGen, what facilitates the segregation and retrieval of information about phenotypes.

Clinvar database is also integrated with many other resources, such as the Genetic Testing Registry (GTR), PubMed (providing evidence of associated literature), dbSNP, dbVar and Gene making it simple to navigate between them to find related information.

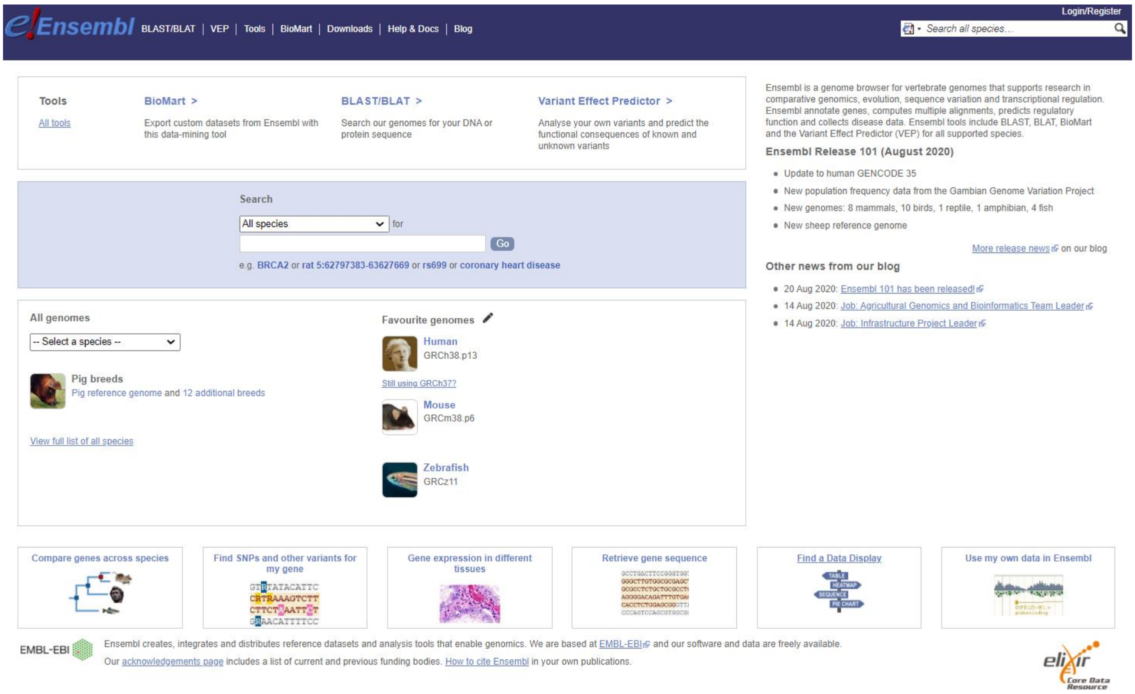
The repository is supported by statements by the American College of Medical Genetics and Genomics (ACMG), American Medical Association (AMA), and the National Society of Genetic Counselors (NSGC) ensuring quality and reliability of information. Consequently, clinical significance terms for Mendelian disorders are reported by ACMG categories.

Data can be accessed through the website or Application programming interface (API) - Entrez system and E-utilities.

ClinVar maintains submissions from over 800 organizations, from 60 countries on five continents. [97]

4.1.1 Ensembl

Ensembl genome database is a bioinformatics research project initiated in 1999 at the European Bioinformatics Institute in response to the imminent completion of the Human Genome Project. [98]. It is a comprehensive source of stable automatic annotation of individual genomes and dissemination platform for integrating and summarizing different types of experimental data against genomes of reference.



Export custom datasets from Ensembl with this data-mining tool

Search our genomes for your DNA or protein sequence

Analyse your own variants and predict the functional consequences of known and unknown variants

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

Pig reference genome and 12 additional breeds

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Ensembl Release 101 (August 2020)

- Update to human GENCODE 35
- New population frequency data from the Gambian Genome Variation Project
- New genomes: 8 mammals, 10 birds, 1 reptile, 1 amphibian, 4 fish
- New sheep reference genome

More release news on our blog

Other news from our blog

- 20 Aug 2020: Ensembl 101 has been released!
- 14 Aug 2020: Job: Agricultural Genomics and Bioinformatics Team Leader
- 14 Aug 2020: Job: Infrastructure Project Leader

Ensembl creates, integrates and distributes reference datasets and analysis tools that enable genomics. We are based at EMBL-EBI and our software and data are freely available.

Our acknowledgements page includes a list of current and previous funding bodies. How to cite Ensembl in your own publications.

Figure 4.5 Ensembl homepage [99]

Ensembl's aim is to deliver useful information to a large community of users, both from industry and academia, who use it as a framework for experimental and computational genome-based investigations. The platform supports research in comparative genomics, variation sequence and transcriptional regulation. [98]

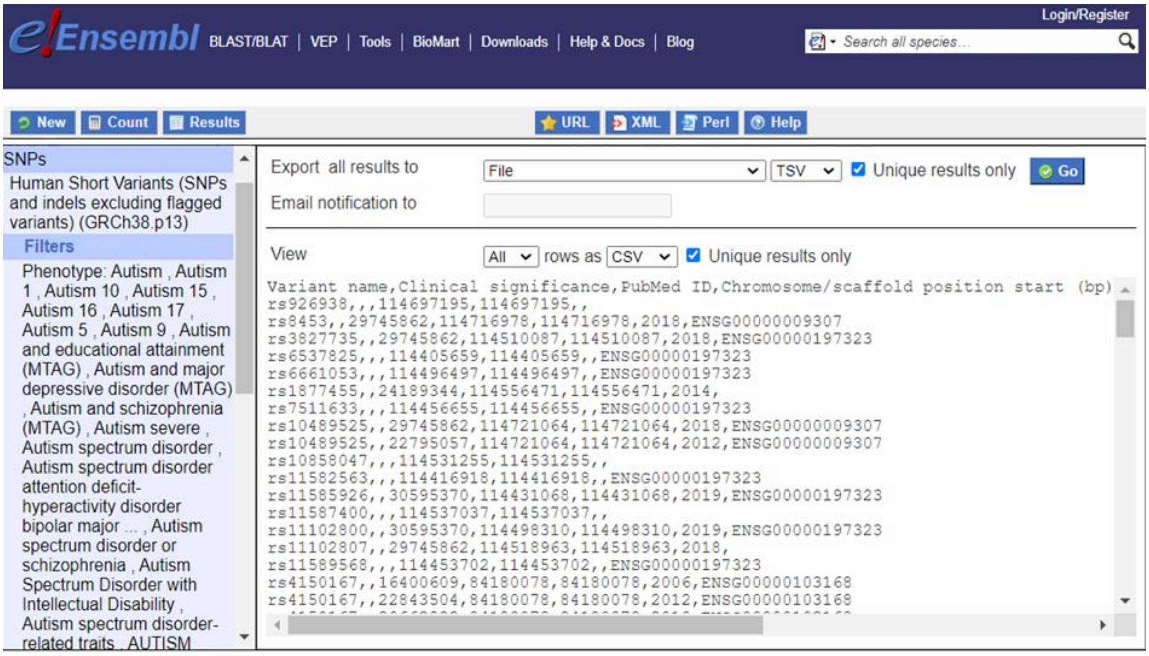
All the data and software generated in the project are freely available without restrictions to enable genomic science and foster the development of rapid research in all fields of human and animal diseases.

Ensembl is one of the three main systems displaying genome information, alongside the UCSC genome browser system and the NCBI genome resources.

Ensembl enhances work with data by providing appropriate tools for bioinformatic analysis such as BLAST/BLAT, BioMart and VEP (Variant Effect Predictor) [100] and uses a MySQL-based relational database system to efficiently store data. Data can be also accessed through the REST server using API.

The BioMart tool offers a wide range of customizable searches for individual needs and is a tool that is certainly missing from Clinvar.

The data-mining tool allows to specify attributes such as Global minor allele frequency (all individuals) or Associated gene with phenotype using specified filters. The filters are divided into appropriate sections: Variant information, Variant synonyms, Phenotype annotation, Variant Set, Variant Citations, Gene attribute, Ensemble Regulatory Features overlapping variants, Ensemble Motif features overlapping variants.



Human Short Variants (SNPs and indels excluding flagged variants) (GRCh38.p13)

Phenotype: Autism , Autism 1 , Autism 10 , Autism 15 , Autism 16 , Autism 17 , Autism 5 , Autism 9 , Autism and educational attainment (MTAG) , Autism and major depressive disorder (MTAG) , Autism and schizophrenia (MTAG) , Autism severe , Autism spectrum disorder , Autism spectrum disorder attention deficit-hyperactivity disorder bipolar major ... , Autism spectrum disorder or schizophrenia , Autism Spectrum Disorder with Intellectual Disability , Autism spectrum disorder-related traits , AUTISM

Export all results to: File, TSV, ☑ Unique results only

Email notification to:

View: All rows as CSV, ☑ Unique results only

```
Variant name,Clinical significance,PubMed ID,Chromosome/scaffold position start (bp)
rs926938,,,114697195,114697195,,
rs8453,,29745862,114716978,114716978,2018,ENSG00000009307
rs3827735,,29745862,114510087,114510087,2018,ENSG00000197323
rs6537825,,,114405659,114405659,,ENSG00000197323
rs6661053,,,114496497,114496497,,ENSG00000197323
rs1877455,,24189344,114556471,114556471,2014,
rs7511633,,,114456655,114456655,,ENSG00000197323
rs10489525,,29745862,114721064,114721064,2018,ENSG00000009307
rs10489525,,22795057,114721064,114721064,2012,ENSG00000009307
rs10858047,,,114531255,114531255,,
rs11582563,,,114416918,114416918,,ENSG00000197323
rs11585926,,30595370,114431068,114431068,2019,ENSG00000197323
rs11587400,,,114537037,114537037,,
rs11102800,,30595370,114498310,114498310,2019,ENSG00000197323
rs11102807,,29745862,114518963,114518963,2018,
rs11589568,,,114453702,114453702,,ENSG00000197323
rs4150167,,16400609,84180078,84180078,2006,ENSG00000103168
rs4150167,,22843504,84180078,84180078,2012,ENSG00000103168
```

Figure 4.6 the BioMart tool in the process of the variant search

4.1.2 OMIM

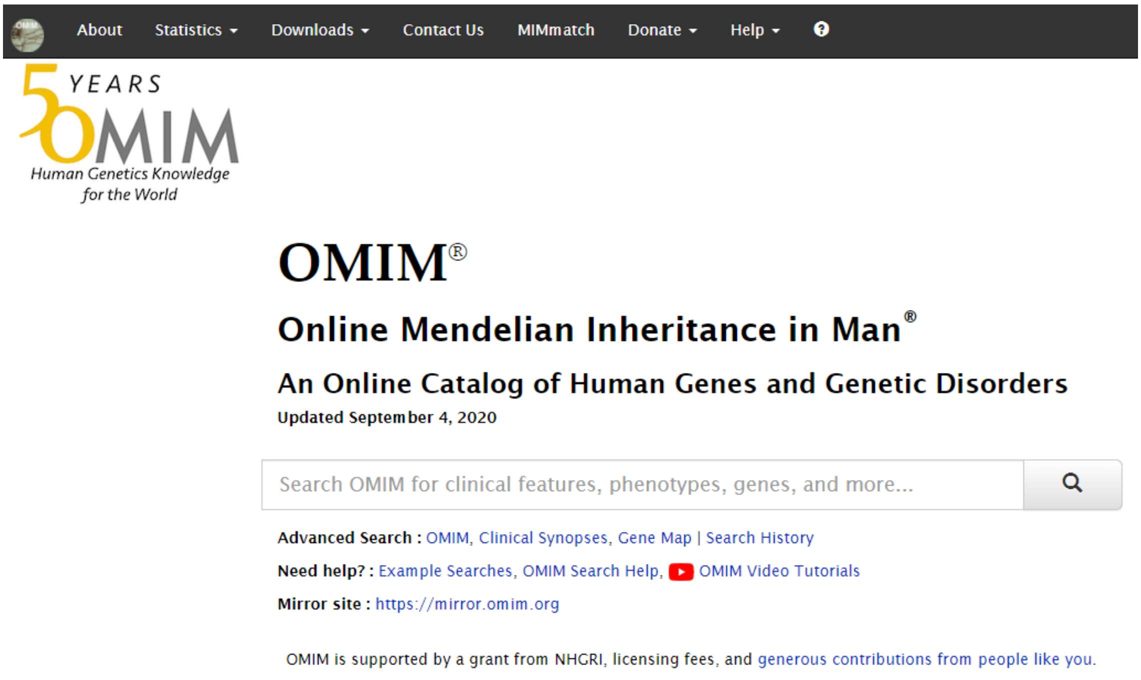
Online Mendelian Inheritance in Man (OMIM) is a comprehensive, continuously updated database of human genes and genetic disorders focusing on the relationship between phenotype and genotype. OMIM is an online continuation of Mendelian Inheritance in Man (MIM) which was initiative curated in the early 1960s at Johns Hopkins University (Baltimore-USA), first by Victor A. McKusick, known as the father of modern genetic medicine. [101]

"Mendelian inheritance" refers to the inheritance patterns controlled by a single gene with two alleles where one can be completely dominant in relation to the other. The name originates from Gregor Mendel, an Austrian monk who defined the basic principles of inheritance. [101]

In 1987 OMIM became widely available on the internet and since 1995 it has been distributed on the World Wide Web by the National Center for Biotechnology Information (NCBI) creating one of the most reliable sources of genes information. Currently, between 60,000 and 90,000 users: clinicians, researchers and others around the world, access OMIM every month. [102]

According to the latest update of statistics OMIM is a collection of more than 25,500 entries focusing mainly on gene description including 6,723 phenotypes for which the molecular basis is known and 4,318 genes with phenotype-causing mutation. [103, 104] In addition, OMIM has numerous links to external resources providing information about DNA and protein sequences, general/generic and locus-specific mutations, HUGO nomenclature, clinical trial databases, variant databases, PubMed references and more.

The names of disorders are often known by many synonyms, but OMIM makes it easier to search for phenotypes by automatically mapping the search terms and covering various ways of referring to the same anomaly. This is particularly helpful when searching for disorders such as Autism Spectrum Disorder, which contains many disease references. Each phenotype is characterized by individual *Phenotype MIM number*.



OMIM®

Online Mendelian Inheritance in Man®

An Online Catalog of Human Genes and Genetic Disorders

Updated September 4, 2020

Search OMIM for clinical features, phenotypes, genes, and more...

Advanced Search : OMIM, Clinical Synopses, Gene Map | Search History

Need help? : Example Searches, OMIM Search Help, OMIM Video Tutorials

Mirror site : https://mirror.omim.org

OMIM is supported by a grant from NHGRI, licensing fees, and generous contributions from people like you.

Figure 4.7 OMIM homepage [105]

4.2 Identification

The information found in the previous stage must be carefully processed in order to select variants that are sufficiently relevant for potential clinical application. For this purpose, the specific workflow has to be used.

The scheme is an integral part of the identification of variants and is in a continuous improvement process to increase the efficiency and accuracy of information quality.

The workflow used in my thesis is an extended version of a base schema and was previously used in the publication *Identification of Relevant Variants in Genome Information Systems: The Early Onset Alzheimer Disease Case* created by Mireia Costa Sánchez, Ana León and Óscar Pastor.

The chart created on the basis of the mentioned study is presented below:

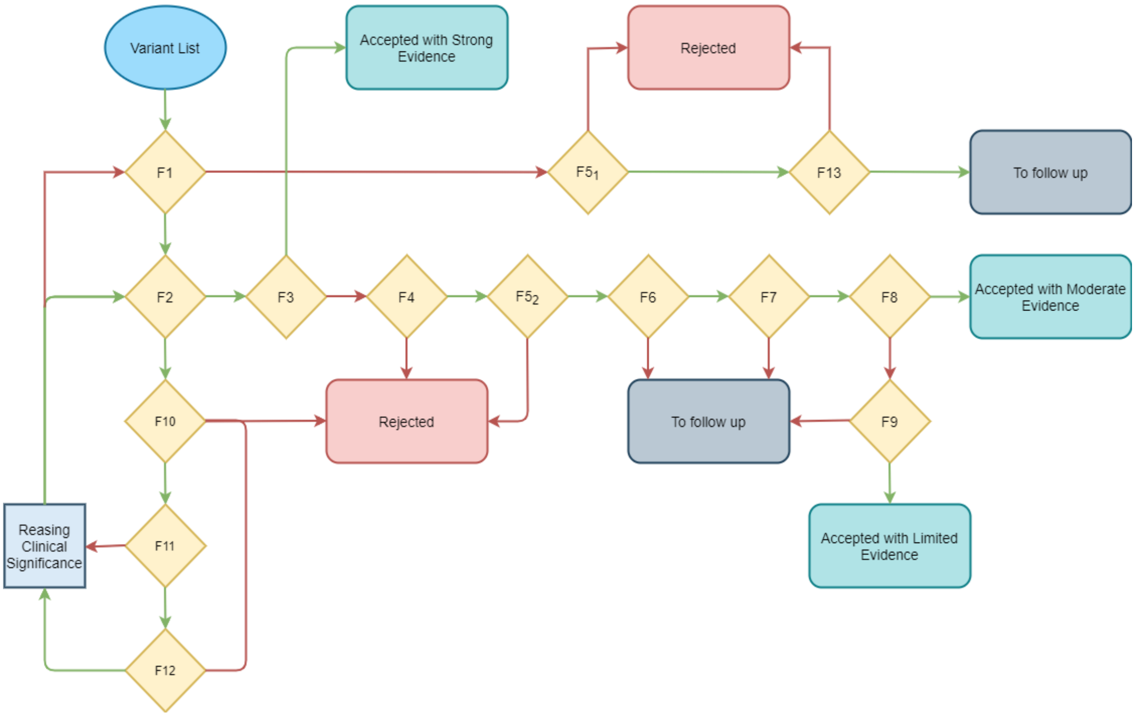

Variant List
Accepted with Strong Evidence
Rejected
F1
$F5_1$
F13
To follow up
F2
F3
F4
$F5_2$
F6
F7
F8
Accepted with Moderate Evidence
F10
Rejected
To follow up
F9
Reasing Clinical Significance
F11
Accepted with Limited Evidence
F12

Figure 4.8 Schema of SILE workflow

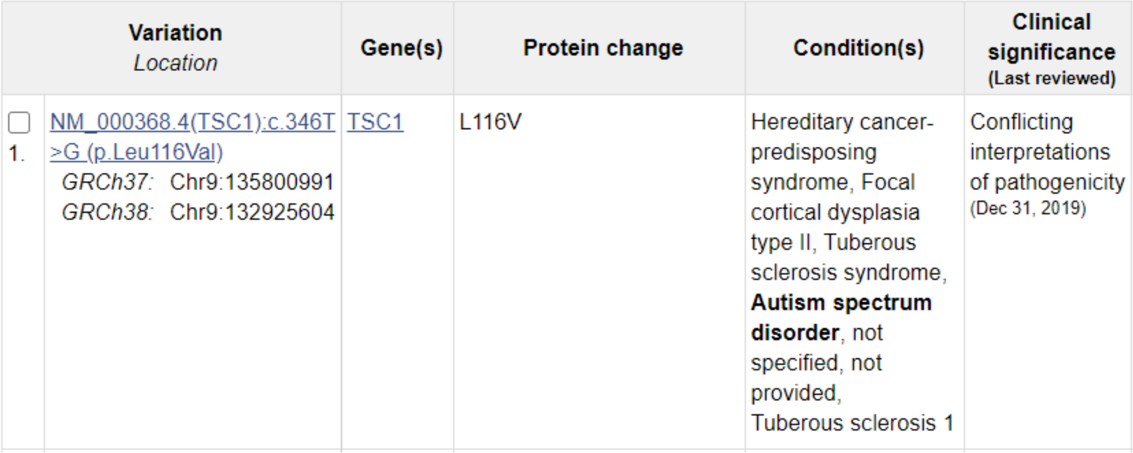
The results found related to a specific phenotype are not completely verifiable only by automatic way, so the individual steps for manual verification of variants were defined to prevent errors.

4.2.1 Conflicting Interpretations

One of the characteristics of the variant that may affect misclassification is "Interpretation". If a variant is associated with more than one phenotype, each variant in which there is a conflict of interpretations must be checked individually. Such information may result from different interpretations assigned to different diseases for the same variant. Considering this inaccuracy, the workflow contains the F10-F12 steps. To correctly assign the interpretation to the phenotype that interests us, each case where conflict of interpretations appears has to be considered and this process is contained in step F10.

For the following, an example of a variant rs199620268 associated with Autism spectrum disorder, in which conflicting interpretation of pathogenicity occurs, is presented.

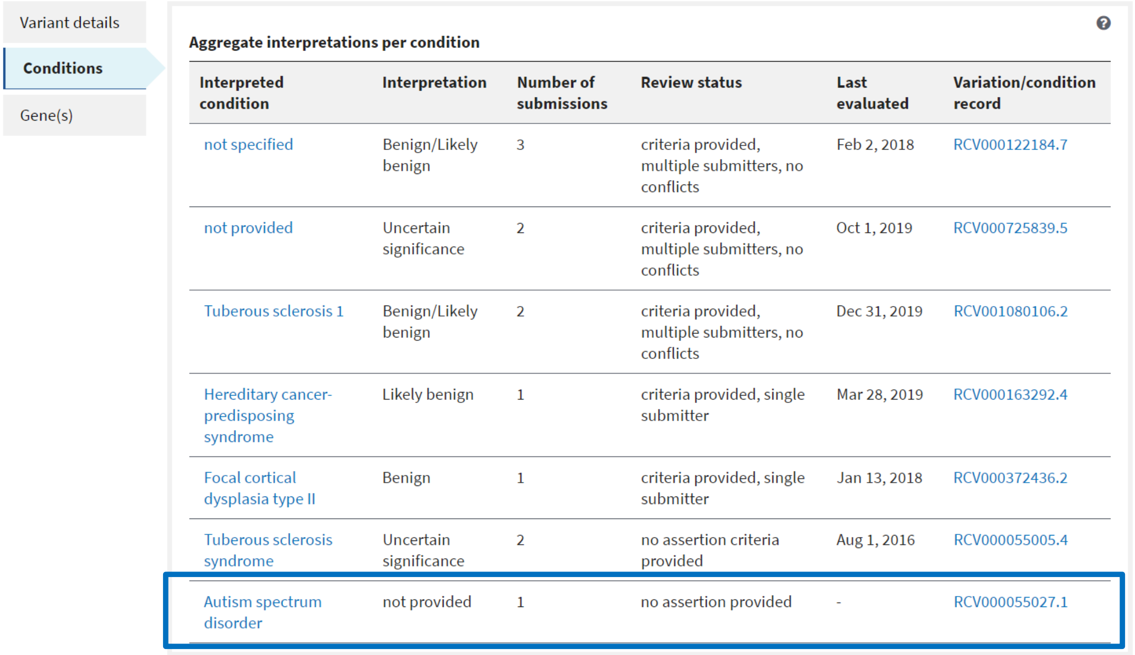
Table 4.3 Fragment of the ClinVar database search results table containing the variant rs199620268 with pathogenicity interpretation conflict



	Variation *Location*	Gene(s)	Protein change	Condition(s)	Clinical significance (Last reviewed)
□ 1.	NM_000368.4(TSC1):c.346T>G (p.Leu116Val) *GRCh37:* Chr9:135800991 *GRCh38:* Chr9:132925604	TSC1	L116V	Hereditary cancer-predisposing syndrome, Focal cortical dysplasia type II, Tuberous sclerosis syndrome, **Autism spectrum disorder**, not specified, not provided, Tuberous sclerosis 1	Conflicting interpretations of pathogenicity (Dec 31, 2019)

After manual check of the variant, it can be determined that the interpretation for the selected phenotype is not provided.

Table 4.4 Table of interpretations per condition of the variant rs199620268



Aggregate interpretations per condition

Interpreted condition	Interpretation	Number of submissions	Review status	Last evaluated	Variation/condition record
not specified	Benign/Likely benign	3	criteria provided, multiple submitters, no conflicts	Feb 2, 2018	RCV000122184.7
not provided	Uncertain significance	2	criteria provided, multiple submitters, no conflicts	Oct 1, 2019	RCV000725839.5
Tuberous sclerosis 1	Benign/Likely benign	2	criteria provided, multiple submitters, no conflicts	Dec 31, 2019	RCV001080106.2
Hereditary cancer-predisposing syndrome	Likely benign	1	criteria provided, single submitter	Mar 28, 2019	RCV000163292.4
Focal cortical dysplasia type II	Benign	1	criteria provided, single submitter	Jan 13, 2018	RCV000372436.2
Tuberous sclerosis syndrome	Uncertain significance	2	no assertion criteria provided	Aug 1, 2016	RCV000055005.4
Autism spectrum disorder	not provided	1	no assertion provided	-	RCV000055027.1

Once the interpretation of each variant is determined, variants are separated into those whose interpretation is clinically actionable.

4.2.1.1 Clinical significance

In order to understand the classification of variants under clinical significance, it is necessary to be familiar with the value options recognized in the ClinVar database. The table below explains the basic values and their application guide in ClinVar. [106]

Table 4.5 The list of terms used for clinical significance by ClinVar

Clinical significance value	Guidance for use in ClinVar SCV records
Benign	As recommended by ACMG/AMP for variants interpreted for Mendelian disorders.
Likely benign	As recommended by ACMG/AMP for variants interpreted for Mendelian disorders.
Uncertain significance	As recommended by ACMG/AMP for variants interpreted for Mendelian disorders.
Likely pathogenic	As recommended by ACMG/AMP for variants interpreted for Mendelian disorders.
Pathogenic	As recommended by ACMG/AMP for variants interpreted for Mendelian disorders.
	Variants that are pathogenic with low penetrance may be submitted as "Pathogenic"; please also include information about the penetrance in a "Comment on clinical significance".
drug response	A general term for a variant that affects a drug response, not a disease. We anticipate adding more specific drug response terms based on a recommendation by CPIC.
association	For variants identified in a GWAS study and further interpreted for their clinical significance.
risk factor	For variants that are interpreted not to cause a disorder but to increase the risk.
protective	For variants that decrease the risk of a disorder, including infections.
Affects	For variants that cause a non-disease phenotype, such as lactose intolerance.
conflicting data from submitters	Only for submissions from a consortium, where groups within the consortium have conflicting interpretations of a variant but provide a single submission to ClinVar.
other	If ClinVar does not have the appropriate term for your submission, we ask that you submit "other" as clinical significance and contact us to discuss if there are other terms we should add.

not provided	For submissions without an interpretation of clinical significance. The primary goal of ClinVar is to archive reports of clinical significance of variants. Therefore submissions with a clinical significance of "not provided" should be limited to: "literature only" submissions that report a publication about the variant, without interpreting the clinical significance "research" submissions that provide functional significance (e.g. undetectable protein level) but no interpretation of clinical significance "phenotyping only" submissions from clinics or physicians that provide additional information about individuals with the variant, such as observed phenotypes, but do not interpret the clinical significance
'-'	This value may not be submitted. It is used in the file variant_summary.txt.gz in the path https://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/. This file reports '-' in the ClinicalSignificance column for an allele that was submitted to ClinVar only in combination with another allele (e.g.a submission with an interpretation of a haplotype or a compound heterozygote) and was not interpreted explictly. ClinVar thus has no interpretation specific to that allele. To find the the interpretation that includes that allele, you can query ClinVar by the AlleleID,e.g. https://www.ncbi.nlm.nih.gov/clinvar/?term=38420[alleleid].

The American College of Medical Genomics has recommended a five-tier classification system and it is used by a majority of research laboratories. According to this system, a variant can be classified as [32]:

1. Pathogenic – the variant contributes directly to the development of disease
2. Likely pathogenic – there is a high likelihood, greater than 90% certainly, that the variant contributes to disease
3. Uncertain Significance – there is not enough information to classify the variant with certainty
4. Likely benign – the variant is not expected to contribute with the disease, however there is not sufficient evidence to prove it
5. Benign – the variant does not contribute with the disease

In accordance with the ACMG/AMP recommendation used in the workflow, variants are considered relevant if the interpretation is Pathogenic, Likely pathogenic. In addition, those with “association” and “risk factor” characteristics are also taken into account.

The rest of the variants are classified as Rejected and are accompanied by other variants that have a clear interpretation from the beginning and belong to a set classified as clinically irrelevant – Uncertain Significance, Likely benign, Benign.

If 'no provided', the variant is considered further in step F5_1 as specified in the section “Variants without Interpretation provided”.

Variations classified for the next step are verified in F11, where it is determined whether the conflict appears in a particular phenotype and whether there is a conflict between the information from the submitters if there are more than one. Variants with a clearly defined feature return to the process, and those with conflict submissions are analyzed in step F12. It is assumed that there has to be an agreement between at least 75% of the submitters, in which case the interpretation of the variant is determined by the majority.

4.2.2 Variants without Interpretation provided

Variants with no interpretation provided may still be sufficiently linked to the disease. It should be checked whether the gene where the specific variant is located is related to the phenotype ($F5_1$). Then the variant-disease relationship can be estimated by analyzing the statistical significance of the associated studies. In the F13 the following steps should be done:

- to check the existence of studies confirming the relationship between the variant and the disease,
- to determine the Minor Allele Frequency (MAF),
- to evaluate compatibility with the disease.

If a variant passes the verifications, it is classified as *To follow up*. There is a likelihood of a variant-disease association, although no clinical significance is presented.

4.2.3 The Variant Review Status

One of the important data quality indicators is the relevance of submission. The Variant Review Status provides information about the overall status based on an analysis of all submissions and is represented graphically by 0-4 scale stars as shown in the figure 4.3. [107]

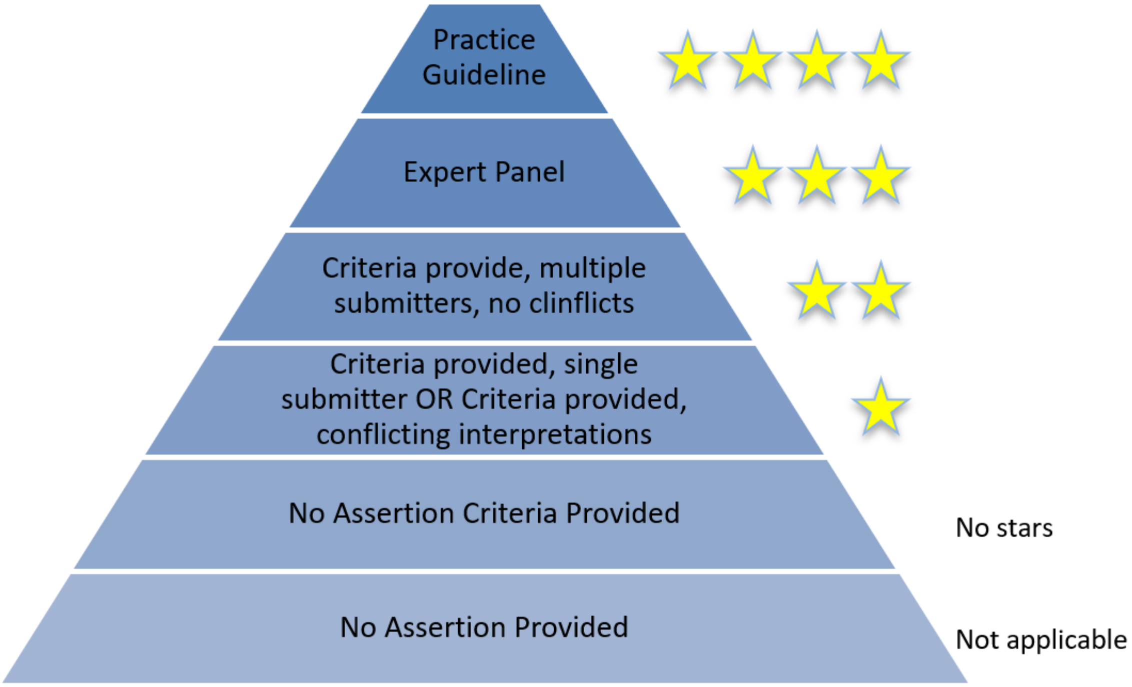

Practice Guideline
★★★★
Expert Panel
★★★
Criteria provide, multiple submitters, no clinflicts
★★
Criteria provided, single submitter OR Criteria provided, conflicting interpretations
★
No Assertion Criteria Provided
No stars
No Assertion Provided
Not applicable

Figure 4.9 The value range of the Variant Review Status

An interpretation with Practice Guidelines or Expert Panel Review Status guarantee the best quality and veracity of data. Variants whose submitters belong to one of the following characteristics are classified as *Accepted with Strong Evidence* and this condition is evaluated in the step F3 of the workflow. [107]

Variants with submissions characterized by one or two stars have to be verified individually. In the step F7, the quality criteria identify the source as relevant and reliable if more than 200 submissions have been carried out by the certain submitter.

4.2.4 Interpretation of the collection method

The collection method describes the origin of the submitted information and is used to interpret the variant clinical significance. Data can be collected as part of clinical testing, research, or literature review. [108] Description of each is provided below.

Table 4.6 Clinvar main collection method description

Collection method	Definition
Clinical testing	Provides a standardized classification and includes variants interpreted through clinical genetic testing, large-volume tests with compliance of results with CLIA, ISO, GLP standards and are routinely re-tested.
Research	Ensures that variants are verified in research projects and may be standardized but are not routinely tested and do not meet the requirements for clinical testing.
Literature only	Provides information about variants extracted from the published literature and does not provide additional testing and verification of the variant's phenotype consistency.

My study focuses on the possible practical use of information in medicine; therefore, the greatest guarantee of the variant-disease relationship is the data originating from clinical testing. The method is validated in step F4 and if the criterion is not met then the variant is considered as *Rejected*.

4.2.5 Genes Relevance

Variants not interpreted by practical guidelines or expert panels have to be verified in regard to the variant-gene-phenotype relationship to determine whether the variant is located in a gene previously associated with the disease.

As it was mentioned before, the appropriate repositories providing data about genes verified within the phenotype has to be reviewed. In my study, the F5 step evaluation is based on Online Mendelian Inheritance in Man.

4.2.6 Assertion Criteria

The initial step in evaluating the assertion criteria is to check whether any of the criteria are provided (F6). Variants with *no assertion criteria provided* are classified as *To follow up*.

Mentioned previously in the chapter *4.2.1.1 Clinical significance*, the American College of Medical Genetics (ACMG) and the Association for Molecular Pathology (AMP), developed guidelines for the assessment of evidence and their application ensures greater consistency and transparency in clinical variant interpretation. Therefore, the next necessary step (F8) in workflow is to verify if the interpretation of the data is specified in accordance with these guidelines. Compliance with ACMG/AMP ensures that the interpretation is made on an internationally accepted criteria which increases the significance of the variant.

4.2.7 The Last Date of Review

The absence of evidence supporting the use of ACMG/AMP criteria does not necessarily mean that the other criteria used by the submitter are not valid, so workflow validates such variants as *Accepted with limited evidence* if the last date of review took place within less than 3 years from the present year (F3). [37]

A summary instruction of the particular workflow steps is given below.

Table 4.7 Instructional table of the workflow.

No.	**Verification instruction**	**Next step to follow if**	
		Yes (green)	**No (red)**
F1	Check if the clinical significance is provided	F2	F4$_1$
F2	Check if there is conflict interpretation of interpretation	F3	F10
F3	Check if the submitter is a Practice Guideline or an Expert Panel	Accepted with Strong Evidence	F4
F4	Check if the method used is clinical testing	F5	Rejected
F5$_1$	Check if the gene where the variant is located is related to the studied phenotype	F13	Rejected
F5$_2$		F6	Rejected
F6	Check if the assertion criteria are provided	F7	To follow up
F7	Check the relevance of the submitter according to the number of submissions	F8	To follow up
F8	Check if the ACMG/AMP guidelines have been used	Accepted with Moderate Evidence	F9
F9	Check the last date of review	Accepted with Limited Evidence	To follow up

F10	Check if the variant is clinically actionable	F11	Rejected
F11	Check if the conflict occurs in the studied phenotype	F1 or F2	F12
F12	Check if a major agreement between submitters confirm clinical importance	F1 or F2	Rejected
F13	Verify the relevance of the associated studies	To follow up	Rejected

In accordance with guidelines described above, the final designation of the relevant variants is as follows:

• Accepted with strong evidence: includes variants with provided Practice Guideline or Expert Panel reference.

• Accepted with moderate evidence: contains variants which, despite the lack of references from the Practice Guideline or Expert Panel, are considered relevant due to the moderate quality level of the rest of the evidence.

• Accepted with limited evidence: contains variants which, despite the lack of a Practice Guideline or Expert Panel reference, are considered relevant due to the limited quality level of the rest of the evidence.

CHAPTER 5. APPLICATION OF THE SILE METHODOLOGY

5.1 Workflow application to ASD

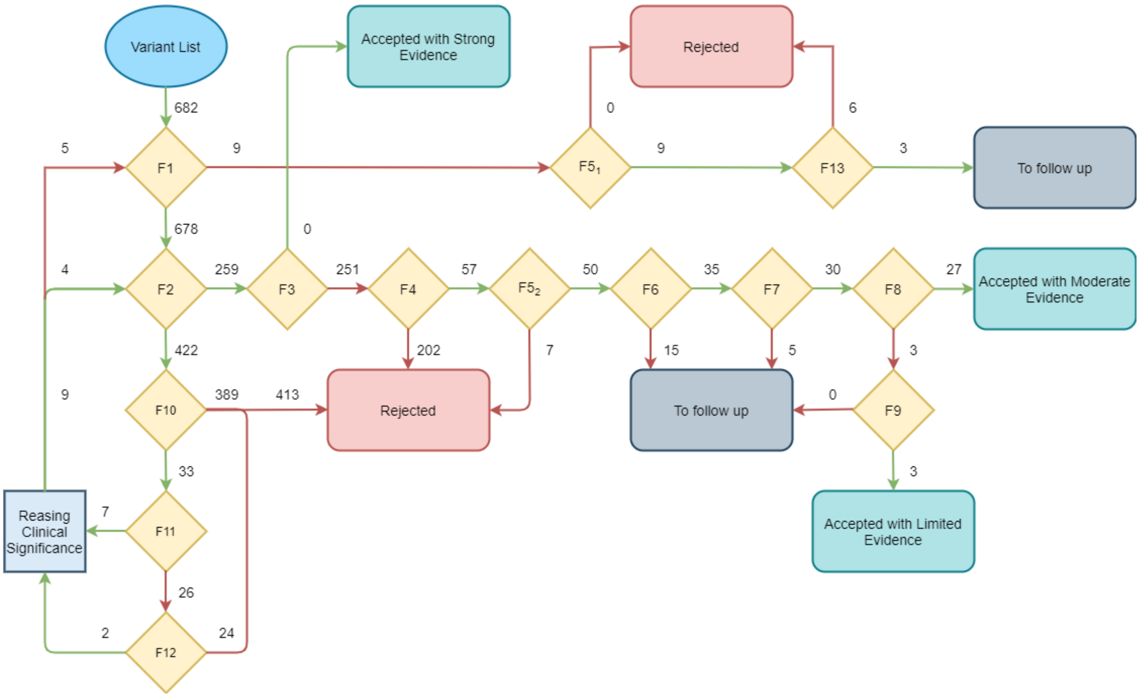


Figure 5.1 Results of the improved workflow application (ASD)

The whole diagram shows the exact number of variants related to ASD that went through the steps of the workflow validation.

- Search

  Before starting the process of selecting variants, it is necessary to create their list by searching data in ClinVar. An essential concern in case of ASD is that the spectrum includes several diseases, so each of them has to be considered in the query when searching for variants.

  Through Autism[Disease/Phenotype] command it was possible to search all the results associated with phenotypes: atypical autism, childhood autism, infantile autism, Kanner's autism, high-functioning autism. The result of the query was 680 variants.

  The search related to Asperger's disorder returned 5 variants but 4 of them have already been found in a previous query.

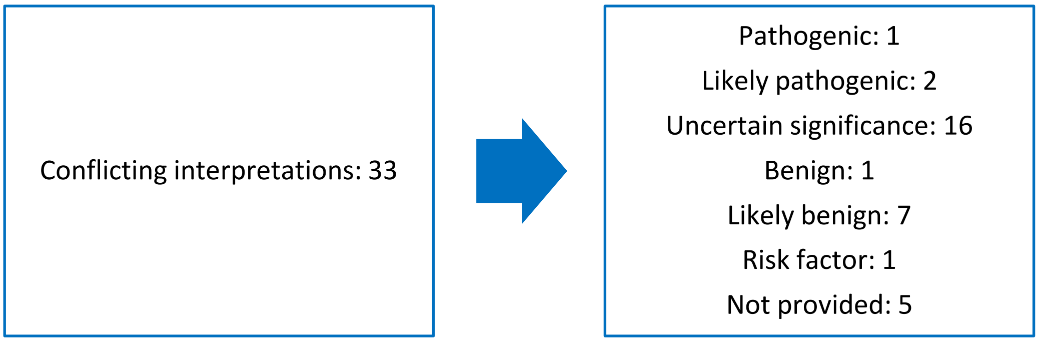
  The search related to childhood disintegrative disorders (CDD) has not returned any results.

  The search related to pervasive developmental disorder not otherwise specified (PDD-NOS) returned 1 variant.

The final list included 682 variants.

- Steps F10-F12

  First, an analysis of steps F10-F12 was performed. After verification of 33 variants showing conflicting interpretations of pathogenicity, the characteristics were updated.



Conflicting interpretations: 33

Pathogenic: 1
Likely pathogenic: 2
Uncertain significance: 16
Benign: 1
Likely benign: 7
Risk factor: 1
Not provided: 5

Figure 5.2 Verification of variants with conflicting interpretation (ASD)

In steps F10-F12, 414 variants were rejected with a division into 28 Benign, 32 Benign/Likely benign, 94 Likely benign, 260 Uncertain Significance. 5 variants with assertion criteria not provided were returned to step F1.

A full picture of the distribution of variants depending on clinical significance was obtained. The overview is shown in the graph below.

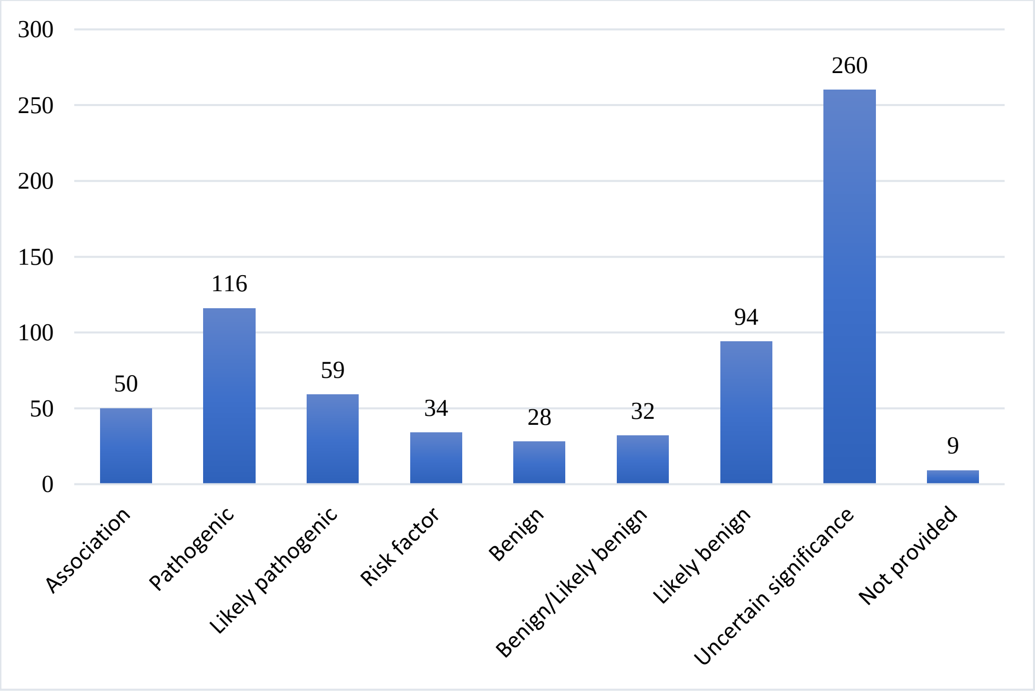


Figure 5.3 Distribution of variants by clinical significance (ASD)

- Steps F5$_1$-F13

  In this section 9 variants with clinical significance "not provided" were reviewed. All variants belong to the phenotype related genes but only 3 of them met the conditions of the step F13.

- Step F3

  Using a filter available from a search engine on the Clinvar website, 8 variants were classified as those being submitted by the Expert Panel, but after a detailed review of each variant it turned out that none of them fulfill the condition for ASD.

- Step F4

  202 variants were not evaluated using clinical testing. The effect of the step F4 on reducing the number of variants depending on the type is shown below.

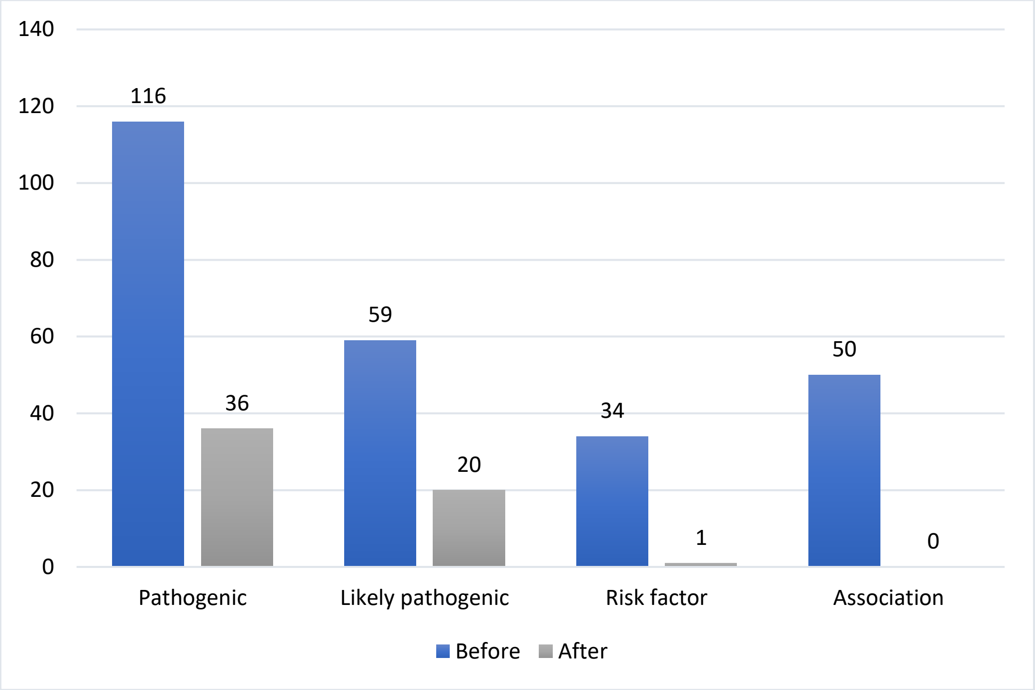


Figure 5.4 Number of variants before and after step F4 (ASD)

- Step F5

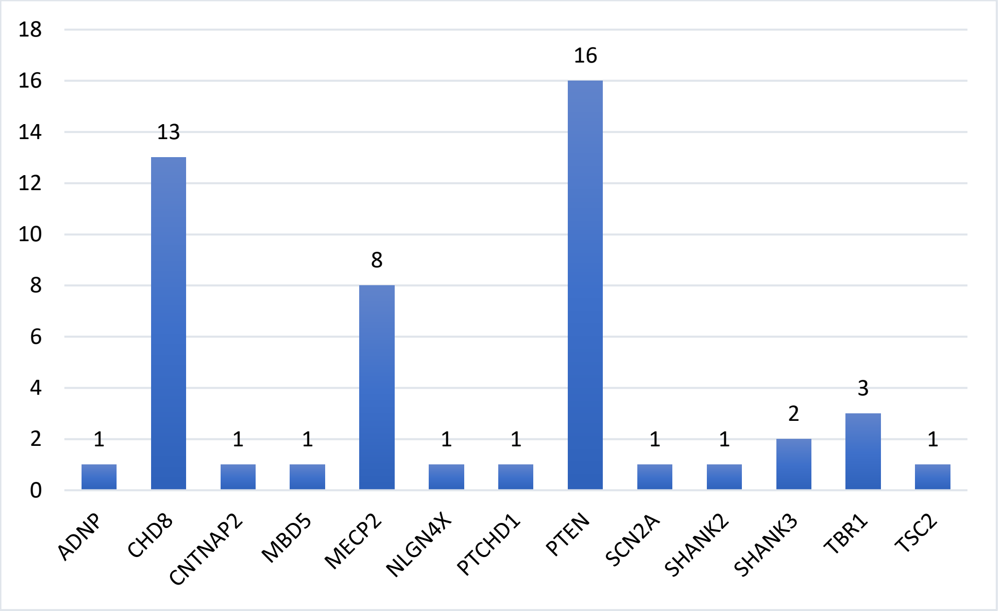


Figure 5.5 Distribution of variants by gene (ASD)

Each of the 20 genes related to the analyzed variants was verified using information from the OMIM database and 13 were selected as relevant.

As can be seen in the diagram (Figure 5.5) the three genes CHD8, MECP2 and PTEN show the greatest connection to Autism Spectrum Disorder.

- Step F6
  During the analysis of the review status information, initially 4 variants showed no assertion criteria provided, and other 4 reviewed by Expert Panel, but after manual verification of each of the variants, it turned out that for this phenotype there were no criteria provided for 15 variants.
- Step F7 and F8
  30 variants went through the filter F7 specifying the relevance of the submitter and only 3 of them failed to meet condition of step F8 describing the use of ACMG/AMP guides. The impact of each step is described in the graph below.

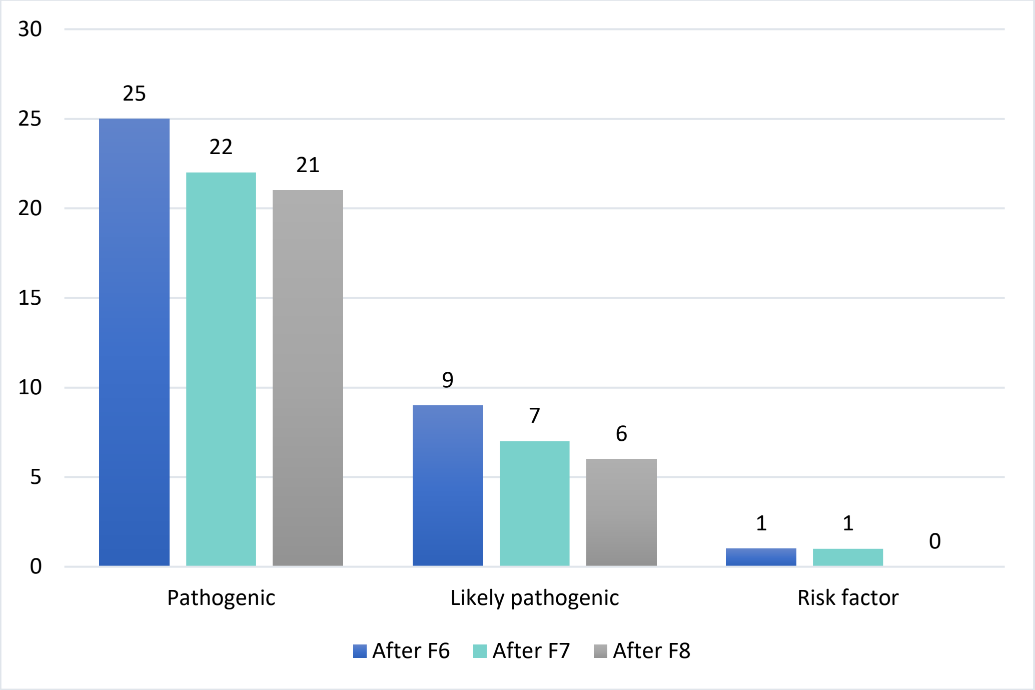


Figure 5.6 Number of variants after steps F6, F7 and F8 (ASD)

- Step F9
  The last evaluation date of all the 3 variants rejected in step F8 is no older than three years from 2020, the current year, and they have been qualified as Accepted with limited evidence.

Summarizing, most of the variants, 629 precisely, were classified as *Rejected*. 27 variants were interpreted as *Accepted with moderate evidence*, 23 *To follow up*, and 3 *Accepted with limited evidence*.

The figure 5.7 shows the results of the variant classification.

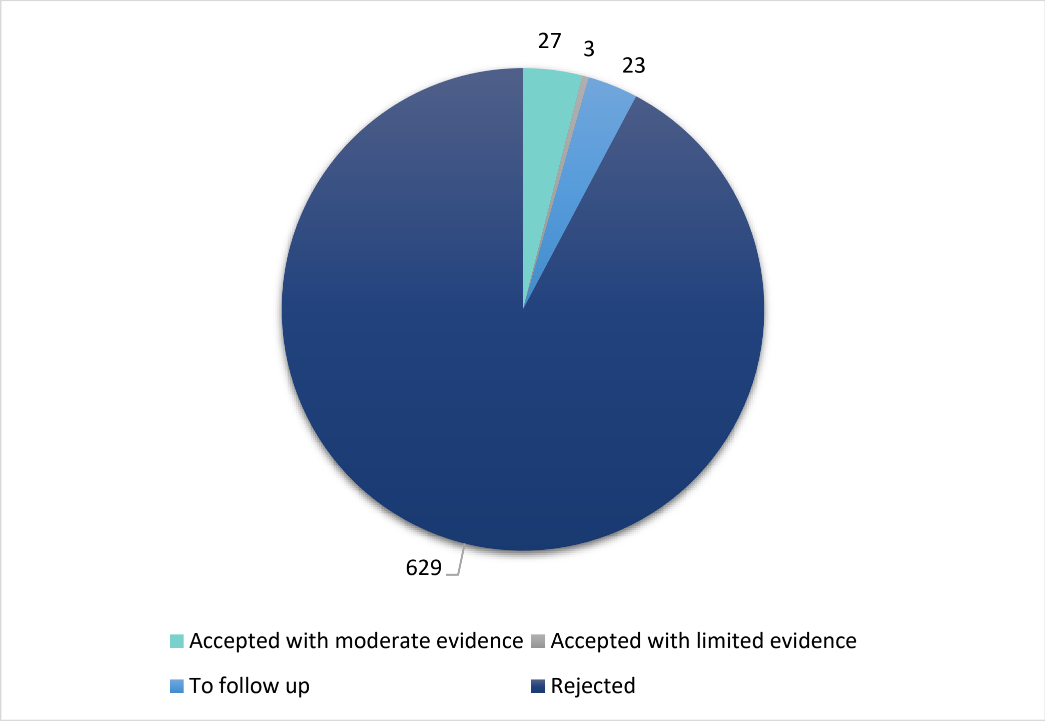


Figure 5.7 Final results of the workflow (ASD)

5.2 Workflow application to ADHD

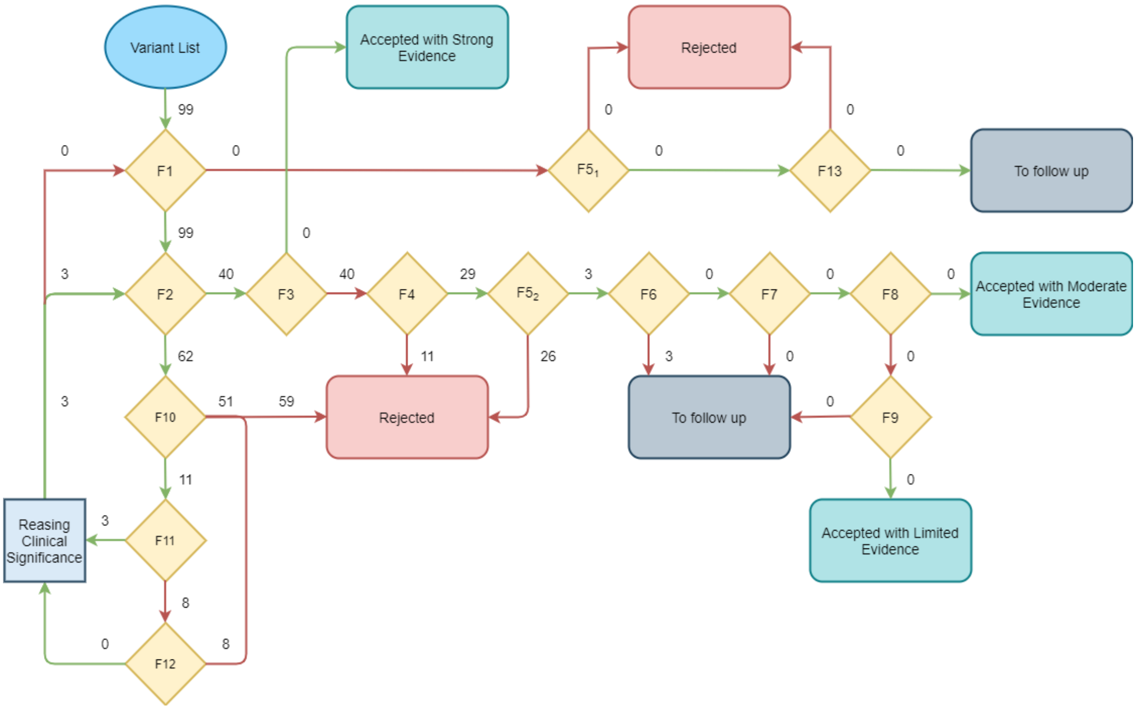
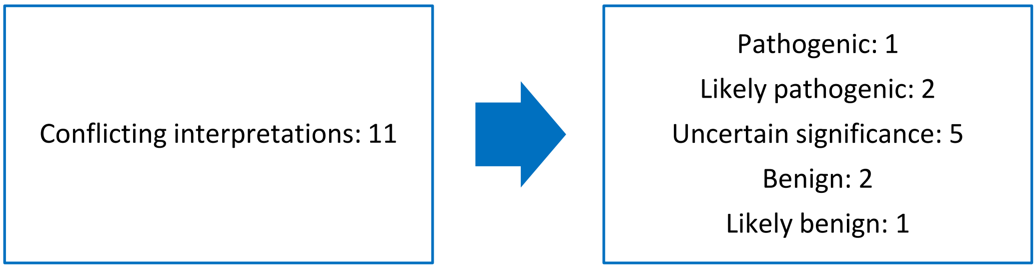


Figure 5.8 Results of the improved workflow application (ADHD)

- Search
  The list of 99 variants was created using the command (Attention[Disease/Phenotype] AND Deficit[Disease/Phenotype] AND Hyperactivity [Disease/Phenotype] AND Disorder[Disease/Phenotype]) in ClinVar.
- Steps F10-F12
  11 variants presenting conflicting interpretations were verified and updated following the figure below.


Conflicting interpretations: 11

Pathogenic: 1
Likely pathogenic: 2
Uncertain significance: 5
Benign: 2
Likely benign: 1

Figure 5.9 Verification of variants with conflicting interpretation (ADHD)

After filtering the variants in steps F10-F12, 59 were finally rejected with a division into 9 Benign, 5 Benign/Likely benign, 1 Likely benign, 44 Uncertain Significance.

The overview of the distribution of variants depending on clinical significance is shown in the graph below.

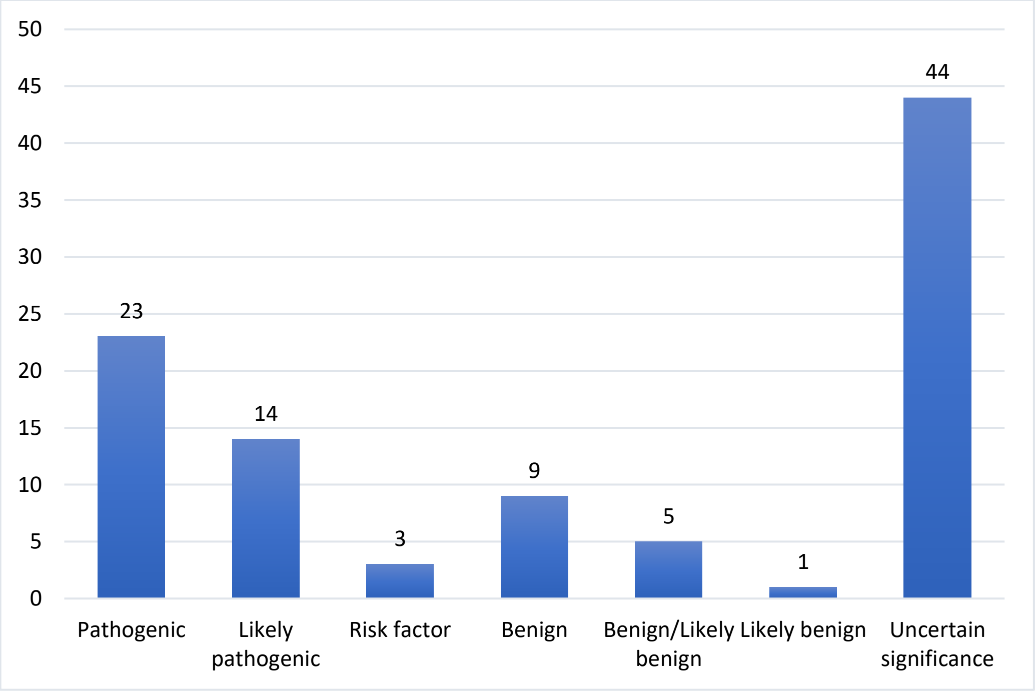


Figure 5.10 Distribution of variants by clinical significance (ADHD)

- Steps F5$_{1}$-F13
  Due to the lack of variants with the clinical significance of "not provided", no steps were taken.

- Step F3

  None of the 40 variants met the requirements to have a Practice guideline or Expert Panel as a submitter.
- Step F4

  29 variants were evaluated using clinical testing and their distribution between clinical significance is shown in the Figure 15.11, 11 other variants were rejected.

  The impact of the step F4 on the number of different types of variants is shown on the graph below.

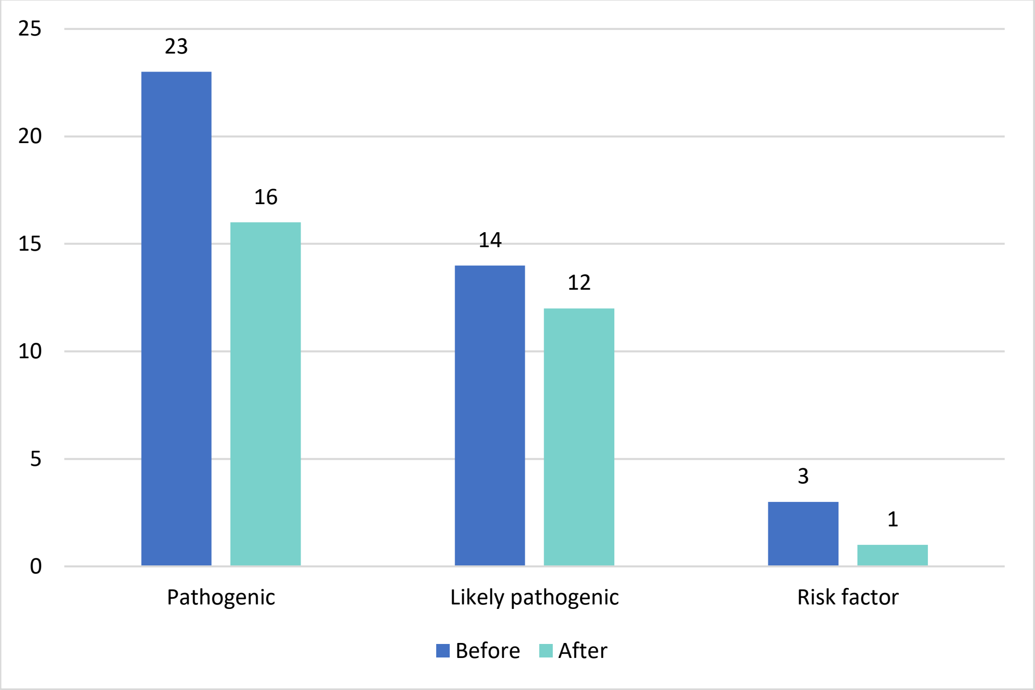


Figure 5.11 Number of variants before and after step F4 (ADHD)

- Step F5

  Unfortunately, most of the genes corresponding to the variants are not related to the discussed ADHD phenotype. According to the OMIM, out of 27 genes, only 2 have a correlation with disease and these are GNB5 associated with two variants, and TPH2 associated with one.
- Step F6

  None of the three variants had assertion criteria provided and all were classified as *To follow up*.
- Step F7, F8 and F9

  All the variants have already been identified, so steps F7, F8 and F9 were not followed.

The final result of the selection of variants refers to only two classifications. None of the variants has been accepted. 3 variants were marked as *To follow up* and all the remaining 96 *Rejected*.

The results are presented in Figure 5.12

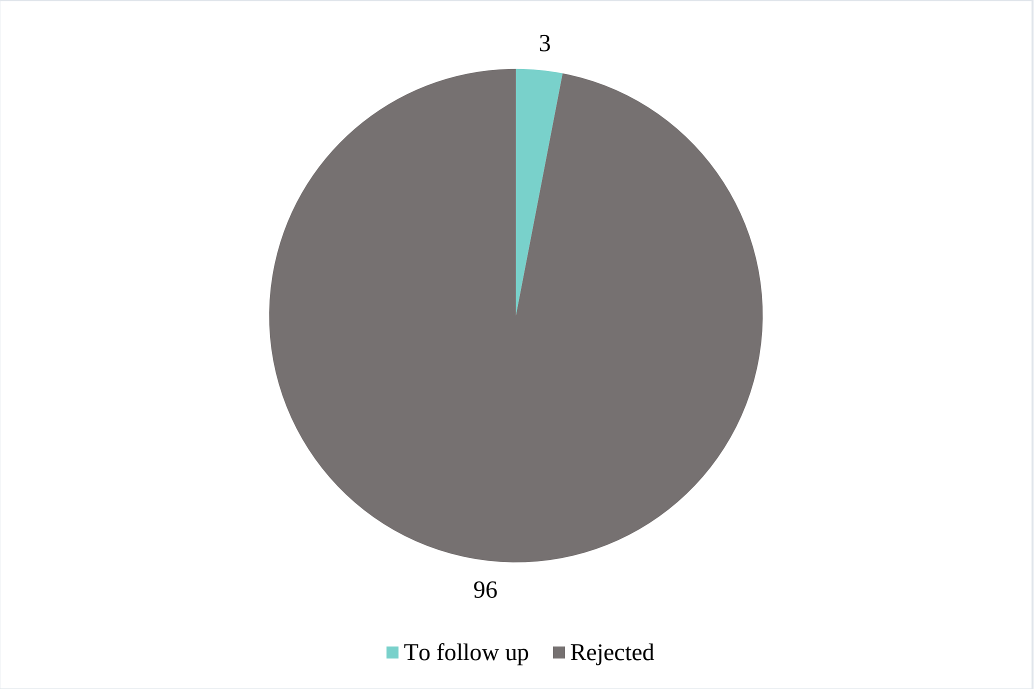


Figure 5.12 Final results of the workflow (ADHD)

5.3 Workflow application to Schizophrenia

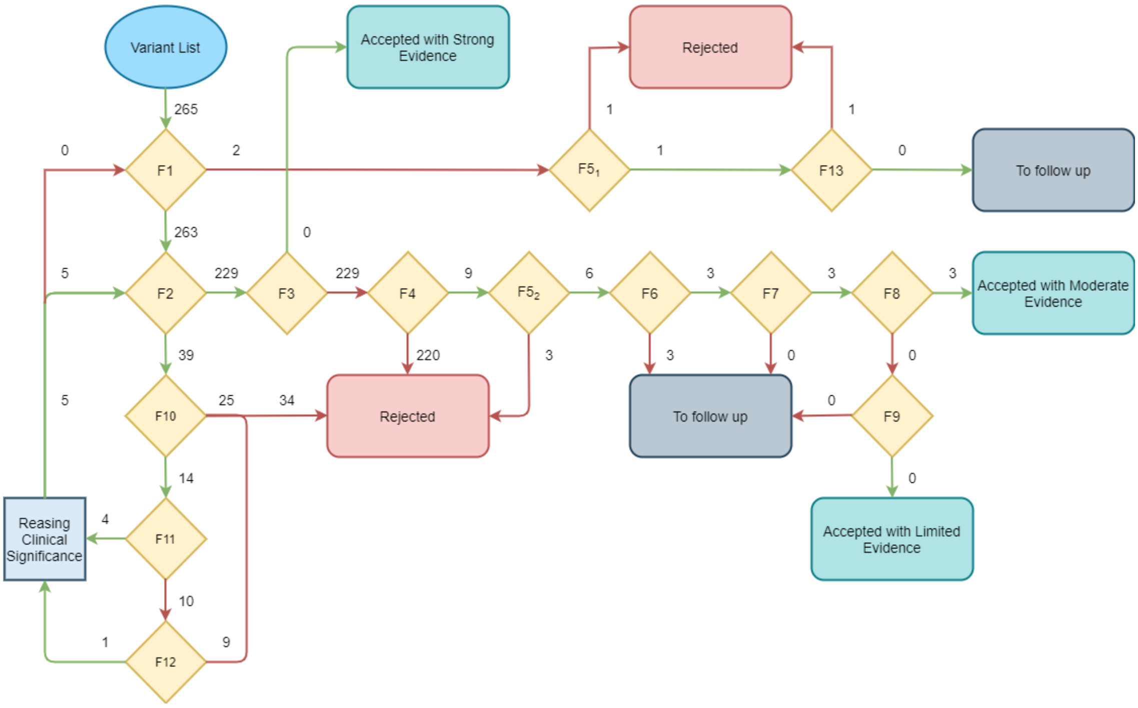


Figure 5.13 Results of the improved workflow application (Schizophrenia)

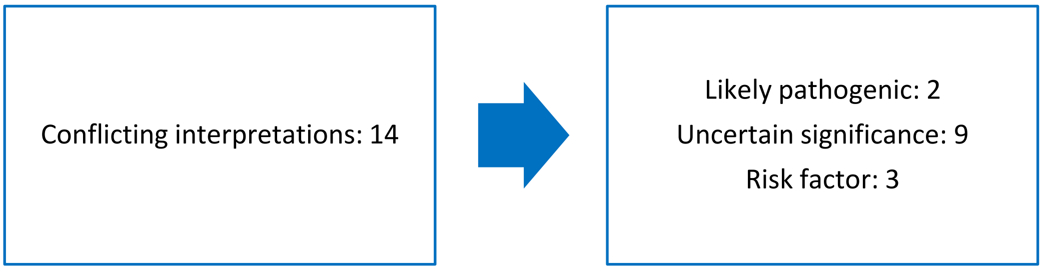
The sequence of steps for identifying the variants related to Schizophrenia is described below.

- Search

  The list of variants was created using Schizophrenia[Disease/Phenotype] query. 265 variants were found.

- Steps F10-F12

  14 variants showed conflicts of interpretation and their characteristics were verified individually.



Conflicting interpretations: 14

Likely pathogenic: 2
Uncertain significance: 9
Risk factor: 3

Figure 5.14 Verification of variants with conflicting interpretation (Schizophrenia)

In steps F10-F12, 34 variants were rejected with a division into 7 Benign, 1 Benign/Likely benign, 1 Likely benign, 25 Uncertain Significance.

A full picture of the distribution of variants depending on clinical significance was obtained. The overview is shown in the graph below.

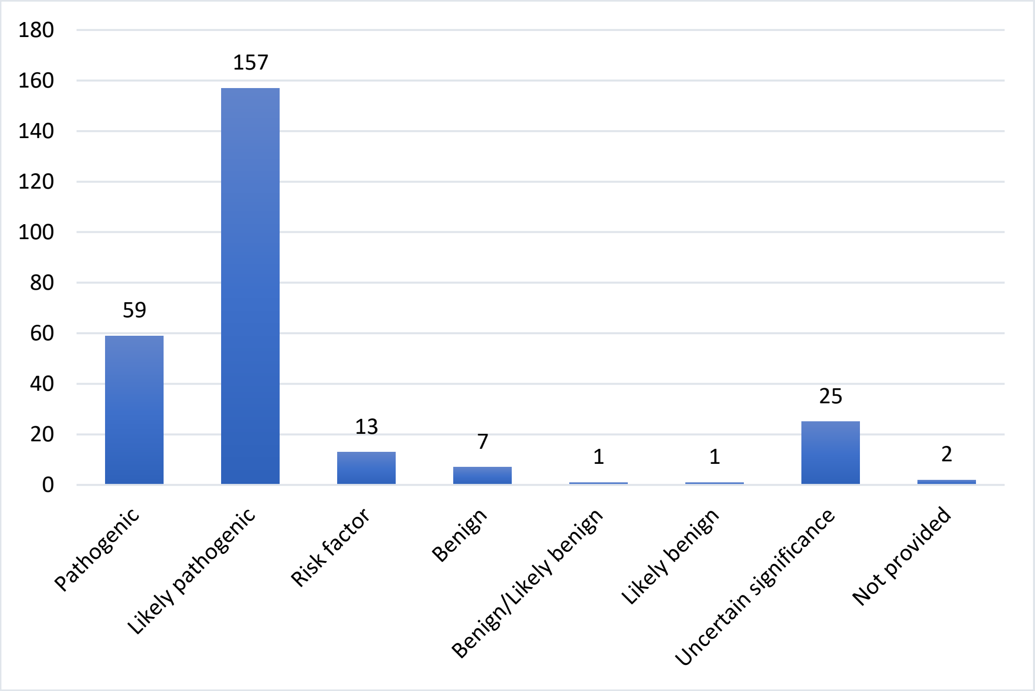


Figure 5.15 Distribution of variants by clinical significance (Schizophrenia)

- Steps F5$_1$-F13

  Both variants were rejected because of not meeting the conditions.
- Step F3

  None of the 229 variants met the criteria to have a Practice guideline or Expert Panel as a submitter.
- Step F4

  A total of 220 variants were not evaluated using clinical testing and the extreme selection of variants is shown in the graph Figure 5.16.

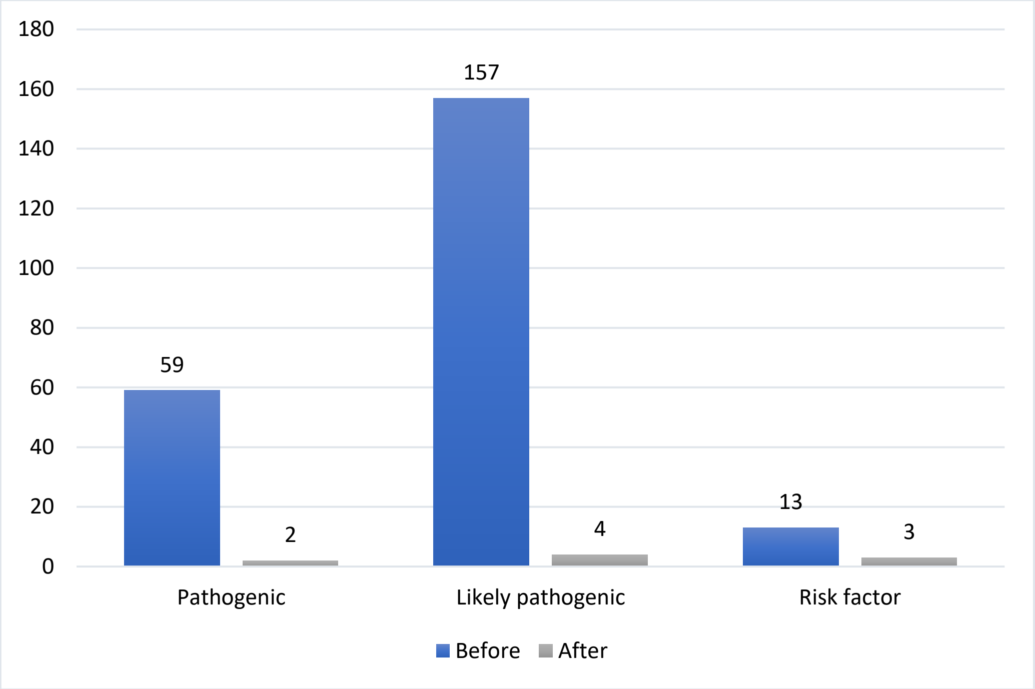


Figure 5.16 Number of variants before and after step F4 (Schizophrenia)

- Step F5

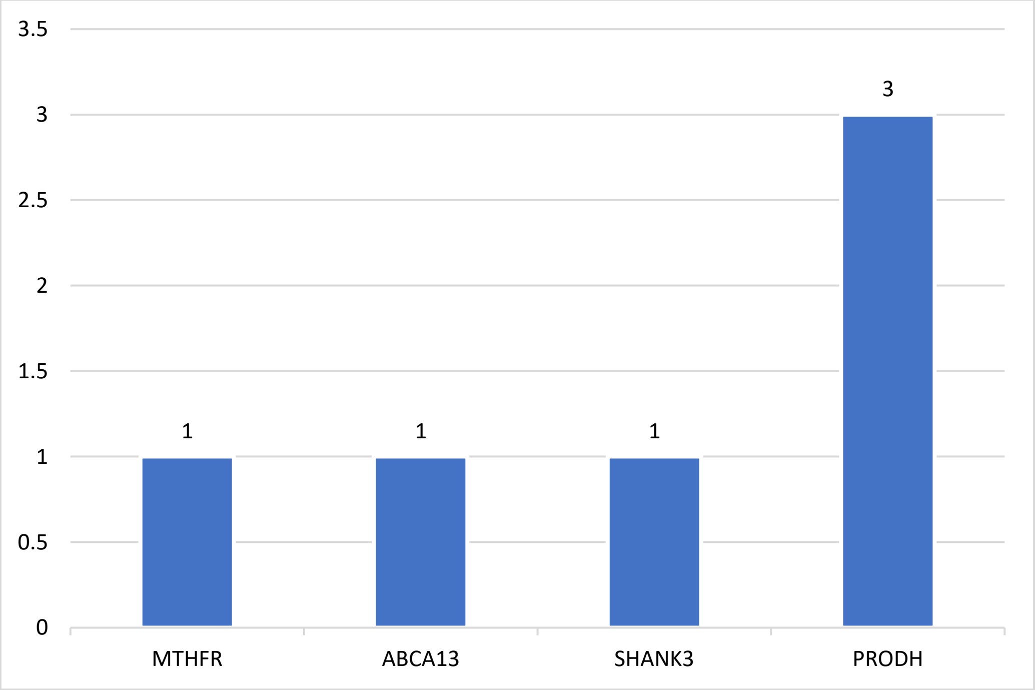


Figure 5.17 Distribution of variants by gene (Schizophrenia)

The above diagram shows the distribution of variants by gene. 3 out of 7 genes were considered unrelated.

- Step F6, F7 and F8

  The analysis of the review status information (F6) identified three variants without assertion criteria provided and have been classified as *To Follow Up*.

The filter F7 and F8 the filters were passed, and the other variants were classified as *Accepted with moderate evidence*.

The impact of the steps F6, F7 and F8 is shown on the graph below.

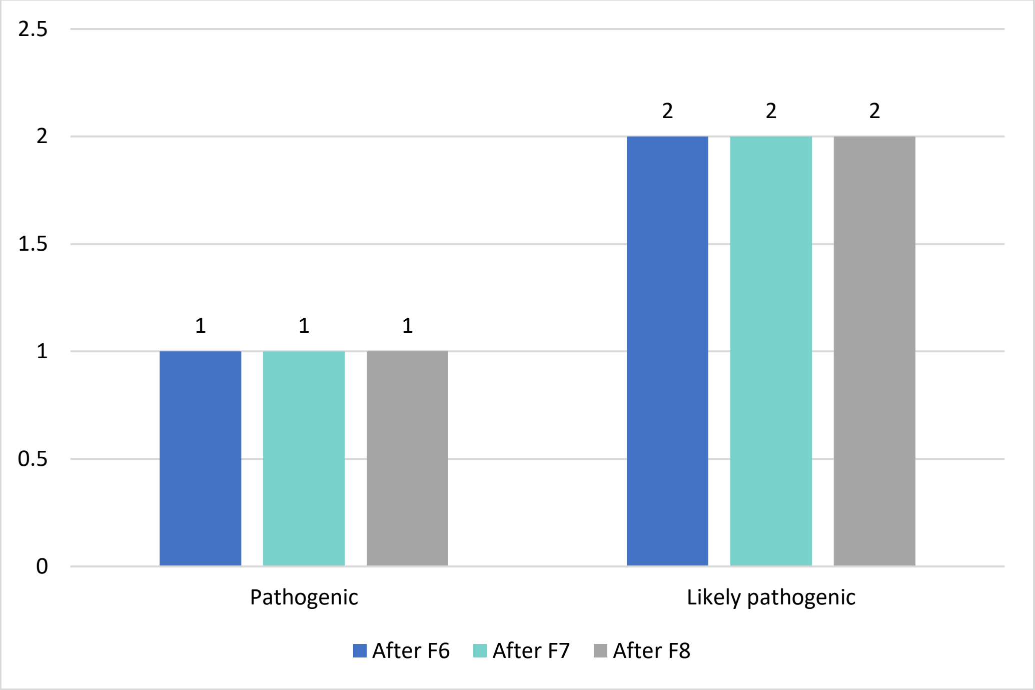


Figure 5.18 Number of variants after step F6, F7 and F8 (Schizophrenia)

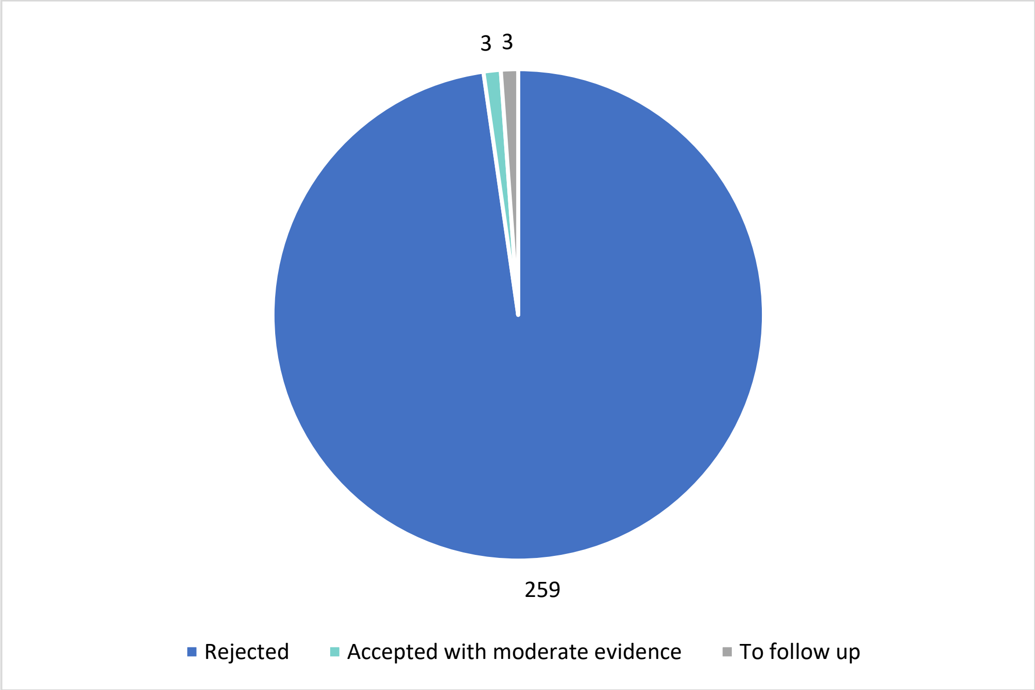


Figure 5.19 Final results of the workflow (Schizophrenia)

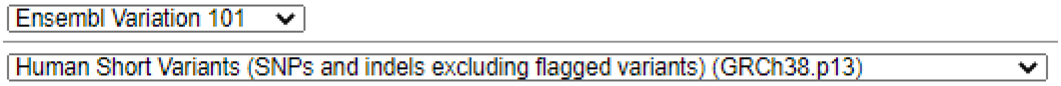
The final list shows that almost none of the variants met the criteria. Three variants were classified as *Accepted with moderate evidence* and three as *To follow up*. The rest 259 were *Rejected*.

5.4 Data completion

The Ensembl BioMart tool was used to complement the expertise.

The data was obtained in the following steps:

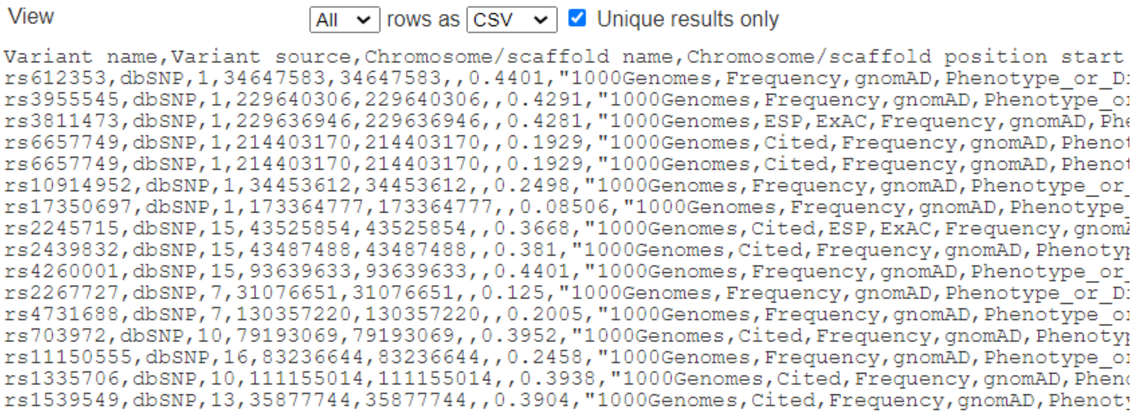
- selecting the appropriate database and dataset



Ensembl Variation 101
Human Short Variants (SNPs and indels excluding flagged variants) (GRCh38.p13)

Figure 5.20 BioMart selection view

- using the Phenotype filter and carefully selecting each variant suitable for the disease
- adding attribute PubMed ID
- downloading data returned in CSV format for further editing



```
Variant name,Variant source,Chromosome/scaffold name,Chromosome/scaffold position start
rs612353,dbSNP,1,34647583,34647583,,0.4401,"1000Genomes,Frequency,gnomAD,Phenotype_or_D
rs3955545,dbSNP,1,229640306,229640306,,0.4291,"1000Genomes,Frequency,gnomAD,Phenotype_o
rs3811473,dbSNP,1,229636946,229636946,,0.4281,"1000Genomes,ESP,ExAC,Frequency,gnomAD,Ph
rs6657749,dbSNP,1,214403170,214403170,,0.1929,"1000Genomes,Cited,Frequency,gnomAD,Pheno
rs6657749,dbSNP,1,214403170,214403170,,0.1929,"1000Genomes,Cited,Frequency,gnomAD,Pheno
rs10914952,dbSNP,1,34453612,34453612,,0.2498,"1000Genomes,Frequency,gnomAD,Phenotype_or_
rs17350697,dbSNP,1,173364777,173364777,,0.08506,"1000Genomes,Frequency,gnomAD,Phenotype_
rs2245715,dbSNP,15,43525854,43525854,,0.3668,"1000Genomes,Cited,ESP,ExAC,Frequency,gnom
rs2439832,dbSNP,15,43487488,43487488,,0.381,"1000Genomes,Cited,Frequency,gnomAD,Phenoty
rs4260001,dbSNP,15,93639633,93639633,,0.4401,"1000Genomes,Frequency,gnomAD,Phenotype_or_
rs2267727,dbSNP,7,31076651,31076651,,0.125,"1000Genomes,Frequency,gnomAD,Phenotype_or_D
rs4731688,dbSNP,7,130357220,130357220,,0.2005,"1000Genomes,Frequency,gnomAD,Phenotype_o
rs703972,dbSNP,10,79193069,79193069,,0.3952,"1000Genomes,Cited,Frequency,gnomAD,Phenoty
rs11150555,dbSNP,16,83236644,83236644,,0.2458,"1000Genomes,Frequency,gnomAD,Phenotype_o
rs1335706,dbSNP,10,111155014,111155014,,0.3938,"1000Genomes,Cited,Frequency,gnomAD,Phen
rs1539549,dbSNP,13,35877744,35877744,,0.3904,"1000Genomes,Cited,Frequency,gnomAD,Phenot
```

Figure 5.21 Search result in BioMart

As a result, from the searched data confronted with the previously selected variants, information about the related literature was extracted, adding the PubMed identifier to the final list and thus making it more complete. The final output is placed in Appendix B.

The amount of additional data extracted from the supplementary database depends on the individual subsequent use of the data. The more complex the project and purpose is, the more detailed the selected data should be.

In addition, a data set was performed to illustrate the complexity of the data and the variety of not fully compatible repositories. This summary, together with a commentary, can be found in the next chapter in the part “The heterogeneity of repositories”.

After completing the identification stage, the next stages that the obtained genomic variants must go through are loading and exploitation.

These stages of the SILE method were not implemented because they are outside the objectives and scope of this master’s thesis.

CHAPTER 6. RESULTS ANALYSIS

6.1 Relations between variants

One of the aims of the study was to check if there is a relationship between the particular phenotypes through the variant-phenotype pathway. In other words, whether there are variants assigned to more than one disorder.

The table below presents information on common variants, found in Clinvar, for the discussed diseases.

Table 6.1 Variant-phenotypes relation

Phenotypes relation	Variant ID	Clinical significance	Identification
ASD and ADHD	143458	Benign/Likely benign	Rejected
	267902	Pathogenic	Rejected
ASD and Schizophrenia	585115	not provided	Rejected
ADHD and Schizophrenia	4011	uncertain significance	Rejected
ASD, ADHD and Schizophrenia	-	-	-

As can be seen, there are few variants that were obtained when searching for two diseases, but these are variants identified as *Rejected*. It means that the dependencies are still at the investment stage and cannot be considered relevant in practice. Nevertheless, this is still promising information for future research.

6.2 Gene-based perspective

The mere fact of not finding clinically relevant common variants does not mean that there are no genetic links. The gene perspective can give a broader look at variant-phenotype interactions.

As the common genetic factors are already documented in the scientific literature, only the genetic relationships of the variants considered at step F5 of the workflow were taken into account.

This assumption was made with the aim to consider only the genes verified for their relationship with the disease and relevant to the variants analyzed in the thesis.

Table 6.2 Gene-phenotypes relation

Phenotypes relation	Gene
ASD and ADHD	-
ASD and Schizophrenia	SHANK3
ADHD and Schizophrenia	-
ASD, ADHD and Schizophrenia	-

The table indicates that only variants relating to ASD and schizophrenia are associated with the shared gene.

Exactly three variations correspond to the conditions and are presented below together with their characteristics of clinical significance and final identification.

Table 6.3 Variants associated with the SHANK3 gene

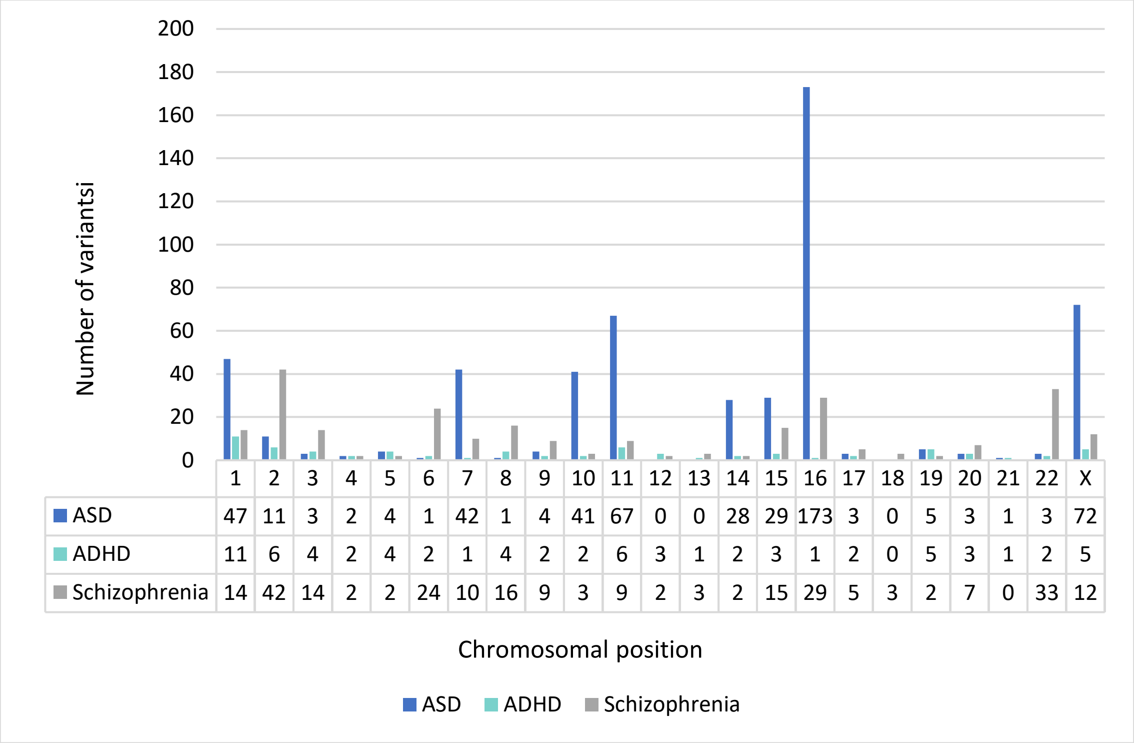
Variant ID	dbSNP	Clinical significance	Identification	Phenotype
208759	rs762292772	Pathogenic	To follow up	ASD
372707	rs1555910143	Pathogenic	To follow up	ASD
397528	rs1555910162	Pathogenic	Accepted with Moderate Evidence	Schizophrenia

SHANK3 gene is a postsynaptic protein an having influence on neural morphology. It is crucial for synaptic transmission and plasticity and because of that each mutation can cause various neuropsychiatric disorders.

This means, there is a significant correlation between the Autism Spectrum Disorder and Schizophrenia because their genetic trigger is located in the same cerebral region.

The genetic factors influencing mental diseases and the relationship between them are very complex. The determinants cannot be simplified to a single gene, because the probability of developing a disease can occur through the influence of one or more mutations and their combination with the environmental factors shortly introduced in the theoretical part.

The following chart shows the distribution of variants within specific chromosomes indicating wide differences in genetic position. The summary was created for each of the variants for which the chromosomal position was determined, the rest was ignored. The distribution of variants in particular chromosomal locations can be seen in a graph below.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X
■ ASD	47	11	3	2	4	1	42	1	4	41	67	0	0	28	29	173	3	0	5	3	1	3	72
■ ADHD	11	6	4	2	4	2	1	4	2	2	6	3	1	2	3	1	2	0	5	3	1	2	5
■ Schizophrenia	14	42	14	2	2	24	10	16	9	3	9	2	3	2	15	29	5	3	2	7	0	33	12

Figure 6.1 Chromosomal location of variants

There is a noticeable large number of variants on the chromosome 16 for ASD. Scientific publications confirm that changes in chromosome 16 are firmly linked to Autism. [109,110]

In other cases, the values do not indicate so visible preferred location. Mutations can occur within almost every chromosome which only proves the complexity of the disease genetic architecture.

6.3 The heterogeneity of repositories

A problem encountered when searching for suitable repositories is the heterogeneity of these databases. It was analyzed whether Ensemble can also be used for analysis.

Despite a very complex database structure and tools providing easy access to a lot of information, which is missing in other repositories, it is not possible to apply the SILE workflow methodology used in this thesis to the Ensembl repository.

Standardization in accordance with the ACMG guideline is a key prerequisite for the database to be considered in the context of acquiring information on variants for clinical practice.

In addition, information about the clinical significance of the variant is necessary to fulfill the quality criteria, and Ensembl does not provide such information itself. The only searches in which this information appears are provided by connection to the Clinvar database and do not bring any new value as this repository was already explored.

However, Ensembl provides an access to relevant information that is not easily navigable in Clinvar. While Clinvar does not allow to query PubMed IDs value from the search engine level of the website, Ensembl enables to access such information through filtering tools. In Clinvar it is necessary to enter each variant individually and manually check if there is any literature associated. This value is important because it informs about the related literature evidence of the variant increasing its significance. As a result, Ensembl was used to complete the schematic diagram with this value.

Another problem with heterogeneity is data discrepancy. As can be seen from the summary of these two databases: Clinvar and Ensembl, a large number of variants were found only in one of the two repositories. This means that the second part of data can be equally important and should also be checked with appropriate tools in the future.

This, however, requires a redefinition of the approach and development of other steps than those used in this work.

The table below shows the number of variants obtained for a given phenotype in the Clinvar and Ensembl databases, as well as how many of these variants were found in both repositories.

Table 6.4 Number of variants found depending on the source

Phenotype	Clinvar	Ensembl	Clinvar and Ensembl
ASD	682	903	410
ADHD	99	242	2
Schizophrenia	265	1525	27

It can be seen that the repositories do not cover each other's datasets what creates the need to extend the analysis of variants to other repositories.

CHAPTER 7. CONCLUSIONS AND FUTURE WORK

Genomic data management is a very complex challenge and requires understanding from the principles of biological structures to advanced data analysis tools and frameworks.

The preparation of this master’s thesis allowed me to contribute in a small part to the development in the field and enabled me to better understand the genomic science that has a chance to revolutionize today's medicine.

The main aim of my master’s thesis was to identify genetic variants associated with mental or behavioral disorders: Autism Spectrum Disorder, Attention Deficit Hyperactivity Disorder and Schizophrenia and the objective was achieved through two phases:

In phase 1, relevant and functional genomic databases were analyzed based on pre-established quality criteria and successfully selected to be later implemented in the variant-disease association study. During this task, however, it was realized that despite the dynamic development of databases full of ready-to-use information, there is no single repository that meets all needs and it is necessary to draw knowledge from many sources.

In phase 2, a number of quality criteria were applied in accordance with the SILE methodology for the purpose of the precise variation identification. This approach allowed me to face the complexity of efficient genomic information management. The direct demand for the correct management of genomic "Big Data" was experienced.

Only a little amount of data passed the workflow filters. Of all the initial data, 7.33% of variants identified for ASD and 1.16% for Schizophrenia are relevant. Additionally, 0.44% of ASD variants, 3.03% of ADHD, and 1.16% of Schizophrenia are considered promising. This is due to the high-quality requirements for variants designed for the clinical practice usage. The repositories store a huge amount of data, but not all of it is relevant.

At some steps, the analysis of data was very time-consuming. The appropriate tools for efficient information filtering or searching are not always available which in consequence leads to manual checking of each individual data to meet the quality criterion.

The wide range of names and categories of the diseases was another factor that made work difficult. It requires a lot of time to analyze each option to ensure whether it belongs within the scope of the disease and whether such data can be included in the study.

All these factors show how necessary development in the field of bioinformatics directed at data management is. This is an essential step towards Precise Medicine, which is increasingly within reach.

The expanding amount of data and development of techniques for its assessment offer the prospect of a change in approach to the treatment of mental illnesses, which are still a neglected field of medicine.

During the development of this project, certain ideas have emerged that might have a contribution to the field of Genomic Data Sciences. There are several objectives to be achieved in the near future:

- Application of the methodology to more repositories extending the number of relevant variants.
- Development of the SILE methodology to improve the process of variants selection.
- Creation of programming tools to facilitate search and identification of desired data.
- Construction and maintenance of database related to links between genomic structures based on the Conceptual Scheme of the Human Genome.

REFERENCES

[1] Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. XThe next-generation sequencing revolution and its impact on genomics. *Cell*. 2013;155(1):27. doi:10.1016/j.cell.2013.09.006

[2] Sujansky W. Heterogeneous database integration in biomedicine. *J Biomed Inform*. 2001;34(4):285-298. doi:10.1006/jbin.2001.1024

[3] Genomic Information Systems - PROS - Research Center on Software Production MethodsPROS – Research Center on Software Production Methods. Accessed September 18, 2020. http://www.pros.webs.upv.es/genomic-information-systems/

[4] What is the PGC? | Psychiatric Genomics Consortium. Accessed September 18, 2020. https://www.med.unc.edu/pgc/

[5] What is precision medicine? - Genetics Home Reference - NIH. Accessed September 16, 2020. https://ghr.nlm.nih.gov/primer/precisionmedicine/definition

[6] About | 1000 Genomes. Accessed September 16, 2020. https://www.internationalgenome.org/about#1000G_PROJECT

[7] ARTICLE A global reference for human genetic variation The 1000 Genomes Project Consortium*. doi:10.1038/nature15393

[8] 1000 Genomes | A Deep Catalog of Human Genetic Variation. Accessed September 18, 2020. https://www.internationalgenome.org/

[9] Homo_sapiens - Ensembl genome browser 101. Accessed September 16, 2020. http://www.ensembl.org/Homo_sapiens/Info/Index

[10] Altshuler DM, Gibbs RA, Peltonen L, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52-58. doi:10.1038/nature09298

[11] International HapMap Project. Accessed September 16, 2020. https://www.genome.gov/10001688/international-hapmap-project

[12] What is the International HapMap Project? - Genetics Home Reference - NIH. Accessed September 16, 2020. https://ghr.nlm.nih.gov/primer/genomicresearch/hapmap

[13] NCBI. Accessed September 16, 2020. https://www.ncbi.nlm.nih.gov/variation/news/NCBI_retiring_HapMap/

[14] Boeke JD, Church G, Hessel A, et al. Logo of the first Human Genome Project. Image courtesy of the U.S. Department of Energy Genomic Science program. Science (80-). 2016;353(6295):126-127. doi:10.1126/science.aaf6850

[15] Jasny BR, Kennedy D. THE HUMAN GENOME. Science (80-). 2001;291(5507):1153-1153.

[16] What is the Encyclopedia of DNA Elements (ENCODE) Project? - Genetics Home Reference - NIH. Accessed September 17, 2020. https://ghr.nlm.nih.gov/primer/genomicresearch/encode

[17] ENCODE: Encyclopedia of DNA Elements – ENCODE. Accessed September 17, 2020. https://www.encodeproject.org/

[18] X Chromosome. Accessed September 17, 2020. https://www.genome.gov/genetics-glossary/X-Chromosome

[19] Nobel Prize in Medicine: 1901-Present | Live Science. Accessed September 17, 2020. https://www.livescience.com/16342-nobel-prize-medicine-history-list.html

[20] Uzman A. Molecular biology of the cell (4th ed.): Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. Biochem Mol Biol Educ. 2003;31(4):212-214. doi:10.1002/bmb.2003.494031049999

[21] Life Sciences Cyberbridge. Accessed September 17, 2020. http://cyberbridge.mcb.harvard.edu/dna_1.html

[22] The Human Genome - Genomes - NCBI Bookshelf. Accessed September 17, 2020. https://www.ncbi.nlm.nih.gov/books/NBK21134/

[23] File:Gene Intron Exon nb.svg - Wikimedia Commons. Accessed September 17, 2020. https://commons.wikimedia.org/wiki/File:Gene_Intron_Exon_nb.svg

[24] Elston RC, Satagopan JM, Sun S. Genetic terminology. Methods Mol Biol. 2012;850:1-9. doi:10.1007/978-1-61779-555-8_1

[25] Mutations: Types and Causes - Molecular Cell Biology - NCBI Bookshelf. Accessed September 17, 2020. https://www.ncbi.nlm.nih.gov/books/NBK21578/

[26] Bruford EA, Braschi B, Denny P, Jones TEM, Seal RL, Tweedie S. Guidelines for human gene nomenclature. Nat Genet. 2020;52(8):754-758. doi:10.1038/s41588-020-0669-3

[27] Blanco A, Blanco G. Proteins. In: Medical Biochemistry. Elsevier; 2017:21-71. doi:10.1016/B978-0-12-803550-4.00003-3

[28] Transcription, Translation and Replication. Accessed September 17, 2020. https://www.atdbio.com/content/14/Transcription-Translation-and-Replication

[29] Translation: DNA to mRNA to Protein | Learn Science at Scitable. Accessed September 18, 2020. https://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393/

[30] Haplotype - an overview | ScienceDirect Topics. Accessed September 17, 2020. https://www.sciencedirect.com/topics/neuroscience/haplotype

[31] Neigenfind J, Gyetvai G, Basekow R, et al. Haplotype inference from unphased SNP data in heterozygous polyploids based on SAT. BMC Genomics. 2008;9(1):356. doi:10.1186/1471-2164-9-356

[32] Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405-424. doi:10.1038/gim.2015.30

[33] mutation | Learn Science at Scitable. Accessed September 17, 2020. https://www.nature.com/scitable/definition/mutation-8/

[34] Griffiths AJ, Miller JH, Suzuki DT, Lewontin RC, Gelbart WM. Genetic variation. Published online 2000.

[35] Phenotype. Accessed September 17, 2020. https://www.genome.gov/genetics-glossary/Phenotype

[36] Zhou Q. Academic Libraries in Research Data Management Service: Perceptions and Practices. OALib. 2018;05(06):1-4. doi:10.4236/oalib.1104693

[37] León Palacio A, Pastor López Ó. Smart Data for Genomic Information Systems: the SILE Method. Complex Syst Informatics Model Q. 2018;0(17):1-23. doi:10.7250/csimq.2018-17.01

[38] Introduction. In: Conceptual Modeling of Information Systems. Springer Berlin Heidelberg; 2007:1-36. doi:10.1007/978-3-540-39390-0_1

[39] Pastor O, Palacio AL, Román JFR, Casamayor JC. Modeling Life: A Conceptual Schema-centric Approach to Understand the Genome MMQEF: a framework to evaluate quality in MDE environments View project. Published online 2017. doi:10.1007/978-3-319-67271-7_3

[40] Fabián Reyes Román J. Diseño y Desarrollo de Un Sistema de Información Genómica Basado En Un Modelo Conceptual Holístico Del Genoma Humano (Doctoral Dissertation).; 2018.

[41] León Palacio A. SILE: A Method for the Efficient Management of Smart Genomic Information. Published online October 18, 2019. doi:10.4995/Thesis/10251/131698

[42] Sturm H, Fernell E, Gillberg C. Autism spectrum disorders in children with normal intellectual levels: Associated impairments and subgroups. Dev Med Child Neurol. 2004;46(7):444-447. doi:10.1017/S0012162204000738

[43] Rommelse NNJ, Franke B, Geurts HM, Hartman CA, Buitelaar JK. Shared heritability of attention-deficit/hyperactivity disorder and autism spectrum disorder. doi:10.1007/s00787-010-0092-x

[44] Hofvander B, Delorme R, Chaste P, et al. Psychiatric and psychosocial problems in adults with normal-intelligence autism spectrum disorders. BMC Psychiatry. 2009;9(1):35. doi:10.1186/1471-244X-9-35

[45] Barneveld PS, Pieterse J, de Sonneville L, et al. Overlap of autistic and schizotypal traits in adolescents with Autism Spectrum Disorders. Schizophr Res. 2011;126(1-3):231-236. doi:10.1016/j.schres.2010.09.004

[46] Chisholm K, Lin A, Abu-Akel A, Wood SJ. The association between autism and schizophrenia spectrum disorders: A review of eight alternate models of co-occurrence. Neurosci Biobehav Rev. 2015;55:173-183. doi:10.1016/j.neubiorev.2015.04.012

[47] Van Der Meer JMJ, Oerlemans AM, Van Steijn DJ, et al. Are autism spectrum disorder and attention-deficit/hyperactivity disorder different manifestations of one overarching disorder? Cognitive and symptom evidence from a clinical and population-based sample. J Am Acad Child Adolesc Psychiatry. 2012;51(11). doi:10.1016/j.jaac.2012.08.024

[48] Taurines R, Schwenck C, Westerwald E, Sachse M, Siniatchkin M, Freitag C. ADHD and autism: Differential diagnosis or overlapping traits? A selective review. ADHD Atten Deficit Hyperact Disord. 2012;4(3):115-139. doi:10.1007/s12402-012-0086-2

[49] American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5) - American Psychiatric Association.; 2013.

[50] Abbott AE, Nair A, Keown CL, et al. Patterns of Atypical Functional Connectivity and Behavioral Links in Autism Differ between Default, Salience, and Executive Networks. Cereb Cortex. 2016;26(10):4034-4045. doi:10.1093/cercor/bhv191

[51] Sripada CS, Kessler D, Angstadt M. Lag in maturation of the brain's intrinsic functional architecture in attention-deficit/hyperactivity disorder. Proc Natl Acad Sci U S A. 2014;111(39):14259-14264. doi:10.1073/pnas.1407787111

[52] McCarthy H, Skokauskas N, Mulligan A, et al. Attention network hypoconnectivity with default and affective network hyperconnectivity in adults diagnosed with attention-deficit/hyperactivity disorder in childhood. JAMA Psychiatry. 2013;70(12):1329-1337. doi:10.1001/jamapsychiatry.2013.2174

[53] Berman RA, Gotts SJ, McAdams HM, et al. Disrupted sensorimotor and social-cognitive networks underlie symptoms in childhoodonset schizophrenia. Brain. 2016;139(1):276-291. doi:10.1093/brain/awv306

[54] Verhoeff B. Autism in flux: a history of the concept from Leo Kanner to DSM-5. Hist Psychiatry. 2013;24(4):442-458. doi:10.1177/0957154X13500584

[55] Kanner, L., & others. Autistic disturbances of affective contact. Nervous Child. 1943;2(3);217–250.

[56] Eisenberg L, Kanner L. Childhood schizophrenia: Symposium, 1955: 6. Early infantile autism, 1943–55. Am J Orthopsychiatry. 1956;26(3):556-566. doi:10.1111/j.1939-0025.1956.tb06202.x

[57] Asperger H. Die "Autistischen Psychopathen" im Kindesalter. Arch Psychiatr Nervenkr. 1944;117(1):76-136. doi:10.1007/BF01837709

[58] Feinstein A. A History of Autism. Wiley-Blackwell; 2010. doi:10.1002/9781444325461

[59] Lai MC, Lombardo M V., Baron-Cohen S. Autism. In: The Lancet. Vol 383. Lancet Publishing Group; 2014:896-910. doi:10.1016/S0140-6736(13)61539-1

[60] "Colvert E, Tick B, McEwen F, Stewart C, Curran SR, Woodhouse E, et al. Heritability of Autism Spectrum Disorder in a UK Population-Based Twin Sample. Colver, E, Tick B, McEwen F, Stewart C, Curran SR, Woodhouse E, et al. Heritability of Autism Spectrum Disorder in a UK Population-Based Twin Sample. JAMA Psychiatry. 2015;72(5):415-23."

[61] "Sandin S, Lichtenstein P, Kuja-Halkola R, Larsson H, Hultman CM, Reichenberg A. The familial risk of autism. JAMA. 2014;311(17):1770-7"

[62] "Gaugler T, Klei L, Sanders SJ, Bodea CA, Goldberg AP, Lee AB, et al. Most genetic risk for autism resides with common variation. Nat Genet. 2014;46(8):881-5."

[63] "Yuen RK, Thiruvahindrapuram B, Merico D, Walker S, Tammimies K, Hoang N, et al. Whole-genome sequencing of quartet families with autism spectrum disorder. Nat Med. 2015;21(2):185-91"

[64] "O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat Genet. 2011;43(6):585-9"

[65] "Wang K, Zhang HT, Ma DQ, Bucan M, Glessner JT, Abrahams BS, et al. Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature. 2009;459(7246):528-33"

[66] "Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, et al. Strong association of de novo copy number mutations with autism. Science. 2007;316(5823):445-9."

[67] "State MW, Sestan N. Neuroscience. The emerging biology of autism spectrum disorders. Science. 2012;337(6100):1301-3"

[68] Barkley RA, Peters H. The Earliest Reference to ADHD in the Medical Literature? Melchior Adam Weikard's Description in 1775 of "Attention Deficit" (Mangel der Aufmerksamkeit, Attentio Volubilis). J Atten Disord. 2012;16(8):623-630. doi:10.1177/1087054711432309

[69] Crichton A. An inquiry into the nature and origin of mental derangement: On attention and its diseases. J Atten Disord. 2008;12(3):200-204. doi:10.1177/1087054708315137

[70] Lange KW, Reichl S, Lange KM, Tucha L, Tucha O. The history of attention deficit hyperactivity disorder. ADHD Atten Deficit Hyperact Disord. 2010;2(4):241-255. doi:10.1007/s12402-010-0045-8

[71] Bradley C. THE BEHAVIOR OF CHILDREN RECEIVING BENZEDRINE. Am J Psychiatry. 1937;94(3):577-585. doi:10.1176/ajp.94.3.577

[72] [DSM-III: the 3d edition of the Diagnostic and Statistical Manual of Mental Disorders from the American Psychiatric Association] - PubMed. Accessed September 17, 2020. https://pubmed.ncbi.nlm.nih.gov/3787052/

[73] American Journal of Psychiatry. Diagnostic and Statistical Manual of Mental Disorders, 3rd ed., revised (DSM-III-R). Am J Psychiatry. 1988;145(10):1301-1302. doi:10.1176/ajp.145.10.1301

[74] Thapar A, Cooper M, Rutter M. Neurodevelopmental disorders. The Lancet Psychiatry. 2017;4(4):339-346. doi:10.1016/S2215-0366(16)30376-5

[75] Biederman J, Mick E, Faraone S V. Age-dependent decline of symptoms of attention deficit hyperactivity disorder: Impact of remission definition and symptom type. Am J Psychiatry. 2000;157(5):816-818. doi:10.1176/appi.ajp.157.5.816

[76] Polanczyk G, De Lima MS, Horta BL, Biederman J, Rohde LA. The worldwide prevalence of ADHD: A systematic review and metaregression analysis. Am J Psychiatry. 2007;164(6):942-948. doi:10.1176/ajp.2007.164.6.942

[77] Poelmans G, Pauls DL, Buitelaar JK, Franke B. Integrated genome-wide association study findings: Identification of a neurodevelopmental network for attention deficit hyperactivity disorder. Am J Psychiatry. 2011;168(4):365-377. doi:10.1176/appi.ajp.2010.10070948

[78] Faraone S V., Perlis RH, Doyle AE, et al. Molecular genetics of attention-deficit/hyperactivity disorder. Biol Psychiatry. 2005;57(11):1313-1323. doi:10.1016/j.biopsych.2004.11.024

[79] Alberts-Corush J, Firestone P, Goodman JT. Attention and impulsivity characteristics of the biological and adoptive parents of hyperactive and normal control children. Am J Orthopsychiatry. 1986;56(3):413-423. doi:10.1111/j.1939-0025.1986.tb03473.x

[80] Hawi Z, Segurado R, Conroy J, et al. Preferential transmission of paternal alleles at risk genes in attention-deficit/hyperactivity disorder. Am J Hum Genet. 2005;77(6):958-965. doi:10.1086/498174

[81] Compi L. The natural history of schizophrenia in the long term. Brit. J. Psychiat. 1980;136;413-420 doi:10.1192/bjp.136.5.413

[82] Jablensky A. The Diagnostic Concept of Schizophrenia: Its History, Evolution, and Future Prospects. Vol 12.; 2010.

[83] Saha S, Chant D, Welham J, Mcgrath J, Hyman SE. A Systematic Review of the Prevalence of Schizophrenia. Published online 2005. doi:10.1371/journal.pmed.0020141

[84] Schizophrenia - Genetics Home Reference - NIH. Accessed September 18, 2020. https://ghr.nlm.nih.gov/condition/schizophrenia

[85] van de Leemput J, Hess JL, Glatt SJ, Tsuang MT. Genetics of Schizophrenia: Historical Insights and Prevailing Evidence. Adv Genet. 2016;96:99-141. doi:10.1016/bs.adgen.2016.08.001

[86] Adult Psychopathology and Diagnosis - Google Books. Accessed September 17, 2020. https://books.google.es/books?hl=pl&lr=&id=Q8hTDwAAQBAJ&oi=fnd&pg=PR10&dq=Etiological+considerations%22.+Adult+psychopathology+and+diagnosis.+(6th+ed.).+New+Jersey:+John+Wiley+%26+Sons&ots=nhUyJHs-jT&sig=PnPIkMNBwWsTbNW3thg-WxyxMeU#v=onepage&q&f=false

[87] Navarrete-Hidalgo M, Reyes Román JF, Pastor López Ó. Design and Implementation of a Geis for the Genomic Diagnosis using the SILE Methodology. Case Study: Congenital Cataract. In: Proceedings of the 13th International Conference on Evaluation of Novel Approaches to Software Engineering. Vol 2018-March. SCITEPRESS - Science and Technology Publications; 2018:267-274. doi:10.5220/0006705802670274

[88] Román JFR, López ÓP. Use of GeIS for Early Diagnosis of Alcohol Sensitivity. In: Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies. SCITEPRESS - Science and and Technology Publications; 2016:284-289. doi:10.5220/0005822902840289

[89] Burriel V, Reyes RJF, Casanoves AH, Iniguez-Jarrin C, Leon A. GeIS based on Conceptual Models for the risk assessment of Neuroblastoma. In: Proceedings - International Conference on Research Challenges in Information Science. IEEE Computer Society; 2017:451-452. doi:10.1109/RCIS.2017.7956581

[90] About | Nucleic Acids Research | Oxford Academic. Accessed September 17, 2020. https://academic.oup.com/nar/pages/About

[91] DATABASES & TOOLS | Human Genome Variation Society. Accessed September 17, 2020. http://www.hgvs.org/content/databases-tools

[92] What is ClinVar? Accessed September 17, 2020. https://www.ncbi.nlm.nih.gov/clinvar/intro/

[93] Landrum MJ, Kattman BL. ClinVar at five years: Delivering on the promise. Hum Mutat. 2018;39(11):1623-1630. doi:10.1002/humu.23641

[94] ClinVar. Accessed September 17, 2020. https://www.ncbi.nlm.nih.gov/clinvar/

[95] Sequence Variant Nomenclature. Accessed September 17, 2020. https://varnomen.hgvs.org/

[96] ClinVar: A Central Repository for Clinically Relevant Variants - Meli…. Accessed September 17, 2020. https://www.slideshare.net/variomeproj/clinvar-a-central-repository-for-clinically-relevant-variants-melissa-j-landrum

[97] Landrum MJ, Lee JM, Benson M, et al. ClinVar: Improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062-D1067. doi:10.1093/nar/gkx1153

[98] Hubbard T, Barker D, Birney E, et al. The Ensembl genome database project. Nucleic Acids Res. 2002;30(1):38-41. doi:10.1093/nar/30.1.38

[99] Ensembl genome browser 101. Accessed September 18, 2020. https://www.ensembl.org/index.html

[100] Ensembl Tools. Accessed September 18, 2020. https://www.ensembl.org/info/docs/tools/index.html

[101] McKusick VA. Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet. 2007;80(4):588-604. doi:10.1086/514346

[101] Chong JX, Buckingham KJ, Jhangiani SN, et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet. 2015;97(2):199-215. doi:10.1016/j.ajhg.2015.06.009

[102] OMIM Turns 50: A Genetic Database's Past, Present, and Future. Accessed September 17, 2020. https://www.hopkinsmedicine.org/research/advancements-in-research/fundamentals/in-depth/omim-turns-50-a-genetic-databases-past-present-and-future

[103] OMIM Entry Statistics. Accessed September 17, 2020. https://www.omim.org/statistics/entry

[104] OMIM Gene Map Statistics. Accessed September 17, 2020. https://www.omim.org/statistics/geneMap

[105] OMIM - Online Mendelian Inheritance in Man. Accessed September 17, 2020. https://www.omim.org/

[106] Representation of clinical significance in ClinVar and other variation resources at NCBI. Accessed September 18, 2020. https://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/

[107] Harrison SM, Riggs ER, Maglott DR, et al. Using ClinVar as a Resource to Support Variant Interpretations. doi:10.1002/0471142905.hg0816s89

[108] Instructions for ClinVar submission spreadsheets. Accessed September 18, 2020. https://www.ncbi.nlm.nih.gov/clinvar/docs/spreadsheet/#collection

[109] OMIM Entry - # 611913 - CHROMOSOME 16p11.2 DELETION SYNDROME, 593-KB. Accessed September 20, 2020. https://www.omim.org/entry/611913

[110] Weiss LA, Shen Y, Korn JM, et al. Association between microdeletion and microduplication at 16p11.2 and autism. N Engl J Med. 2008;358(7):667-675. doi:10.1056/NEJMoa075974

APPENDIX A

Table A. Subcategories of mental disorders discussed in the master's thesis, sorted according to OMIM

Acronym	Phenotype MIM number	Alternative titles; symbols
	Schizophrenia	
SCZD	# 181500	SCHIZOPHRENIA, SCHIZOPHRENIA WITH OR WITHOUT AN AFFECTIVE DISORDER
SCZD1	% 181510	SCHIZOPHRENIA 1, SCHIZOPHRENIA 1 WITH OR WITHOUT AN AFFECTIVE DISORDER, SCHIZOPHRENIA SUSCEPTIBILITY LOCUS, CHROMOSOME 5-RELATED
SCZD2	% 603342	SCHIZOPHRENIA 2, SCHIZOPHRENIA SUSCEPTIBILITY LOCUS, CHROMOSOME 11q-RELATED
SCZD3	% 600511	SCHIZOPHRENIA 3, SCHIZOPHRENIA 3 WITH OR WITHOUT AN AFFECTIVE DISORDER, SCHIZOPHRENIA SUSCEPTIBILITY LOCUS, CHROMOSOME 6-RELATED
SCZD4	# 600850	SCHIZOPHRENIA 4, SCHIZOPHRENIA SUSCEPTIBILITY LOCUS, CHROMOSOME 22q11-RELATED
SCZD5	% 603175	SCHIZOPHRENIA 5, SCHIZOPHRENIA 5 WITH OR WITHOUT AN AFFECTIVE DISORDER, SCHIZOPHRENIA SUSCEPTIBILITY LOCUS, CHROMOSOME 6q-RELATED
SCZD6	% 603013	SCHIZOPHRENIA 6, SCHIZOPHRENIA SUSCEPTIBILITY LOCUS, CHROMOSOME 8p-RELATED
SCZD7	% 603176	SCHIZOPHRENIA 7, SCHIZOPHRENIA 7 WITH OR WITHOUT AN AFFECTIVE DISORDER, SCHIZOPHRENIA SUSCEPTIBILITY LOCUS, CHROMOSOME 13q-RELATED
SCZD8	% 603206	SCHIZOPHRENIA 8, SCHIZOPHRENIA 8 WITH OR WITHOUT AN AFFECTIVE DISORDER, SCHIZOPHRENIA SUSCEPTIBILITY LOCUS, CHROMOSOME 18-RELATED
SCZD9	# 604906	SCHIZOPHRENIA 9, SCHIZOPHRENIA 9 WITH OR WITHOUT AN AFFECTIVE DISORDER, SCHIZOPHRENIA SUSCEPTIBILITY LOCUS, CHROMOSOME 1q42-RELATED
SCZD10	% 605419	SCHIZOPHRENIA 10, SCHIZOPHRENIA SUSCEPTIBILITY LOCUS, CHROMOSOME 15q15-RELATED, CATATONIA, PERIODIC
SCZD11	% 608078	SCHIZOPHRENIA 11, SCHIZOPHRENIA SUSCEPTIBILITY LOCUS, CHROMOSOME 10q-RELATED

Acronym	Phenotype MIM number	Alternative titles; symbols
SCZD12	% 608543	SCHIZOPHRENIA 12, SCHIZOPHRENIA SUSCEPTIBILITY LOCUS, CHROMOSOME 1p-RELATED
SCZD13	% 613025	SCHIZOPHRENIA 13, SCHIZOPHRENIA SUSCEPTIBILITY LOCUS, CHROMOSOME 15q13-q14-RELATED
SCZD14	% 612361	SCHIZOPHRENIA 14, SCHIZOPHRENIA SUSCEPTIBILITY LOCUS, CHROMOSOME 2q32-RELATED
SCZD15	# 613950	SCHIZOPHRENIA 15, SCHIZOPHRENIA 15 WITH OR WITHOUT AN AFFECTIVE DISORDER, SCHIZOPHRENIA SUSCEPTIBILITY LOCUS, CHROMOSOME 22q13-RELATED
SCZD16	# 613959	SCHIZOPHRENIA 16, SCHIZOPHRENIA SUSCEPTIBILITY LOCUS, CHROMOSOME 7q36.3-RELATED, CHROMOSOME 7q36.3 DUPLICATION SYNDROME, 362-KB
SCZD17	# 614332	SCHIZOPHRENIA 17, CHROMOSOME 2p16.3 DELETION SYNDROME
SCZD18	# 615232	SCHIZOPHRENIA 18, SCHIZOPHRENIA 18 WITH OR WITHOUT AN AFFECTIVE DISORDER
SCZD19	# 617629	SCHIZOPHRENIA 19, SCHIZOPHRENIA 19 WITH OR WITHOUT AN AFFECTIVE DISORDER
Attention Deficit Hyperactivity Disorder		
ADHD	# 143465	ATTENTION DEFICIT-HYPERACTIVITY DISORDER, HYPERACTIVITY OF CHILDHOOD
ADHD1	% 608903	ATTENTION DEFICIT-HYPERACTIVITY DISORDER, SUSCEPTIBILITY TO, 1
ADHD2	% 608904	ATTENTION DEFICIT-HYPERACTIVITY DISORDER, SUSCEPTIBILITY TO, 2
ADHD3	% 608905	ATTENTION DEFICIT-HYPERACTIVITY DISORDER, SUSCEPTIBILITY TO, 3
ADHD4	% 608906	ATTENTION DEFICIT-HYPERACTIVITY DISORDER, SUSCEPTIBILITY TO, 4
ADHD5	% 612311	ATTENTION DEFICIT-HYPERACTIVITY DISORDER, SUSCEPTIBILITY TO, 5
ADHD6	% 612312	ATTENTION DEFICIT-HYPERACTIVITY DISORDER, SUSCEPTIBILITY TO, 6
ADHD7	# 613003	ATTENTION DEFICIT-HYPERACTIVITY DISORDER, SUSCEPTIBILITY TO, 7
Autism Spectrum Disorder		
AUTS1	% 209850	AUTISTIC DISORDER, AUTISM, SUSCEPTIBILITY TO, 1, INCLUDED; INCLUDED, AUTISM SPECTRUM DISORDER, INCLUDED; ASD, INCLUDED

Acronym	Phenotype MIM number	Alternative titles; symbols
AUTS3	% 608049	AUTISM, SUSCEPTIBILITY TO, 3
AUTS5	# 606053	INTELLECTUAL DEVELOPMENTAL DISORDER WITH AUTISM AND SPEECH DELAY; IDDAS; PHRASE SPEECH DELAY, AUTISM-RELATED, AUTISM-RELATED SPEECH DELAY, AUTISM, SUSCEPTIBILITY TO, 5, FORMERLY
AUTS6	% 609378	AUTISM, SUSCEPTIBILITY TO, 6
AUTS7	% 610676	AUTISM, SUSCEPTIBILITY TO, 7
AUTS8	% 607373	AUTISM, SUSCEPTIBILITY TO, 8
AUTS9	% 611015	AUTISM, SUSCEPTIBILITY TO, 9
AUTS10	% 611016	AUTISM, SUSCEPTIBILITY TO, 10
AUTS11	% 610836	AUTISM, SUSCEPTIBILITY TO, 11
AUTS12	% 610838	AUTISM, SUSCEPTIBILITY TO, 12
AUTS13	% 610908	AUTISM, SUSCEPTIBILITY TO, 13
AUTS14A	# 611913	AUTISM, SUSCEPTIBILITY TO, 14A, INCLUDED; CHROMOSOME 16p11.2 DELETION SYNDROME, 593-KB
AUTS14B	# 614671	AUTISM, SUSCEPTIBILITY TO, 14B, INCLUDED; CHROMOSOME 16p11.2 DUPLICATION SYNDROME
AUTS15	# 612100	AUTISM, SUSCEPTIBILITY TO, 15
AUTS16	# 613410	AUTISM, SUSCEPTIBILITY TO, 16
AUTS17	# 613436	AUTISM, SUSCEPTIBILITY TO, 17
AUTS18	# 615032	AUTISM, SUSCEPTIBILITY TO, 18
AUTS19	# 615091	AUTISM, SUSCEPTIBILITY TO, 19
AUTS20	# 618830	AUTISM, SUSCEPTIBILITY TO, 20
AUTSX1	# 300425	AUTISM, SUSCEPTIBILITY TO, X-LINKED 1
AUTSX2	# 300495	AUTISM, SUSCEPTIBILITY TO, X-LINKED 2; MENTAL RETARDATION, X-LINKED, INCLUDED
AUTSX3	# 300496	AUTISM, SUSCEPTIBILITY TO, X-LINKED 3
AUTSX4	# 300830	AUTISM, SUSCEPTIBILITY TO, X-LINKED 4; CHROMOSOME Xp22 DELETION SYNDROME
AUTSX5	# 300847	AUTISM, SUSCEPTIBILITY TO, X-LINKED 5
AUTSX6	# 300872	AUTISM, SUSCEPTIBILITY TO, X-LINKED 6
ASPG1	% 608638	ASPERGER SYNDROME, SUSCEPTIBILITY TO, 1
ASPG2	% 608631	ASPERGER SYNDROME, SUSCEPTIBILITY TO, 2;
ASPG3	% 608781	ASPERGER SYNDROME, SUSCEPTIBILITY TO, 3
ASPG4	% 609954	ASPERGER SYNDROME, SUSCEPTIBILITY TO, 4

Acronym	Phenotype MIM number	Alternative titles; symbols
ASPGX1	# 300494	ASPERGER SYNDROME, X-LINKED, SUSCEPTIBILITY TO, 1
ASPGX2	# 300497	ASPERGER SYNDROME, X-LINKED, SUSCEPTIBILITY TO, 2

APPENDIX B

Table B. The final list of selected and segregated variants

Variation ID	dbSNP ID	Gene	Chromosome	Clinical Significance	Bib. ID/Year
Attention Deficit Hyperactivity Disorder					
426541	rs1085307675	GNB5	15	Pathogenic	
254029	rs761399728	GNB5	15	Pathogenic	
3163	rs120074176	TPH2	12	risk factor	71979053 (2008)
Autism Spectrum Disorder					
5493	rs121908445	CNTNAP2	7	Risk factor	18179895 (2008)
11815	rs61750240	MECP2	X	Pathogenic	20301670 (2020)
801771	rs1574152522	TBR1	2	Likely pathogenic	
189500	rs121913293	PTEN	10	Pathogenic	28526761 (2017)
520408	rs1555314317	CHD8	14	Likely pathogenic	
973313		CHD8	14	Pathogenic	
869427		CHD8	14	Pathogenic	
811149	rs556977377	CHD8	14	Pathogenic	
807384	rs1555313027	CHD8	14	Pathogenic	
488481	rs1454466097	CHD8	14	Pathogenic	
437423	rs1555314788	CHD8	14	Pathogenic	
429840	rs1131691627	CHD8	14	Pathogenic	
39630	rs397514552	CHD8	14	Pathogenic	23160955 (2012)
11844	rs179363900	MECP2	X	Pathogenic	23262346 (2013)
143406	rs61752992	MECP2	X	Pathogenic	27799067 (2016)
11829	rs61749721	MECP2	X	Pathogenic	27884167 (2016)
11823	rs28934908	MECP2	X	Pathogenic	25741868 (2015)
11819	rs61751362	MECP2	X	Pathogenic	28785396 (2017)
11811	rs28934906	MECP2	X	Pathogenic	26467025 (2016)
436440	rs1555912102	PTCHD1	X	Likely pathogenic	
560715	rs1564568303	PTEN	10	Likely pathogenic	
428235	rs370795352	PTEN	10	Likely pathogenic	
689447	rs1257124719	PTEN	10	Likely pathogenic	

Variation ID	dbSNP ID	Gene	Chromosome	Clinical Significance	Bib. ID/Year
428253	rs1114167667	PTEN	10	Pathogenic	
211972	rs797045904	PTEN	10	Pathogenic	25741868 (2015)
7833	rs121909231	PTEN	10	Pathogenic	28526761 (2017)
468676	rs1224040268	PTEN	10	Pathogenic	
216987	rs863224909	PTEN	10	Pathogenic	25326637 (2014)
224144	rs869312704	TBR1	2	Likely pathogenic	30268909 (2018)
807510	rs1574152672	TBR1	2	Pathogenic	
7850	rs121909240	PTEN	10	Likely pathogenic	15805158 (2005)
7849	rs121909239	PTEN	10	Likely pathogenic	15805158 (2005)
7848	rs121909238	PTEN	10	Pathogenic	15805158 (2005)
64814	rs199620268	TSC1	9	not provided	24728327 (2014)
65285	rs397515209	TSC2	16	not provided	24033266 (2013)
65216	rs201206500	TSC2	16	not provided	
374229	rs1057518991	ADNP	20	Likely pathogenic	
827815	rs1594340060	CHD8	14	Likely pathogenic	
666292	rs1594331875	CHD8	14	Likely pathogenic	
635468	rs1594329885	CHD8	14	Likely pathogenic	
422245	rs1064795655	CHD8	14	Likely pathogenic	
143369	rs267608327	MECP2	X	Pathogenic	
374392	rs756651509	NLGN4X	X	Pathogenic	
39670	rs398123323	PTEN	10	Pathogenic	23757202 (2013)
7827	rs121909227	PTEN	10	Pathogenic	28086757 (2017)
7819	rs121909224	PTEN	10	Pathogenic	29784605 (2018)
7813	rs121909219	PTEN	10	Pathogenic	
620503	rs1565526121	SHANK2	11	Likely pathogenic	
431141	rs1135401811	SCN2A	2	Likely pathogenic	25741868 (2015)
521004	rs1553518509	MBD5	2	Pathogenic	
208759	rs762292772	SHANK3	22	Pathogenic	17173049 (2007)
372707	rs1555910143	SHANK3	22	Pathogenic	
207710	rs796053483	TSC2	16	Likely pathogenic	
Schizophrenia					
210065	rs797045205	ABCA13	7	Likely pathogenic	25741868 (2015)
3521	rs1801131	MTHFR	1	Likely pathogenic	21919968 (2012)

Variation ID	dbSNP ID	Gene	Chromosome	Clinical Significance	Bib. ID/Year
397528	rs1555910162	SHANK3	22	Pathogenic	
4013	rs2870984	PRODH	22	Risk factor	25741868 (2015)
4008	rs2904551	PRODH	22	Risk factor	12217952 (2002)
4006	rs3970559	PRODH	22	Risk factor	24033266 (2013)