UNIVERSITAT DE
BARCELONA

Integrating structural and X-chromosome variants in genetic studies of complex diseases

Daniel Matías Sánchez

Aquesta tesi doctoral està subjecta a la llicència ***Reconeixement- NoComercial – SenseObraDerivada 4.0. Espanya de Creative Commons***.

Esta tesis doctoral está sujeta a la licencia ***Reconocimiento - NoComercial – SinObraDerivada 4.0. España de Creative*** **Commons.**

This doctoral thesis is licensed under the ***Creative Commons Attribution-NonCommercial-NoDerivs 4.0. Spain License.***

Integrating structural and X-chromosome variants in genetic studies of complex diseases

PhD Thesis

Daniel Matías Sánchez
May 2023

Facultat de Biologia, Universitat de Barcelona
Programa de doctorat Biomedicina (HDK05)

Integrating structural and X-chromosome variants in genetic studies of complex diseases

Memòria presentada per Daniel Matías Sánchez per
optar al grau de doctor per la Universitat de Barcelona

Tesi realitzada al

Barcelona Supercomputing Center (BSC)

Doctorand
Daniel Matías Sánchez

Director
David Torrents Arenales

David Torrents

Director
Cecilia Salvoro

Cecilia Salvo

Tutor
Àlex Sànchez Plà

Alex Sanchez

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

LIBERTAS PERFVNDET · OMNIA LVCE
UNIVERSITAT DE BARCELONA

Agradecimientos

Durante el desarrollo, ya no solo de esta tesis, sino también de mi carrera profesional como bioinformático, ha habido tantas personas que podría llenar todas las páginas de esta tesis solo agradeciendo a cada una de ellas todo lo que me han aportado, enseñado y hecho crecer.

En primer lugar, quiero empezar por la persona que me dio la oportunidad de entrar en este mundo: David. Siendo honestos, podríamos decir que mi expediente no era el más adecuado para realizar un doctorado, siempre rondando el típico 7 de media, pero esto a él nunca le ha importado. Aún recuerdo el día que, después de una clase suya en la universidad, me acerqué a preguntarle que requisitos tenía que cumplir para realizar el doctorado en su grupo, y su respuesta fue "Ganas". Actualmente pocos te dan la oportunidad de hacer un doctorado sin un expediente que les garantice una beca, y, en el caso de David, esto no le importó y me dio la posibilidad, no solo de hacer un doctorado, sino de entrar en un mundo profesional que me apasiona.

Las siguientes personas a las que tengo que agradecerles que hoy esté donde estoy son Montse, Jordi e Iván. La primera, por hacer que mi primer contacto con una terminal no fuera una pesadilla, enseñándome desde el día uno todo un mundo de computación y programación y, además, por convertirse en una gran amiga. De Jordi qué decir, nos hemos reído, discutido, peleado, chillado… todo lo que haces con un compañero de proyecto. He aprendido mucho de él durante los dos años que trabajamos juntos y no tengo más que palabras de agradecimiento para él. Por último, pero no menos importante en este bloque, Iván, una persona que hizo que volviera a interesarme por la "odiada" bioestadística que años atrás en la universidad me había hecho pasarlo mal. Siempre dispuesto a ayudar y a aportar ideas, gracias por cada cosa que me has enseñado y por tener con nosotros tanta paciencia. Además, me gustaría incluir a Rafael de Cid y a su grupo por la oportunidad de trabajar con ellos y el buen trato siempre recibido.

Me gustaría destacar en estos agradecimientos a la persona que es, con toda seguridad la más especial y uno de los pilares de mi desarrollo profesional: Cecilia. No solo me dejó formar parte de su proyecto, sino que me acogió y formó desde cero en todo lo que ha podido. Has sido amiga, compañera, profesora y por último codirectora de esta tesis. Nunca sabré como agradecerte que cogieras a un chaval que había programado cuatro cosas y le enseñaras como ser un "científico", a tener un pensamiento crítico, a darle mil vueltas a las cosas, a mirar cada resultado desde distintos puntos de vista y que me trasmitieras la motivación, pasión y amor que tú tienes por la genética. Este párrafo se queda muy corto para de verdad transmitirte todo lo que me has enseñado y ayudado, y si hoy en día tengo las oportunidades profesionales que tengo es en gran parte, gracias a ti.

No me quiero dejar tampoco a cada uno de los miembros del grupo con los que he compartido estos años experiencias tanto científicas como personales. En primer lugar, agradecer a Lorena, con la que he compartido cientos de conversaciones frikis de todo tipo. Gracias por ayudarme siempre que lo he necesitado, eres la nobleza en persona. A Ignasi, por la paciencia que tuvo conmigo y aguantar mis “chapas” en cada comida. A Luisa, por nuestras conversaciones sobre nuestro futuro profesional y las ganas que teníamos de ser ricos y no tener que preocuparnos por nada más. A Álvaro, Miguel, Elias, Mercé y Marta, porque, aunque hemos coincidimos menos tiempo, también habéis formado parte de esta etapa y me llevo un pedacito de cada uno de vosotros.

Este párrafo seguramente es el que más corto se me quede para darle las gracias a una de las personas, no solo más especiales en el tiempo en el que estado realizando esta tesis, sino actualmente de mi vida, Ana. La verdad es que nuestra primera toma de contacto no fue la mejor; no nos dirigimos la palabra prácticamente durante los tres primeros meses, pese a estar sentados a poco más de 5 metros. Nunca me hubiera imaginado que esa persona que tenía pinta de tener mal humor iba a ser una de mis mejores amigas y la persona más importante que me llevo de mi estancia en Barcelona. Nada escrito con palabras puede reflejar lo que eres para mí. Gracias por ser amiga y compañera, por estar en las buenas pero, sobre todo, en las malas, por apoyarme en cada decisión que he tomado y por decirme lo que necesitaba escuchar en cada momento. Porque ni una pandemia mundial, o estar a 400 km, ha hecho que nos distanciemos; lo mejor que me llevo de todo este periodo no es un título de doctor, sino el hecho de haberte conocido.

Tampoco quiero dejarme fuera de estos agradecimientos a personas que me han hecho crecer mucho a nivel personal y profesional. En primer lugar, a Ramón Catalá, por dejarme formar parte de tellmeGen durante los dos últimos años y aceptar que realizara la tesis doctoral en paralelo al trabajo, teniendo comprensión y paciencia y dejándome flexibilidad siempre que la he necesitado para que este trabajo fuera posible. También agradecer a mi “familia” de tellmeGen: Lucia, Blanca, Marta, Marián, Silvia, Raquel, Diego, Rafa, Jesús, Yara y Celia, por cada momento bueno que hemos pasado juntos y por hacer que cada momento de agobio fuera más llevadero con vosotros.

Pero si en la familia tellmeGen hay alguien a la que le debo miles de agradecimientos, es a mi “minion” y una de mis mejores amigas, Judith. Tú aún crees que has aprendido mucho de mí, pero no eres consciente de todo lo que me has enseñado. Gracias por tener la paciencia necesaria para aguantarme en cada momento de estrés y de agobio, por leerte esta tesis y aconsejarme y ayudarme como si fuera tuya, por cada cotilleo, cerveza, comida, cena y viaje que hemos compartido.

A mi familia, que me dieron la oportunidad de estudiar lo que he querido y donde he querido sin hacer preguntas, garantizándome el mejor futuro que han podido y sacrificándose para que nunca me faltara de nada. Gracias por apoyarme en cada una de mis decisiones y aventuras, aunque ello supusiera vivir en la distancia.

Por último, no quiero terminar sin darle las gracias al eje central de mi vida, amiga y pareja, Lucrezia. Nunca podré agradecerte lo que haces y eres para mí. Sé que aguantarme no es fácil, y menos en momentos de estrés como los que estoy pasando ahora mientras escribo estas palabras, pero tú estás ahí siempre, sin una mala cara y siempre dándome cariño y amor para que me sienta lo mejor posible. He hecho que tu vida dé mil vueltas, que cambiemos de ciudad, de comunidad y hasta de país, y siempre has estado ahí apoyando cada una de mis decisiones y sacrificando cosas por mi bien profesional y personal. Nunca podré expresar con palabras lo que eres para mí y lo que te quiero.

Thesis Trajectory

Before diving into the content of the thesis, I would like to initially provide some context regarding the development of this thesis within the computational genomics group at the Barcelona Supercomputing Center (BSC), from both a professional and personal perspective.

When I joined the BSC in April 2018, I began working with Jordi Valls, who was a PhD student in the group at the time, on the GCAT project in collaboration with the principal investigator Rafael del Cid from the Institute for Health Science Research Germans Trias i Pujol (IGTP).

The aim of this project was to generate a haplotype reference panel with a specific focus on the identification of structural variants. Jordi, who had already spent a year in the group, had been working on creating an *in silico* sample with known variants introduced by himself to have an example sample for benchmarking variant callers. He was also researching and pre-selecting the variant callers that we would use in the project, which proved to be a fundamental source of information at the beginning of the project development.

Once I joined the group, we divided the project tasks. In the first phase, he focused on processing and performing the necessary quality control on the whole genome sequencing (WGS) samples from GCAT, while I began working on the execution of the variant callers, their fine-tuning, and the benchmarking study to determine which programs were best suited for the research. Following this, Jordi focused on running the different variant callers that I had selected on all the GCAT samples. In parallel, Iván Galván, who was a biostatistician at IGTP at the time, joined the project, and he and I worked together on the development and application of various logistic regression models for filtering high-quality variants in the creation of the reference panel. In the third phase of this project, Jordi and Iván worked on evaluating the imputation performance of the reference panel created, while I focused on validation using both experimental information and comparisons of our variant set with other available databases.

Although during the different project phases, each of the three team members focused on a specific part, we all directly collaborated in each other's tasks and provided support, making it a collaborative effort in the end. As a result, this project culminated in the publication of an article in the journal Nucleic Acid Research (NAR), in which all three of us are listed as first authors or co-authors[1]. Finally, this project, along with the development of the *in silico* sample, is the central core of Jordi Valls' thesis, so part of the information presented about this project in this thesis has been previously presented.

The second phase of my thesis involved working with PhD Cecilia Salvoro on what we called the “X chromosome project”. The aim of this project was to recover the role of the X chromosome on complex diseases, something that had been neglected until now due to computational limitations. In this project, our main strategy was to use case-control cohorts obtained from the database of genotypes and phenotypes (dbGaP) that covered a wide range of phenotypes, and perform X chromosome association studies (XWAS), analyzing males and females separately. Later, we decided to expand our study by also including the UK Biobank.

During the first phase of the project, I worked on phenotype curation, quality control, phasing, imputation, and association of all dbGaP cohorts and phenotypes, while Cecilia worked on the UK Biobank. It should be noted that in these processes, we used the GCAT reference panel, which allowed us to introduce variants that would not have been possible to analyze without this panel.

In the second phase, my focus was on calculating a significance threshold for the study and determining which loci were significant in the project, as well as comparing them with loci described in previous analysis of the UK Biobank.

The third phase of the project was marked by various personal changes for both me and Cecilia. In July 2021, I joined a company as a full-time bioinformatician, while Cecilia also left the BSC and joined an external company, causing the final phase of the project to slow down as we both worked part-time on it. During this phase, Miguel Pérez, now PhD student at the group, joined the project to provide support. In this last phase, we worked on conducting a preliminary functional analysis of the signals obtained and tried to determine significant functional differences between males and females.

Currently, we are working to finish writing the paper of the project, while there is an open analysis that consists of studying the impact of heterozygous genotypes on association analyses in women. Cecilia and I initiated this analysis, which is now being carried out by Miguel Perez and will be part of his future thesis, while I am providing him with support.

In conclusion, the two projects I have worked on during these five years are large-scale projects in which various individuals have been involved, and I would like to thank them for their efforts. Without their contribution, these two projects and this thesis could not have been completed.

Abstract

In recent years, the genetics field has placed a significant emphasis on identifying and characterizing genetic factors contributing to complex diseases, alongside environmental factors. Genome-wide association studies (GWAS) have emerged as one of the principal methodologies for this purpose, as they analyze extensive genetic and phenotypic data from multiple individuals to identify genetic variations associated with specific traits. This approach has advanced our understanding of the genetic architecture of complex diseases, allowing the development of prevention strategies and genetic risk estimation. However, despite progress, much information remains to be uncovered, leading to a heritability discrepancy, which refers to the difference between heritability estimated in population studies and that explained by known genetic variations.

Many methodological and statistical limitations are slowing down the identification of the genetic variation associated with the risk to develop complex diseases. Current GWAS rely on Single Nucleotide Polymorphisms (SNP) arrays that have a limited number of variants. To overcome this, the number of variants analyzed can be augmented through imputation of pre-existing genetic variants from reference panels. However, reference panels frequently exclude rare variants and structural variants (SVs) which results in these variants not being considered in the imputation process leading to potential missed associations.

Another element neglected in most studies of complex diseases is the X chromosome, which is one of the two sex chromosomes and has unique biology that results in different copy number in females and males. When examining the SNP-trait associations reported in the National Human Genome Research Institute's (NHGRI) GWAS catalog, a clear shortfall in the representation of the X chromosome becomes apparent. Still, only 0.5% of the known associations map on chromosome X. This under-representation is primarily due to the methodological challenges associated with its analysis. The unique pattern of inheritance and the effects of allelic inactivation in females can result in allelic imbalances between the sexes and decrease the statistical power during genetic association studies.

In this thesis, we aim to address these challenges by creating a comprehensive genetic resource, consisting of a haplotype map, particularly enriched in well characterized, and phased SVs; and deal with the gap in X-chromosome analysis by designing, implementing and applying a targeted methodology for the study of the role of the X-chromosome across multiple phenotypes.

The haplotype map was generated using 785 Illumina high coverage (30x) whole-genomes from the Iberian GCAT Cohort with multiple variant

identification methods and Logistic Regression Models (LRMs) for their validation. The resulting catalog includes 35,431,441 variants, including 89,178 SVs (≥50 bp), 30,325,064 SNVs and 5,017,199 indels, across all individuals in the cohort. The haplotype panel demonstrates improved imputation capabilities, with 14,360,728 SNVs/indels and 23,179 SVs being imputed, representing a 2.7-fold increase in SVs compared to other available genetic variation panels. This panel's significance is highlighted by the imputation of a rare Alu element located in a new locus associated with Mononeuritis of the lower limb, a rare neuromuscular disease. This study represents the first in-depth characterization of genetic variation in the Iberian population and the first haplotype panel that systematically includes SVs in genome-wide genetic studies.

The X-Chromosome targeted strategy was designed and applied to nearly 800,000 individuals across 600 phenotypes from publicly available cohorts (UK Biobank and dbGaP). This pipeline includes the data collection process, a specific and fundamental quality control for the X-chromosome analysis and the phasing, imputation and association process, which was performed by splitting females and males and then meta-analyzing the results, thus allowing to detect sex-differences.

Our analysis of nearly 500,000 X-linked variants, including SVs, resulted in 96 significant associations with 77 traits, with 75 of these being novel. By incorporating sex-specific analyses, we identified 41 loci with different behavior between males and females. These findings give us insight into the level of missing information and the X chromosome's potential role in complex diseases, as well as its contribution to sex-specific risk and manifestation.

In conclusion, this work highlights the importance of considering SVs and the chromosome X in genetic studies, particularly in the context of exploring the genetic architecture of human complex diseases. The findings offer a valuable asset for further examination of the genetic components that contribute to complex diseases, marking a progression towards a more complete comprehension of the genetic landscape and its effects on human health.

Table of contents

Abbreviations and acronyms

1000G	The 1000 Genomes Project
AC	Allelic count
AS	Assembly strategy
BP	Base pairs
BAM	Binary Alignment Map
BSC	Barcelona Supercomputing Center
BQSR	Base Quality Scores
BWA	Burrows-Wheeler Aligner
CGH	Compative Genomic Hybridization
CHR	Chromosome
CIGAR	Concise Idiosyncratic Gapped Alignment Report
CDCV	Common disease-common variant
CNV	Copy Number Variant
dbGaP	Database of Genotypes and Phenotypes
DGV	Database of Genomic Variants
DNA	Deoxyribonucleic acid
EGA	European Genome-phenome Archive
ExAC	Exome Aggregation Consortium
FDR	False Discovery Rate
FN	False Negative
FP	False Positive
GA4GH	Global Alliance for Genomics and Health
GIAB	Genome in a bottle
GL	Genotype Likelihood
GMT	Gene Matrix Transposed
GMMAT	Generalized linear Mixed Model Association Test
GNOMAD	Genome Aggregation Database

GoNL	Genome of the Netherlands
GRM	Genetic Relationship Matrix
GWAMA	Genome-Wide Association Meta-analysis
GWAS	Genome-Wide Association Study
HGDP	Human Genome Diversity Panel
HGP	Human Genome Project
HGSVC	Human Genome Structural Variant Consortium
HRC	Haplotype Reference Consortium
HWE	Hardy-Weinberg Equilibrium
ICD	International Classification of Disease
ID	Identifiers
IBD	Identity by Descent
INDEL	Small Insertion and deletion
LD	Linkage Disequilibrium
LMM	Logistic Mixed Models
LOCO	Leave-one-chromosome-out
LRM	Logistic Regression Model
M	Million
MAC	Minor Allelic Count
MAF	Minor Allele Frequency
MEI	Mobile Element Insertion
NCBI	National Center for Biotechnology Information's
NGS	Next-generation Sequencing
NH	Non-Homologous
NHGRI	National Human Genome Research Institute
NIH	National Institute of Health
OR	Odd Ratio
PAR	Pseudoatosomal regions
PC	Principal Component
PCA	Principal Component Analysis
PIR	Phasing Informative Read

POPRES	Population Reference Sample Project
PRS	Polygenic Risk Scores
QC	Quality Control
ROC	Receiver Operating Characteristic
SAIGE	Scalable and Accurate Implementation of Generalized Mixed Model Software
SBS	Sequencing by Synthesis
SNP	Single Nucleotide Polymorphisms
SD	Standard Deviation
SR	Split Read
SV	Structural Variants
TP	True positive
UCSC	Univeristy of California Santa Cruz
UK	United Kingdom
VCF	Variant Calling File
WGS	Whole Genome Sequencing
WTCCC	Wellcome Trust Case Control Consortium
XCI	X Chromosome Inactivation
XIST	X-inactive specific transcript
XWAS	Chromosome X Wide Association Study

Introduction

1 Brief History of Genetics

Genetics, as defined by the National Institute of Health (NIH), (https://www.nih.gov/), is the study of heredity and variation of inherited traits. It is a scientific discipline that deals with the mechanisms of inheritance and the way traits are passed down from one generation to the next. Genetics encompasses a wide range of topics, including the structure and function of genes, the organization of the genetic material within cells, the mechanisms of mutation and genetic variation, and the study of complex traits that are influenced by multiple genes and environmental factors.

The history of genetics can be divided into three major stages: from its beginnings until the early 20th century, when the first hypotheses on heredity and evolution emerged; the second half of the 20th century, known as the DNA era, when the structure of DNA was discovered; and the 21st century, from the development of sequencing technology and the completion of the human genome project to the present day (Figure 1).

1.1 From the origins to the mid-20th century

The study of genetics can be traced back to the ancient Greeks, who were the first to propose the concept of heredity. They believed that traits were inherited and determined by the mixture of fluids or “humors” in the body[2].

In the 19th century, the science of genetics began to develop. Gregor Mendel (1822-1884) published his experiments on pea plants in 1866, establishing the laws of inheritance and laying the foundation for the science of genetics[2]. He explained how traits were passed from one generation to another through what we now know as a genotype, and how genetic material could create new variations. His work demonstrated that genes are the fundamental units of inheritance, providing the basis for the modern understanding of how characteristics are passed down from one generation to the next[3]. However, the “Mendelian Laws of Inheritance” were disregarded for several decades due to limited understanding of the mechanisms of heredity and the presence of alternative inheritance theories[4].

Charles Darwin (1809-1882) also greatly influenced the field of genetics with his theory of evolution in 1859. He showed that species evolve over time through natural selection, where advantageous traits are passed on to offspring and disadvantageous traits are eliminated[5]. Darwin also proposed the theory of pangenesis in his book “The Variation of Animals and Plants Under Domestication”. This theory suggested that every part of an organism produced small particles called “gemmules” that were believed to be units of inheritance and carried in the blood to the reproductive organs[6].

Over the next few decades, the history of genetics saw several key developments and discoveries where researchers made significant progress in understanding the mechanisms of heredity. One key event was in 1900, where Carl Correns (1864-1933), Hugo de Vries (1848-1935), and Erik Tschermak (1871-1962) independently rediscovered Mendel's work[7]. This led to a renewed interest in Mendel's theories and helped to lay the foundation for the development of the science of genetics. Between 1903 and 1905, Walter Sutton (1877-1916) and Theodor Boveri (1862-1915) proposed the idea that genes were located on chromosomes and confirmed their role in heredity[8]. In addition, in the early 1900s, Thomas Hunt Morgan (1856-1945) conducted experiments on *Drosophila melanogaster* to study the inheritance of traits. He discovered that some traits were linked, meaning that they were always inherited together[9]. He also found that the frequency of recombination, or the exchange of genetic material between chromosomes, was a measure of the distance between two linked genes[10]. Also, in 1908, Godfrey Harold Hardy (1877-1947) and Wilhelm Weinberg (1862-1937) hypothesized that the frequency of alleles in a population will remain constant from one generation to the next in the absence of evolutionary forces, what is known as “Hardy-Weinberg Equilibrium” (HWE)[9].

Finally, the concept of “population genetics” as we know it today was born at same time as the field of genetics was taking shape. It was first introduced by the mathematician and statistician Ronald A. Fisher (1890-1962) as a way to integrate evolutionary biology and genetics. In his work, Fisher applied quantitative methods to the study of evolution and genetics, and he showed how the principles of Mendelian inheritance could be used to understand the evolution of populations over time[11,12]. Population genetics has been a central component of genetics research since its inception, and has had a major impact on our understanding of the genetic basis of diseases, the evolution of species, and the conservation of biodiversity.

1.2 Second half of the 20th century: The DNA era

While the first decades of the 20th century saw great advances in the field of genetics, the second half of the century is marked by the discovery of DNA (deoxyribonucleic acid). For this reason, this period is known as the “DNA era”.

The “DNA era” conventionally starts in the 1950s when scientists made significant breakthroughs in understanding the structure and function of DNA. This period was marked by a rapid increase in knowledge and technological advancements that transformed the field of molecular biology and led to the development of new techniques for studying the structure and function of DNA having.

The discovery of DNA, however, occurred before this time. It was made by Swiss physician and biologist Friedrich Miescher (1844-1895) in 1869. Miescher was studying the composition of white blood cells and discovered a new type of nucleic acid that he called “nuclein”[13]. Later, it was determined that this nucleic acid was DNA. In parallel, between 1884 and 1885, Albrecht von Kölliker (1817-1905), Oscar Hertwig (1849-1922), August Weismann (1834- 1914) and Eduard Strasburger (1844-1912) evidenced that the cell nucleus contains the basis for inheritance[8]. But it was not until 1944 that Oswald Avery (1877-1955), Colin MacLeod (1909-1972), and Maclyn McCarty (1911-2005) demonstrated that a genetic transformation in bacteria could be accomplished by exposing them to purified DNA. Their work showed that DNA was the genetic material responsible for passing on traits from one generation to the next[8].

The key event that marked the beginning of the "DNA era" was the discovery of the double-helix structure of DNA. In 1952, Rosalind Franklin (1920-1958) and Maurice Wilkins (1916-2004) used X-ray crystallography to determine the structure of DNA. Then, James Watson (1928-) and Francis Crick (1916-2004) used Franklin and Wilkins' X-ray diffraction data, along with other data, to build a model of the structure of DNA. They found that the two chains of nucleotides were held together by hydrogen bonds between the nitrogenous bases and that the bases were paired in a specific way: adenine with thymine and cytosine with guanine[14].

Shortly after, in 1961, Marshall W. Nirenberg (1927-2010) and J. Heinrich Matthaei (1929-) deciphered the genetic code, figuring out how the sequence of DNA nucleotides translates into the sequence of amino acids in a protein. This was done through a series of experiments involving decoding the codons[15]. This knowledge allowed the discovery and application of recombinant DNA technologies in the 1970s[16].

1.3 The sequencing era

The discovery of DNA structure in the 1950s set the stage for the field of DNA sequencing. Sequencing is the process of reading and determining the order of nucleotides or bases in a DNA molecule.

The first entire genome sequence, the bacteriophage Φ-X174 genome, was determined in the late 1970s by Frederick Sanger (1918-2013) using the "Sanger Method" which involved a chemically interruption of DNA chain synthesis at specific points and then separating the fragments by size[17], marking a significant milestone in the field of genomics. As sequencing technologies advanced and the potential applications of genomics became more apparent, an ambitious goal emerged: to sequence the entire human genome. This led to the establishment of the Human Genome Project (HGP).

The Human Genome Project (HGP), launched in 1990, aimed to sequence the entire human genome using a combination of the Sanger method and newer, high-throughput sequencing approaches such as the "shotgun sequencing" method. The first draft of the human genome was published in 2001[16,18] and was completed in 2003. This landmark achievement made it possible to study the human genome at a level of detail never possible before.

Parallel to the HGP, scientists recognized the need for a more efficient and cost-effective way to identify genetic variants in the human genome. This led to the development of genotyping arrays, also known as DNA chips or microarrays. The idea was to use these arrays to detect single nucleotide polymorphisms (SNPs)[19].

In the late 2000s and early 2010s, the advent of next-generation sequencing (NGS) technologies revolutionized DNA sequencing. NGS allows for the simultaneous sequencing of millions of DNA fragments, making it possible to sequence entire genomes in a matter of days, rather than years.

The completion of the human genome project, together with the emergence of new technology, led to the emergence of other international projects such as HapMap, which aims to study human genetic variation[20] or ENCODE, which aims to identify the functional elements of the human genome[21].

Today, DNA sequencing continues to advance at a rapid pace with new technologies and applications being developed all the time. For example, long-read sequencing technologies have been developed allowing DNA molecules of up to 100,000 base pairs (bp) to be sequenced and enabling the precise determination of genomic structures changes such as the arrangement of chromosomes[22].

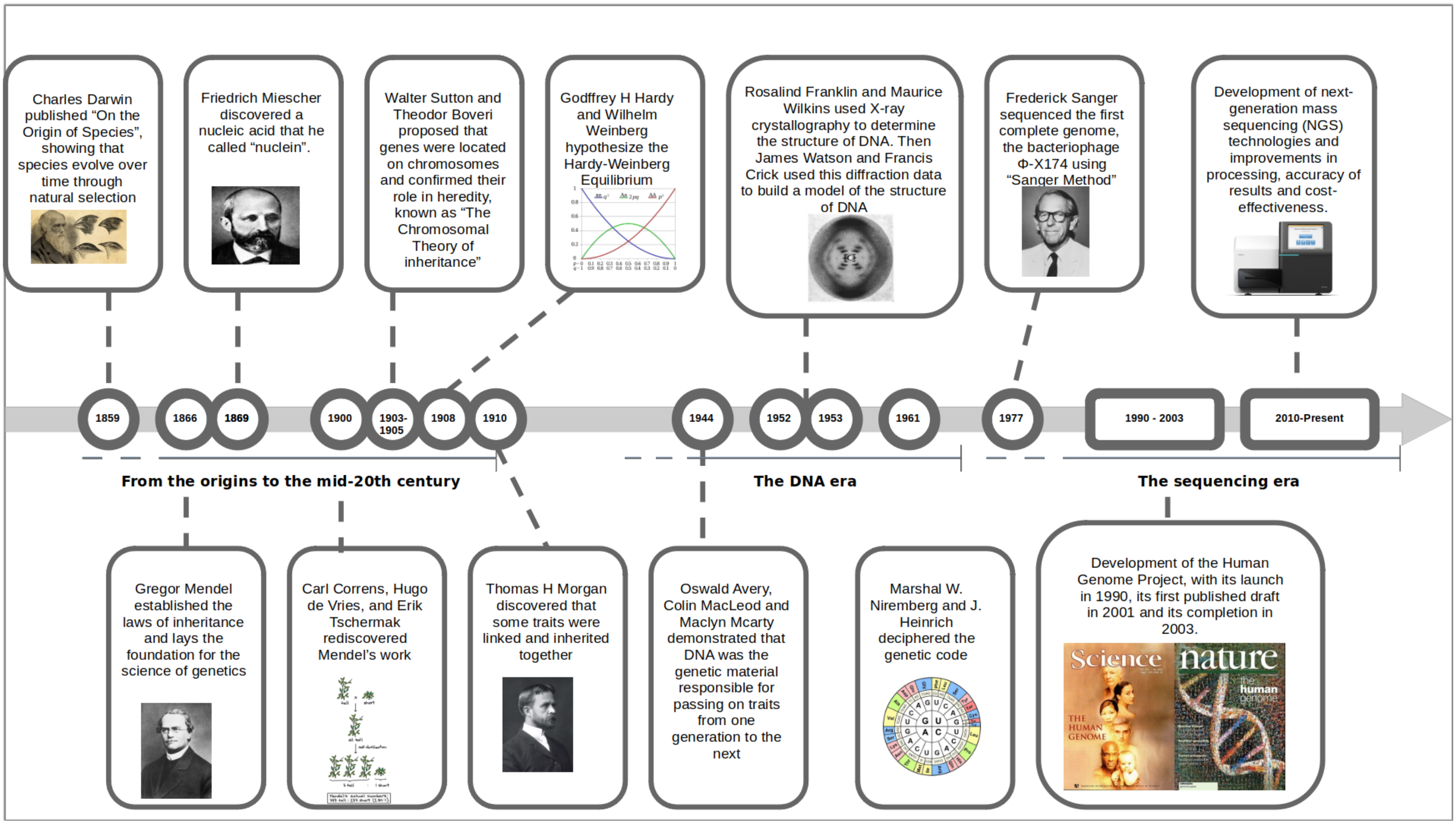

Charles Darwin published "On the Origin of Species", showing that species evolve over time through natural selection
Friedrich Miescher discovered a nucleic acid that he called "nuclein".
Walter Sutton and Theodor Boveri proposed that genes were located on chromosomes and confirmed their role in heredity, known as "The Chromosomal Theory of inheritance"
Godffrey H Hardy and Wilhelm Weinberg hypothesize the Hardy-Weinberg Equilibrium
Rosalind Franklin and Maurice Wilkins used X-ray crystallography to determine the structure of DNA. Then James Watson and Francis Crick used this diffraction data to build a model of the structure of DNA
Frederick Sanger sequenced the first complete genome, the bacteriophage Φ-X174 using "Sanger Method"
Development of next-generation mass sequencing (NGS) technologies and improvements in processing, accuracy of results and cost-effectiveness.
1859
1866
1869
1900
1903-1905
1908
1910
1944
1952
1953
1961
1977
1990 - 2003
2010-Present
From the origins to the mid-20th century
The DNA era
The sequencing era
Gregor Mendel established the laws of inheritance and lays the foundation for the science of genetics
Carl Correns, Hugo de Vries, and Erik Tschermak rediscovered Mendel's work
Thomas H Morgan discovered that some traits were linked and inherited together
Oswald Avery, Colin MacLeod and Maclyn Mcarty demonstrated that DNA was the genetic material responsible for passing on traits from one generation to the next
Marshal W. Nirenberg and J. Heinrich deciphered the genetic code
Development of the Human Genome Project, with its launch in 1990, its first published draft in 2001 and its completion in 2003.
Science
nature
THE HUMAN GENOME
the human genome

Figure 1. Genetics historical overview. The history of genetics starts with the first theories of inheritance and finishes with the current scenario and next-generation sequencing technologies.

2 Basis of genetics

2.1 Definition and types of genetic variants

Genetic information is stored in DNA molecules. In the case of humans, the genome contains a sequence of approximately 3 billion nucleotides, organized into 23 chromosome pairs. However, not all the physical positions in the genome are identical among all individuals, there are differences that give rise to genetic variation. Genetic variants refer to the different versions or forms in the DNA sequence between individuals in a population. The latest estimations from the 1000G-Phase3 release mentioned that a human genome typically differs from the human reference sequence in around 5 M positions[23]. While, in 2017, it was published that more than 644 million high-quality polymorphisms from sequenced human genomes are known[24].

Depending on their nature or origin, we can make a distinction between germline variants and somatic variants. Germline variants refer to genetic variants present in the parent’s germline cells, and therefore present in any cell of the offspring. In contrast, somatic variants are genetic variations that occur in an individual's non-reproductive cells and are not present in all the cells of an individual. These somatic variants are not passed on to offspring (Figure 2a). In this thesis, I will always refer to germline variants when talking about genetic variants.

There is still no consensus on the criteria to classify genetic variants, however, one of the most commonly used is based on size (Figure 2b). Using this classification, we can differentiate between (i) Single nucleotide variants (SNVs), single base pair differences in the DNA sequence; (ii) Small insertions and deletions (INDELs), differences in the DNA sequence caused by the insertion or deletion of one or more base; and (iii) Structural variants (SVs), large-scale differences in the DNA structure, including inversions, deletions, duplications, translocations and insertions[25]. Even within this classification, there is debate about the size thresholds that define a variant as one type or another. For example, in the 1000 Genomes Project, any variant between 2-50 base pairs (bp) was considered an INDEL, and any variant larger than 50 bp was considered a SV[26]. However, in later projects such as gnomAD[27], the size windows were modified, defining INDELs as variants between 2-30 bp and SVs as variants larger than 150 bp. This change introduced the term "mid-size INDEL" to encompass variants between 30-150 bp. While the size thresholds for differentiating between various types of genetic variants continue to be a subject of debate, for the purposes of this thesis, we will classify any variant larger than 50bp as SV.

The frequency with which these genetic variants are present within a population is another way of classifying genetic variants (Figure 2c). Each of these variants is located at a unique position on the genome, known as a locus.

At each locus, a variant can have different versions, termed as alleles. We can distinguish the major allele, which is the most prevalent in a population, a the minor allele, the second most common. With this, variants can be classified according to the frequency of the minor allele (Minor Allele Frequency, MAF) in common variants (MAF > 5%), low-frequency variants (1% ≤ MAF < 5%) and rare variants (MAF < 1%)[26,28–30].

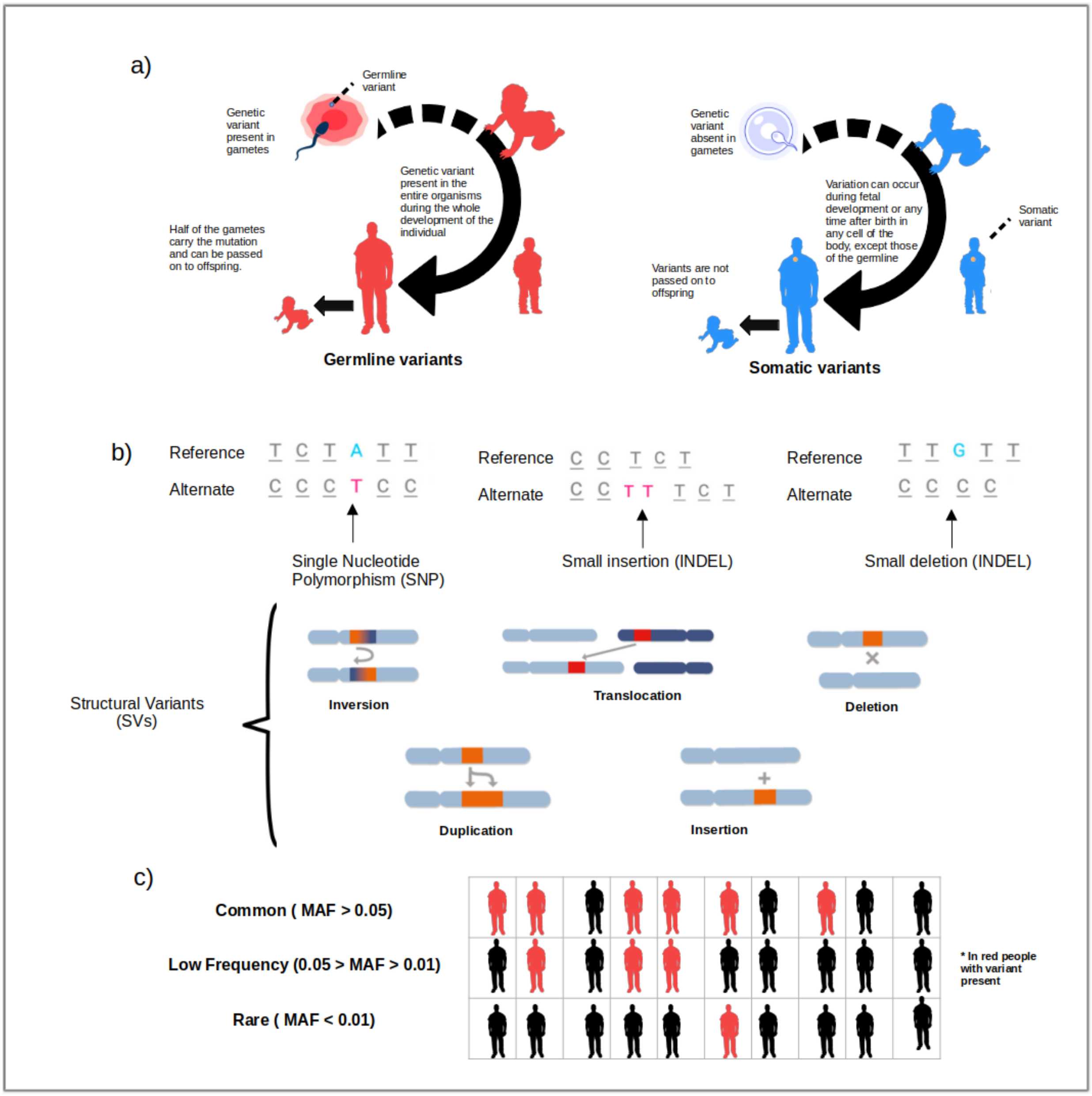

a)
Germline variant
Genetic variant present in gametes
Genetic variant present in the entire organisms during the whole development of the individual
Half of the gametes carry the mutation and can be passed on to offspring.
Germline variants
Genetic variant absent in gametes
Variation can occur during fetal development or any time after birth in any cell of the body, except those of the germline
Somatic variant
Variants are not passed on to offspring
Somatic variants
b)
Reference T C T A T T
Alternate C C C T C C
Single Nucleotide Polymorphism (SNP)
Reference C C T C T
Alternate C C T T T C T
Small insertion (INDEL)
Reference T T G T T
Alternate C C C C
Small deletion (INDEL)
Structural Variants (SVs)
Inversion
Translocation
Deletion
Duplication
Insertion
c)
Common (MAF > 0.05)
Low Frequency (0.05 > MAF > 0.01)
Rare (MAF < 0.01)
* In red people with variant present

Figure 2. Types of genetic variants depending on different features. a) Genetic variation can be classified based on their origin into germline or somatic. Germline variants are present in an individual's germ cells and can be passed on to their offspring. Somatic variants occur in an individual's non-reproductive cells and are not passed on to the offspring. **b)** Genetic variation can be classified according to the size of the variant into SNVs, affecting only one nucleotide; INDELs, insertions or deletions smaller than 50bp; and SVs, large-scale differences in the DNA structure, including inversions, deletions, duplications, and insertions. **c)** Genetic variants can be classified according to their frequency in the population into common, low-frequency and rare variants.

2.2 Linkage Disequilibrium

Although millions of variants have been described in the human genome, they do not provide unique and independent information. This is primarily due to the non-uniform nature of recombination. Recombination occurs during meiosis when homologous chromosomes, which carry identical sets of genes, pair up and exchange genetic material. This process results in the shuffling of alleles at the same locus. However, recombination is not uniformly distributed across the genome; there are regions, known as recombination hotspots, where recombination occurs at a higher frequency[31].

This non-uniform recombination leads to the phenomenon of linkage disequilibrium (LD), which helps explain the observed correlation patterns in human genetic variation[32]. In essence, LD is a correlation that quantifies the degree of association between alleles in a population and, occurs when the frequency of a specific combination of alleles at multiple loci is higher than what would be expected. On top of this, distance plays a critical role since the closer the loci are on a chromosome, the more likely they are to be inherited together.

A set of alleles at different loci in the same chromosome that are inherited together is defined as a haplotype. Because of the nature of recombination described above, haplotype blocks, also named LD blocks, can be identified in human populations[33,34].

3 Methods and technologies to identify human genetic variability

Advancements in technology have greatly facilitated the identification of human genetic variability, which is crucial in comprehending the genetic basis of human phenotypes, including complex diseases. Over the last several decades, the methods for recognizing and categorizing genome variability have improved with regards to accuracy, sensitivity, and cost-effectiveness. The identification of genetic variants involves reading the genome or portions of it and, the use of various analysis techniques on the readings obtained. Although there are multiple technologies for genome reading, two stand out as the most widely utilized: genotyping arrays and sequencing technologies.

3.1 Genotyping arrays

Genotyping arrays marked a new era in the field of genetics, providing researchers with a powerful and efficient tool for analyzing large numbers of genetic variants in a population of individuals.

Genotyping arrays are designed to detect specific variants in a individual’s DNA. This is done by hybridizing a sample of DNA with a set of probe sequences, which are anchored to a solid surface, typically a glass slide or a chip. The probes are designed to bind to specific locations in the genome and detect variants of interest, mainly single-nucelotide polymorphisms (SNPs).

Genotyping arrays have become an important tool in the field of genetics and genomics because they provide a cost-effective and efficient way to genotype large numbers of individuals. This makes it possible to study large populations, including those from diverse ethnic backgrounds, as they could easily provide a bunch of genetic markers evenly distributed in the genome. Additionally, the arrays can be used to genotype individuals in a variety of different studies, including population genetics, association studies, family-based studies, and case-control studies[35].

Advances in genotyping array technology have made it possible to genotype entire genomes, including a larger number of variants to be analyzed, and providing an even more comprehensive understanding of an individual's genetic information. However, despite their widespread use, there are a number of limitations associated with this technology that should be considered when using genotyping arrays in research studies.

One of the main limitations of genotyping arrays is the accuracy of the results. The accuracy of the results depends on the quality of the DNA sample and the specificity of the probes used in the assay. If the DNA sample is of low quality or has been degraded, the results of the genotyping assay may be incorrect or incomplete. Additionally, if the probes used in the assay are not specific enough, the results may not accurately reflect the true genotype of the individual. This can lead to false positive or false negative results, which can have serious implications for genetic association studies and other types of research. Nevertheless, in recent years, there has been a significant advancement in genotyping technologies. This progress has been marked by a considerable improvement in the accuracy of genetic variant identification, with current methods approaching a remarkable accuracy rate.

Another limitation of genotyping arrays is the coverage of the genome. While genotyping arrays are capable of genotyping large numbers of individuals, they typically only cover a subset of pre-defined variants, around 1 million SNPs. This is a much smaller number of SNPs compared to the total number of genetic variants present in the human genome. Furthermore, due to their reliance on predetermined alleles, which exhibit different degrees of polymorphism among various ancestries, genotyping arrays cannot be applied universally across all populations.

A third limitation of genotyping arrays is the variability in performance between different platforms and manufacturers, being Illumina, with its Infinium technology[36], and Thermo Fisher Scientific with Affymetrix[37], the market leaders. This can lead to differences between arrays, including differences in accuracy, sensitivity, and specificity. This variability in performance can have serious implications for genetic association studies, where results from different arrays may not be comparable[38].

Finally, genotyping arrays are not capable of detecting all types of genetic variants. For example, they are not capable of detecting SVs, such as deletions, duplications, and insertions, which can have a significant impact on gene function and disease susceptibility[39]. In order to detect these types of genetic variants, researchers must use other technologies, such as whole genome sequencing or targeted sequencing.

3.2 Sequencing technologies

The information obtained from sequencing is critical for understanding the genetic code and can be used to study the human genome and the genomes of other species. The process of sequencing is performed on DNA fragments and as a result, sequence reads are obtained, which are sequences of DNA letters obtained from a single pass of a sequencing machine. The size of the sequenced fragments is determined by the type of machine and sequencing technologies used. There are various sequencing technologies available, each with its unique strengths and weaknesses.

The choice of technology depends on factors such as the specific research question, the size of the genome being studied, and the amount of sequencing that needs to be performed. Currently, three generations of sequencing technologies can be distinguished (Figure 3).

3.2.1 First-Generation Sequencing technologies

First-Generation Sequencing, also known as Sanger sequencing, is one of the first and most widely used sequencing technologies. It is a chain-termination method, in which DNA is replicated in the presence of a specific set of nucleotides that are modified to prevent the extension of the growing DNA strand. This results in a series of DNA fragments with different lengths, which can be separated by electrophoresis and visualized by autoradiography (now replaced by fluorescence). Sanger sequencing is still widely used, especially for sequencing smaller regions of DNA[17].

3.2.2 Second-Generation Sequencing technologies

As technology has advanced, new sequencing techniques have been developed, giving rise to what is known as NGS technologies, a concept first introduced in 2005[40], giving rise to second-generation sequencing. Second-generation sequencing, also known as high-throughput sequencing, revolutionized the field of genomics by making it possible to sequence large amounts of DNA sequences at a much faster pace and lower cost compared to first-generation sequencing. These methods typically involve the simultaneous sequencing of many DNA fragments in parallel, which results in massive amounts of data[41]. While it took years to sequence a genome with first-generation sequencing technologies, more recent second-generation technologies can do so in less than 24 hours. Additionally, NGS technologies allow for the sequencing of complete genomes or exomes.

Second-generation sequencing is currently the most widely used, mainly due to its advantages over first generation. Illumina sequencing, also known as Illumina high-throughput sequencing, is the most popular second-generation technology that uses a process called bridge amplification to generate clusters of DNA fragments on a surface[42]. The DNA fragments are then sequenced by reading the fluorescent signals generated by the incorporation of fluorescent labeled nucleotides. Illumina sequencing is highly scalable and cost-effective, making it a popular choice for large-scale genome sequencing projects such as the 1000 Genomes Project[29]. However, second-generation sequencing technologies have some limitations, manly related with the read length. While NGS can generate a vast amount of data, the individual sequences, or reads, are relatively short. This makes it challenging to assemble complex genomes or map reads back to repetitive regions of the genome.

3.2.3 Third Generation Sequencing technologies

Third-Generation Sequencing technologies refer to a new generation of DNA sequencing methods that aim to overcome the limitations of the previous first- and second-generation sequencing technologies[43]. The most prominent technologies include Oxford Nanopore technologies and PacBio sequencing, which are also known as long-read sequencing technologies.

Pacific Biosciences sequencing, also known as single-molecule real-time sequencing, is a NGS technology where each DNA molecule is attached to a DNA polymerase enzyme, and this complex is then placed into a tiny well, known as zero-mode waveguides, to detect the fluorescence signals generated by the incorporation of nucleotides[22]. Pacific Biosciences sequencing provides longer read lengths than second-generation sequencing technologies, allowing for the detection of SVs and improved genome assembly.

Oxford Nanopore sequencing, also known as real-time nanopore sequencing, is a NGS technology that uses a nanopore to detect the electrical signals generated by DNA molecules as they pass through at pore[44]. Oxford Nanopore sequencing provides long read lengths and the ability to sequence DNA in real-time, making it a popular choice for the sequencing of complex genomes and the detection of SVs. However, it is currently more expensive and less scalable than other NGS technologies.

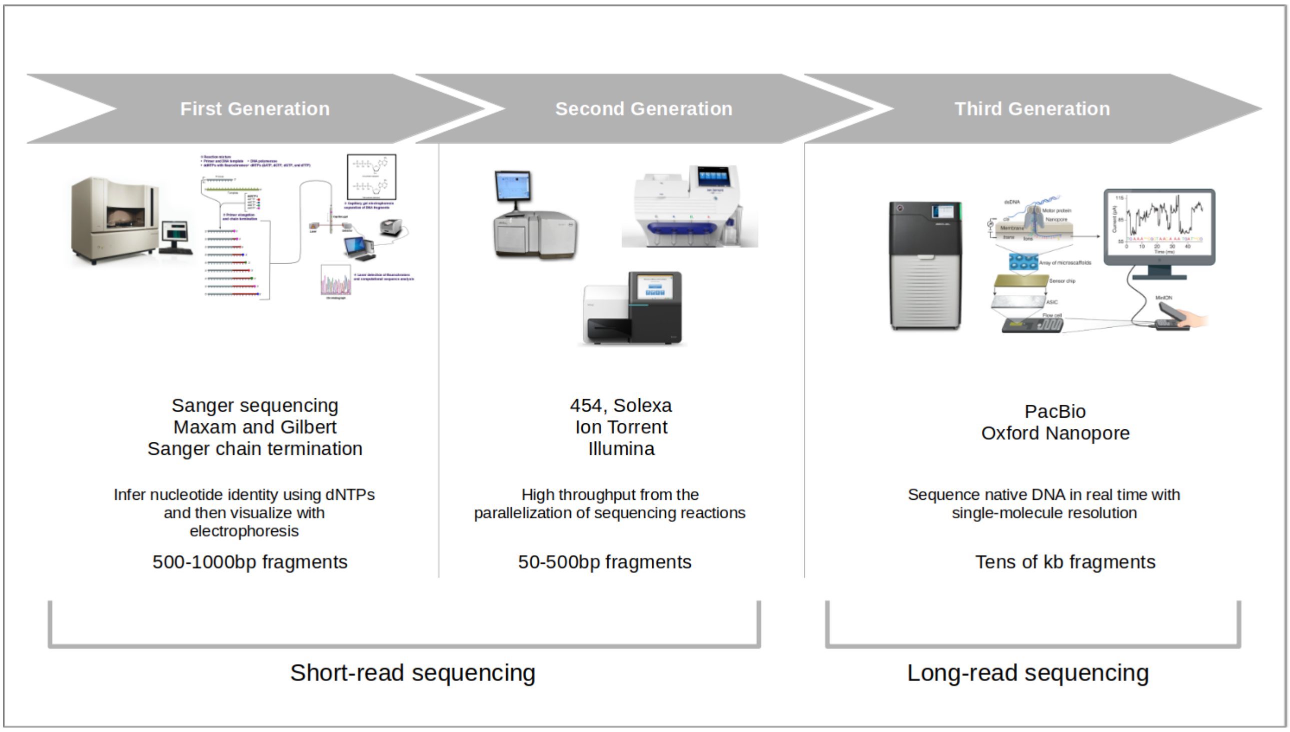


Figure 3. Sequencing technologies. Sequencing technologies can be divided into 3 main generations based on the technology used. The first generation corresponds to the first technologies based on Sanger's discoveries. The second generation, also known as next-generation sequencing, allowed massive sequencing in a shorter space of time and at a lower cost. The third generation, known as long-read generation, made it possible to sequence fragments of tens of thousands of kb for the first time.

3.3 From sequencing to analysis: BAM files

Sequencing technologies generate large amounts of genomic data by reading the genome and recording the information in files. These files are referred to as FASTQ files, which are a complex file format that store raw sequencing data. Each FASTQ file contains four lines of information for each read, including the read identifier and description, the actual nucleotide sequence, a symbol indicating the start of the quality scores, and the quality scores for each nucleotide in the sequence. The nucleotide sequence is made up of the four building blocks of DNA (A, T, C, and G), and the quality scores measure the accuracy of the base calling for each nucleotide. The quality scores are usually recorded in Phred format, a logarithmic scale that reflects the accuracy of the nucleotide base calls[45].

FASTQ files, therefore, contain short reads with no information about where in the genome they are located. To deal with this, a process called alignment is necessary. This process aims to match the different reads with their correct location using a reference genome. The reference genome is a curated version of the human genome, based on the DNA of a few individuals, and is considered the standard against which all other genomes are compared. When a sample genome is sequenced, millions of DNA reads are generated. This alignment results in information that is stored in Binary Alignment Map files (BAMs), which contain information about the location and quality of each aligned read[46]. There are different software tools available for performing this alignment process, with the Burrows-Wheeler Aligner (BWA)[47] being the most widely used (Figure 4a).

Alignment algorithms need to address multiple issues, among which sequence complexity is a significant factor. The difficulty in aligning reads arises when the reference genome lacks proper annotation or if there is high genomic complexity in a region, such as repetitive sequences or regions with high C-G content. In addition, the inability to align reads in regions missing from the reference genome, such as gaps or SVs, can also result in alignment errors[48]. Finally, the alignment of short reads onto sex chromosomes is a challenge due to the high similarity between the X and Y chromosomes, producing technical artifacts and affecting downstream analyses[49].

Despite its limitations, the process of alignment and generation of BAM files has allowed the development of methodologies that allow the detection of genetic variants by comparing the genetic information of an individual and a reference genome. These methodologies are known as variant calling methodologies.

3.4 Methodologies to detect and genotype genetic variants in NGS

Detection of genetic variants from NGS data is crucial as it provides valuable information about an individual's genome. The process of detecting genetic variants can be divided into two parts: (i) identifying which variants exist and where they are located within an individual's genome, known as variant calling, and (ii) determining the number of alleles affected for each variant, known as its genotype, through the process of genotype calling. (Figure 4b).

3.4.1 Variant calling process

Variant calling is the process of identifying changes (variants) between a sample genome and a reference genome. This process starts from the data obtained in the read alignment process. In this context, coverage is an important factor in the variant calling step because it directly affects the accuracy and sensitivity of variant detection. Coverage refers to the number of times a given nucleotide in a genome is sequenced during the sequencing process. In other words, it represents the depth of sequencing data that has been obtained for a particular region of the genome. High coverage provides more confident and accurate identification of variants, while low coverage can lead to false-negative results or lower confidence in variant calls. Once the reads are aligned, a quality control step is performed to filter out low-quality data and ensure the accuracy of the results. This can involve removing reads that are contaminated with adapter sequences, reads that contain poor-quality nucleotides, and reads that have a low mapping quality score.

Once these steps have been completed, the variant calling algorithms are applied to identify the variants present in the sample genome. The accuracy of detecting genetic variants is directly influenced by the characteristics of the variant, from its type to its size or position in the genome; or characteristics of the sequencing process, such as coverage, read size or insert size[50,51].

Currently, there are more than 150 variant calling algorithms, which can be classified according to the type of variants they detect, creating two main groups; algorithms for detecting SNVs and INDELs; and algorithms for detecting SVs[52].

The detection of SNVs and INDELs is a crucial aspect of genomic analysis. SNVs are the most prevalent type of variants in human with a median of ~3.3-4 million SNVs per genome, followed by INDELs, with numbers between 492 and 850 thousand per genome[53,54]. Most of these variants are likely neutral or have small effects, which do not result in noticeable changes in diseases. However, a minority of these variants can have clinically relevant impacts[55].

Variant callers for SNV detection are highly accurate, thanks to the small size of SNVs and ease of read mapping. On the other hand, the detection of INDELs is more challenging due to low concordance between sequencing platforms, alignment errors in repetitive regions, differences in INDEL size, and variations between variant callers[56]. In this context, The Global Alliance for Genomics and Health (GA4GH) has developed a pipeline to standardize the SNV and INDEL representations, known as normalization, allowing the comparison of outputs from different variant callers[57].

Several variant callers are available for the detection of SNVs and INDELs, including Haplotype Caller[58], FreeBayes[59], Platypus[60], VarScan2[61], Strelka2[62], or Deepvariant[63]. These tools use different approaches, such as assembly strategies (AS), deep learning, or Split Reads strategies (SR), to detect genetic variants. Nonetheless, the ability to identify variants is not uniform across all variant callers, as they exhibit disparities in precision, recall, and the range of detectable variant sizes. Despite such inconsistencies, the accuracy of SNVs and INDELs variant callers is generally high, leading to accurate identification of these variant types in the current state.

SVs, on other hand, involve larger changes in the genomic structure and typically have a higher impact on human phenotypes. These changes can take the form of deletions, duplications, insertions, or rearrangements of genomic segments, which can range in size from a few kilobases to several megabases. They can impact gene expression, topologically associating domains, and protein-coding genes, leading to altered gene function and causing rare or complex diseases[64]. Despite their importance, until recently, the analysis of SVs has been neglected, mainly due to technical challenges.

The advent of second-generation sequencing methods improved SV detection, but its accuracy remained limited due to short reads and the low coverage sequencing methods used[65,66]. Nevertheless, different projects emerged with the aim of describing and identifying SVs within a human genome. The most noteworthy are 1000G-Phase3[26], which identified around 2,500 SVs per genome using samples with low coverage (3x), and the recent gnomAD-SV project, which using genomes with higher coverage (30x) identified more than 7000 SVs per individual[67]. These types of studies have increased our knowledge of SVs and their impact, leading to their incorporation into important databases such as dbVar[26] or Database of Genomic Variants[68] (DGV), including about 19 and 36 million variants respectively.

More recently, the use of third-generation sequencing techniques, which employ longer reads, has greatly improved the detection and study of SVs, with the Human Genome Structural Variant Consortium (HGSVC) standing out[69,70]. However, due to the high cost of these techniques, their systematic use is still unfeasible. This has led to efforts being focused on improving the detection of SVs through the design and improvement of different variant callers and the increase of coverage used during the second-generation sequencing process, which is currently more economically accessible.

Currently, the field of SVs calling does not have a widely accepted standard, as no single variant caller can accurately detect all types and sizes of variants[51,71]. In response to this challenge, various variant callers have been developed in the scientific community to detect specific types of SVs, including deletions, duplications, insertions, inversions, translocations, and transposable elements, by utilizing paired-end read and alignment information[72].

Some of these software tools, such as Lumpy[73] or Wham[74], allow for the detection of a range of SV types, while others, such as PopIns[75] or MELT[76], are more specialized, focusing on the detection of large insertions or transpositions, respectively. Additionally, the detection strategy used by a variant caller can greatly impact its performance, accuracy, and computational requirements. For example, Delly2[77] combines split-read, discordant-read, and read-depth strategies, leading to a need for high computational resources. On the other hand, variant callers as Manta[78] use only assembly strategies, which is known to be one of the most accurate methods, resulting in a lower computational burden than multi-strategy variant callers.

3.4.2 Genotype calling process

Variant calling and genotype calling are closely related but distinct steps in the analysis of genomic data. While variant calling aims to identify positions with differences from a reference sequence, genotype calling is the process of determining the genotype for each variant that has already been called in the variant calling step[79]. An accurate genotyping is crucial for enhancing the understanding of genetic variability in a population, as it provides a more accurate representation of the variant allele frequency, allowing to increase the performance of genetic variability reference panels[80].

The accuracy of genotyping depends directly on the coverage of the sequenced sample on which the calling process is performed[81]. Generally, in samples with low coverage (<20x), probabilistic approximations are used to infer the genotype of each of the variants, obtaining a genotyping likelihood.

Genotype likelihood (GL) is a numerical representation of the probability of observing a particular genotype given the sequence data. It provides a measure of confidence in the assigned genotype based on the sequencing quality, coverage, and alignment of the data. Genotype likelihoods are often used in probabilistic genotype calling frameworks, where they are combined with prior information, such as allele frequencies and LD patterns, to determine the most likely genotype for each individual. The resulting genotype calls are typically expressed as a probability score, reflecting the degree of confidence in the assignment.

In genotype calling in samples with high coverage (>20x), the number of reads observed for each individual is counted and fixed cutoffs are applied. For example, a Phred-type quality score of Q20 can be used and, if the proportion of the non-reference reads falls between 20% and 80%, a heterozygous genotype is called, otherwise a homozygous genotype. This method is widely used and effective when sequencing depth is high[79].

Nowadays, most software that performs variant calling is also capable of performing genotyping. However, as in the variant calling, the strategies and accuracies obtained between SNVs or INDELs and SVs differ dramatically.

In the case of SNVs and INDELs, genotyping accuracy is high and consistent with the variant calling process. Additionally, some tools provide the option to perform a joint re-genotyping of the entire sample set after individual genotyping. This re-genotyping, based on LD patterns, can correct any errors from the initial genotyping, enhancing accuracy even further.

In contrast, variant calling of SVs is challenging, often due to poor resolution in detecting the breakpoint position of the event or variation during the variant calling. This leads to lower precision in genotype calling by variant calling software[82]. To overcome this, tools specifically designed for genotyping SVs have been developed. These tools commonly use both the output of the variant calling (the vcf file) and alignment (BAM), providing more information to improve genotype estimation. Examples of these tools are BayesTyper[83] and SVJed[84].

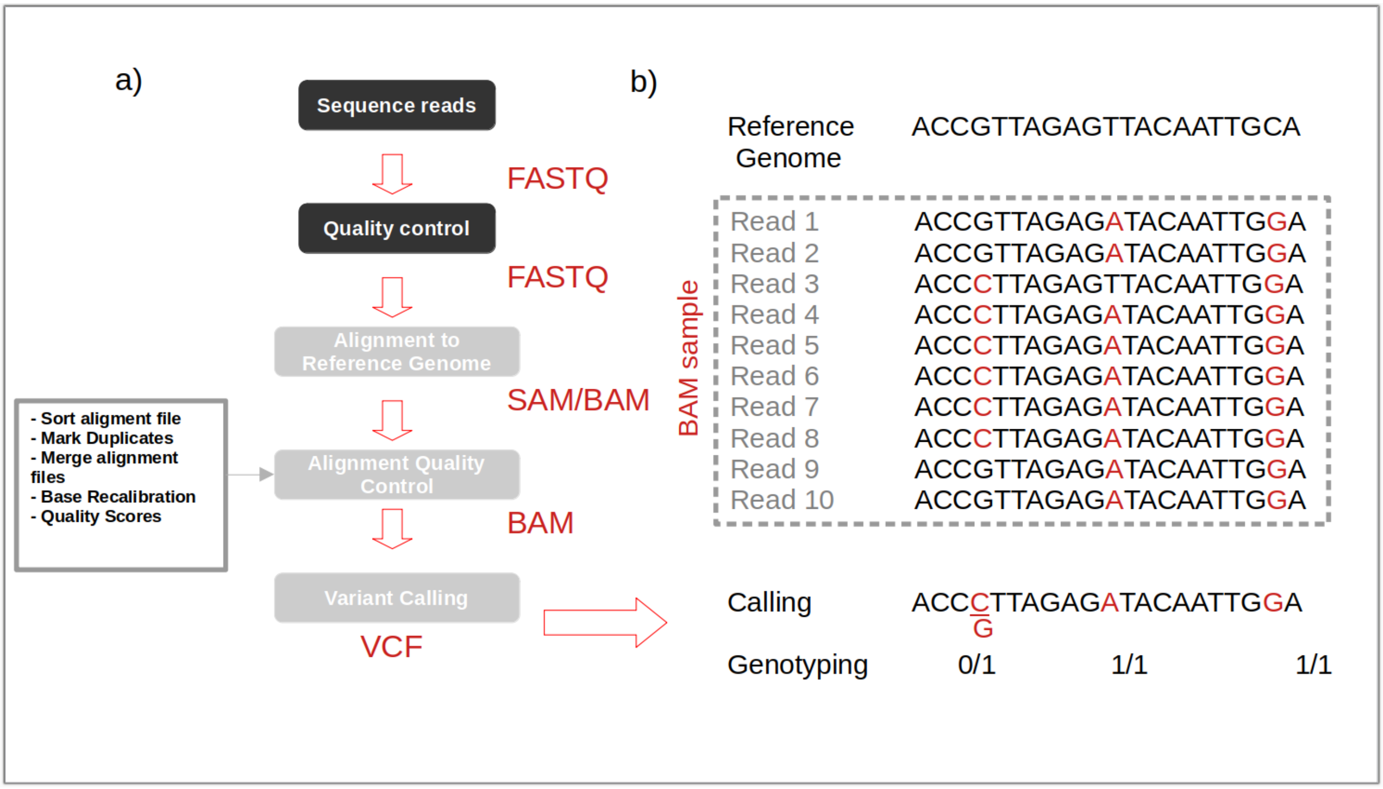

a)
Sequence reads
FASTQ
Quality control
FASTQ
Alignment to Reference Genome
SAM/BAM
- Sort aligment file
- Mark Duplicates
- Merge alignment files
- Base Recalibration
- Quality Scores
Alignment Quality Control
BAM
Variant Calling
VCF
b)
Reference Genome
ACCGTTAGAGTTACAATTGCA
BAM sample
Read 1 ACCGTTAGAGATACAATTGGA
Read 2 ACCGTTAGAGATACAATTGGA
Read 3 ACCCTTAGAGTTACAATTGGA
Read 4 ACCCTTAGAGATACAATTGGA
Read 5 ACCCTTAGAGATACAATTGGA
Read 6 ACCCTTAGAGATACAATTGGA
Read 7 ACCCTTAGAGATACAATTGGA
Read 8 ACCCTTAGAGATACAATTGGA
Read 9 ACCGTTAGAGATACAATTGGA
Read 10 ACCGTTAGAGATACAATTGGA
Calling ACCCTTAGAGATACAATTGGA
G
Genotyping 0/1 1/1 1/1

Figure 4. Pipeline for variant detection in whole-genome sequencing samples. a) From the sequencing stage, a result from a sample goes through different checks and steps to produce a Binary Alignment Map file (BAM), including a post-sequencing quality control on the generated reads, an alignment of the reads with a reference genome and a post-alignment quality control. **b)** Once the BAM file is generated, variant calling is performed. In this process, the reads present in the BAM are compared against a reference genome to determine the variants present in the sample. Finally, the ratio of reads with and without the variant is estimated to determine the genotype of the variant. The result is variant calling file (VCF) containing all variants present in the sample.

3.5 Main projects studying human genetic variability and population genetics

The study of population genetics has undergone major advances thanks to new sequencing and genetic analysis technologies. This has led to the emergence of important initiatives and projects building genetic variability reference panels, collections of genetic data from individuals representing a set or a particular population. These panels are used as a reference for understanding patterns of genetic variation within and between populations.

The first project who aimed at characterizing human genetic variation on a global scale was the HapMap Project (2002-2010). It involved the genotyping of over 1 million single nucleotide variants in over 270 individuals from four different populations. The results of the project were used to develop genotyping arrays that have become widely used in genetic association studies[20].

In 2002, the Human Genome Diversity Panel (HGDP), aimed to characterize the genetic diversity of human populations around the world by sequencing the genomes of over 1,000 individuals from 52 different populations. The results of the project have provided insight into the genetic diversity of populations and the evolutionary history of human populations[85].

In 2008, the 1000 Genomes Project aimed to create a high-resolution map of human genetic variation by sequencing the genomes of over 2,500 individuals from 26 different populations. The project generated a wealth of data that has been used to understand the patterns of genetic variation in different populations and the genetic basis of disease[26].

In 2014, the Exome Aggregation Consortium (ExAC) aimed to aggregate and harmonize exome sequencing data from over 60,000 individuals to create a comprehensive reference dataset for human genetic variation[86].

As sequencing technologies advanced and whole-genome sequencing became more cost-effective, the Genome Aggregation Database (gnomAD) project was established in 2016 including and built on top of ExAC. GnomAD expanded the scope to include both whole-exome and whole-genome sequencing data, providing a more comprehensive view of human genetic variation. This project contains data from over 140,000 individuals including information on over 20 million genetic variants, including SNVs, copy number variants (CNVs), and SVs. This resource provides a rich dataset for studying the distribution and impact of genetic variation in different populations, including the role that rare variants play in human disease[27].

4 Characterization of inheritance of human traits

Human inherited diseases exhibit a wide range of heterogeneity, but certain historical classifications have been established to provide a theoretical framework for studying them. This knowledge requires a deep understanding of both, heritability and genetic architecture of the particular trait.

4.1 Contribution of genetics to traits: the concept of heritability

The understanding of inheritance of traits and diseases was furthered by the realization that relatives tend to be more alike in their traits than randomly selected individuals from the population. This observation led to the creation of the concept of heritability[87].

Heritability is a quantitative estimate of the degree to which the variation in a trait is due to genetic variance, and it is expressed as a value between 0 and 1. A heritability estimate of 1 means that 100% of the variation in a trait is due to genetic factors, while a heritability estimates of 0 means that the trait is not influenced by genetics at all[87,88].

The heritability of a trait can be estimated through various studies. One widely used method is twin studies, which compare the similarities and differences of traits between identical twins (who share same genetic information) and fraternal twins (who share only half of the genetic information). Another method is family studies, which examine the inheritance of traits within families. Additionally, heritability can be analyzed through population design studies. These studies, which use a population of unrelated individuals to capture only the proportion of phenotypic variance explained by genetic variants, can help avoid inflated heritability estimates due to environmental factors shared between related individuals[89,90].

It is important to note that heritability estimates are population-specific, meaning that they only apply to the specific population that was studied. For example, a heritability estimate for a particular trait in one population may not be the same as the heritability estimate for that same trait in a different population[87]. This is because the frequency of certain genes and environmental factors can vary between populations and can influence the heritability of a trait. Heritability estimates are also influenced by the age of the individuals being studied. For example, the heritability of a trait may be different in children compared to adults, as environmental factors can play a more important role in the development of a trait during early life[91]. In addition, heritability may also vary between the sexes, with the heritability of a trait being different in males and females[92].

4.2 Genetic architecture of human diseases and traits

The study of human genetics and the connection between genotype and phenotype has been a challenge for the scientific community. This has given rise to efforts to understand the genetic architecture of human diseases. The genetic architecture of human traits refers to the number, type, frequencies and effect sizes of genetic variants that determine the trait[93].

One of the key components of the genetic architecture of human diseases and traits are the inheritance patterns. Traits and diseases can be classified into two main different categories based on their genetic architecture, monogenic or Mendelian diseases and complex diseases. (Figure 5).

4.2.1 Monogenic diseases

Monogenic diseases are diseases where mutations, typically rare, in a single gene are enough to cause the diseases[94]. They are also called Mendelian diseases as they are passed from generation to generation in a predictable manner, following the laws of inheritance described by Mendel. These diseases follow different inheritance models, including autosomal dominant, recessive or co-dominant. In these diseases, a single variant may be enough to produce a pathological phenotype[94].

Studies of monogenic diseases have shown that the variants involved in the development of these traits have a substantial effect size, with penetrance playing an important role in disease expression. In these cases, the environment often plays a minor role. Over the past few decades, research efforts have focused on monogenic diseases and have resulted in significant advances in understanding the genetic basis of these conditions. Some of the most significant milestones include the discovery of the *BRCA2* gene, which is associated with a significantly increased risk of breast and ovarian cancer[95], or the discovery of the Huntington gene, which is associated with Huntington's disease[96,97].

NGS techniques have enabled the acceleration of gene identification in monogenic phenotypes, leading to an increase in the pace of disease gene discovery. As a result, 2,937 genes were reported for 4,163 monogenic phenotypes between 2005 and 2014. However, the underlying causal genes for approximately 50% of all known monogenic disorders are still unknown[98]. Despite that, the study of monogenic diseases has greatly enriched our understanding of gene function, regulation, human phenotypes, and physiology. This has led to the development of new treatments and early diagnostics for diseases[86,98].

4.2.2 Complex diseases

Complex diseases, also known as multi-factorial diseases, result from a combination of multiple genetic and environmental factors, that can increase or reduce the risk of developing the disease[31]. Examples of complex diseases include heart disease, diabetes, asthma, Alzheimer's disease, and some forms of cancer.

The genetic basis of complex diseases are not well understood. Unlike monogenic diseases, in which mutations in a single gene are enough to cause the disease, complex diseases are the result of a combination of numerous low-penetrant variants that arise from multiple loci, which, together with the environment, collectively contribute to the susceptibility to the disease[99].

The model that was initially widely accepted for years to explain the genetics of complex diseases is based on the common disease-common variant (CDCV) hypothesis. This hypothesis was built upon Fisher's infinitesimal rationale, which posits that multiple genetic variants with a minor allele frequency of at least 1-5%, have individual, modest effects on disease susceptibility. However, when considered collectively, these variants can impart a substantial risk for the manifestation of a complex phenotype[16,99,100].

Nevertheless, despite the efforts of the scientific community to study complex diseases and the impact of the cooperation of multiple common variants, a big fraction of the heritability for most common diseases remains unknown[101]. This led to the emergence of alternative hypotheses to the CDCV hypothesis. The most notably was the common disease-rare variant (CDRV) hypothesis, which suggests that complex diseases are caused by a large number of rare genetic variants with larger effect sizes, which individually have a small contribution to disease risk but collectively have a significant impact[102]. The debate over the relative contributions of common and rare genetic variants to complex diseases has been a longstanding one, with the CDCV and RVCD hypotheses representing the two main viewpoints. However, it is now widely accepted that both common and rare genetic variants play a role in the genetic basis of complex diseases.

Additionally, other theories propose that other factors may also be involved in the development and cause of complex diseases. One of these theories is the epistatic hypothesis, which suggests that complex diseases result from the interactions between multiple genes, rather than the effect of a single gene, leading to disease susceptibility[99,103,104].

Due to their chronic nature, complex diseases, as defined by the World Health Organization, are a critical public health concern and the leading cause of mortality worldwide, compromising economies and life quality. Furthermore, as populations age, the prevalence of complex diseases tends to increase, which in turn can strain healthcare systems and lead to increased healthcare costs[105]. In this context, comprehending the interplay of genetic and environmental components in their manifestation is imperative for the advancement of preventive measures and therapeutic approaches. Despite the substantial progress made in understanding the genetics of complex diseases, much more research is needed to fill the gap in our knowledge about how genetics contribute to their development.

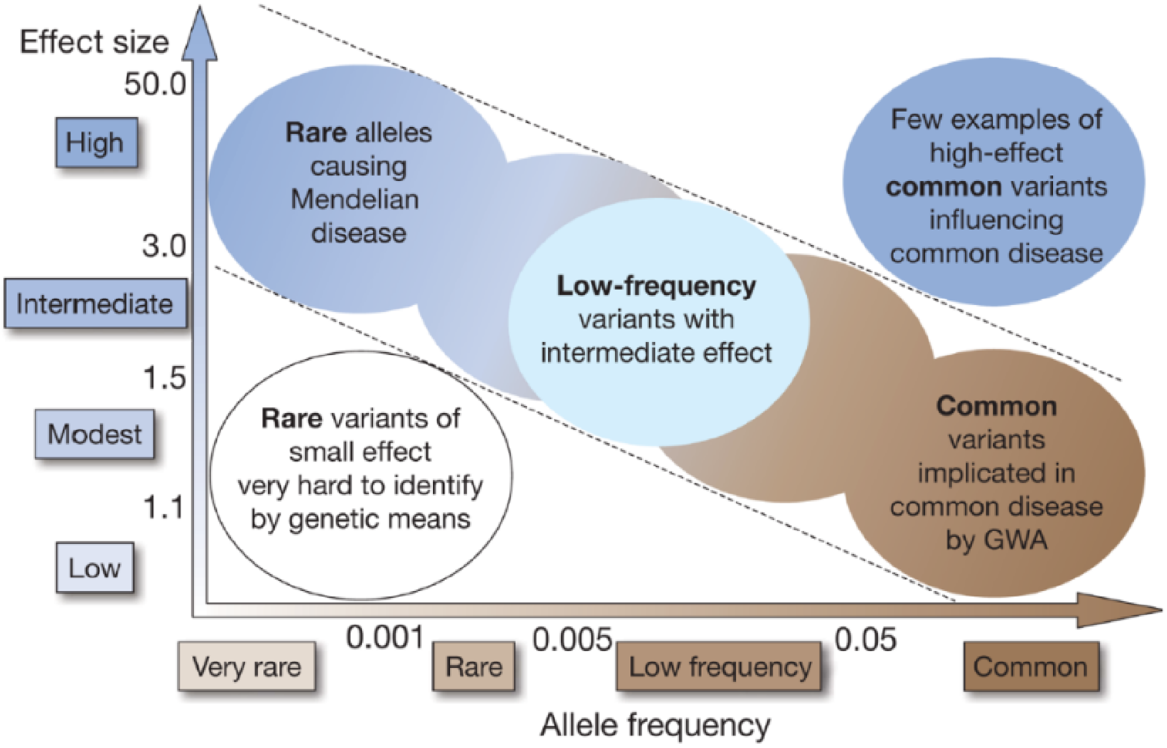

Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

Figure 5. Genetic architecture of rare and complex diseases.
In rare/Mendelian diseases large effect sizes are expected. For complex diseases, the common disease/common variant hypothesis have been accepted. GWA = genome-wide association. Figure from Manolio et al.[99]

5 Genome wide association studies

The advancement in the use of family tree studies has greatly improved our understanding of the inheritance patterns of rare diseases. By 2003, over 1,400 genes responsible for monogenic diseases had been identified[106]. However, this progress has not been reflected in complex traits. Unlike monogenic diseases, only 8 regions with small effect sizes were linked to complex diseases through linkage studies in 2003[107].

As previously discussed, monogenic diseases are characterized by high penetrance, where the variants associated with the disease have a strong impact on its development. Unlike monogenic diseases, complex diseases do not exhibit clear inheritance patterns within families. These complex diseases demonstrate a polygenic architecture that complicates their investigation and comprehension. This complexity poses significant challenges for genomics experts, as they endeavor to unravel the underlying genetic factors and their role in the emergence of these multifaceted conditions.

This indicates that family-based linkage analyses were not appropriate for complex diseases, highlighting the need for new methodologies to study complex diseases more effectively. This resulted in the development of population-based association studies, including the genome-wide association study (GWAS). The first GWAS was conducted as part of the HapMap project in 2005[108], but it was not until the publication of the Wellcome Trust Case Control Consortium (WTCCC) that this technique was widely accepted and began to be used extensively.

5.1 GWAS definition

GWAS are a statistical and computational approach used in genetics and genomics to identify genetic variants associated with a specific trait or disease. This method involves the simultaneous analysis of hundreds of thousands or millions of genetic markers across the entire genome in large sample sizes of individuals, typically using a case-control design where individuals with the trait or disease of interest are compared to those without it (Figure 6)[109].

The main goal of GWAS is to identify specific genomic regions, known as loci, through the analysis of different genetic markers which demonstrate a statistical correlation with the trait or disease being examined. For binary traits, this association typically means that a particular variant has a different frequency between cases and controls, suggesting a connection between the locus and the trait under investigation.

These loci contain causal variants, genetic variations that directly influence the trait or disease. However, pinpointing the causal variants can be challenging due to LD, which, as previously explained, is a phenomenon where genetic variants are inherited together more often than would be expected by chance. This complicates the identification of the specific variant responsible for the observed association, making it difficult to discern which one is the true causal variant inside the locus.

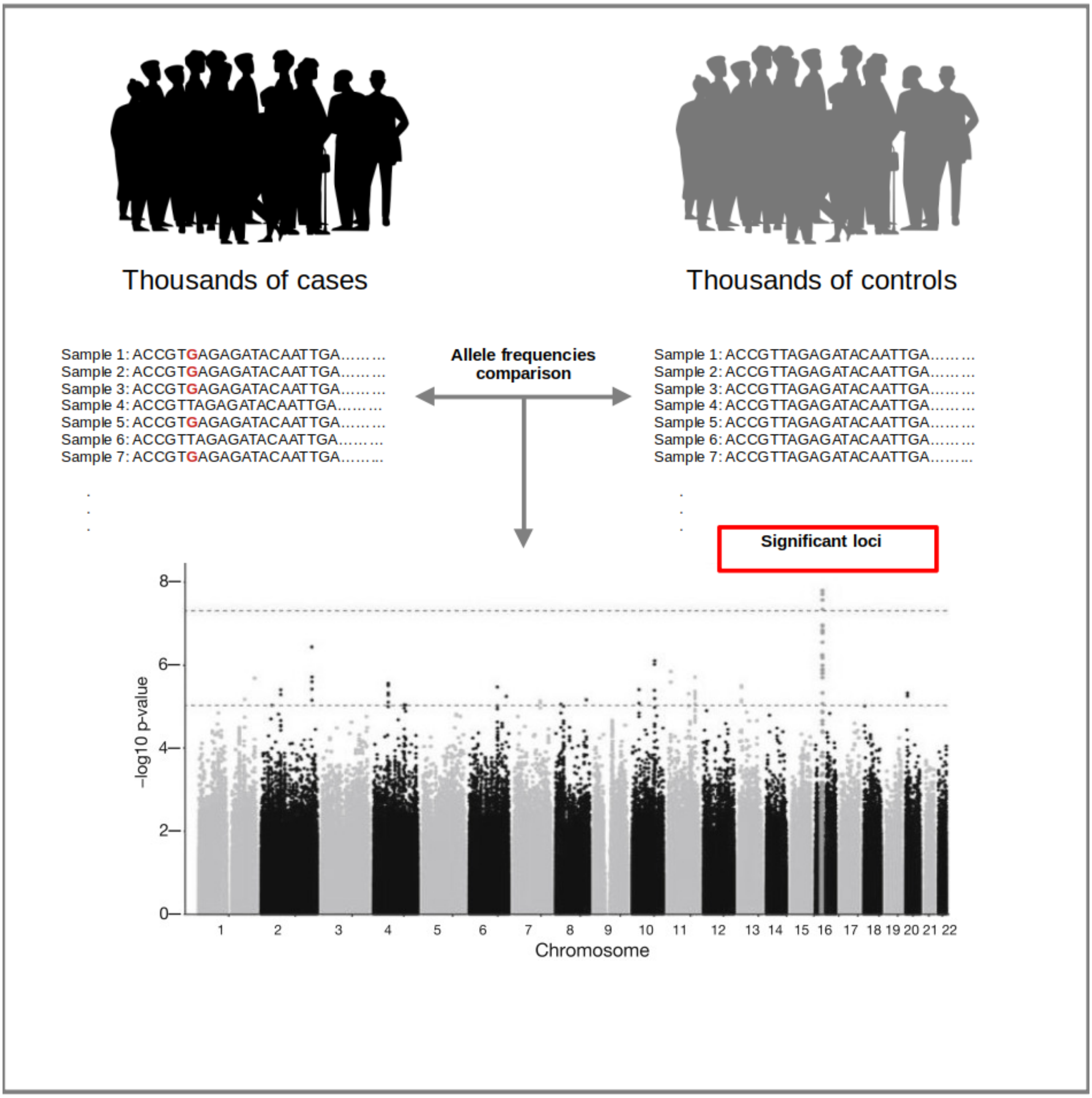

Thousands of cases
Thousands of controls
Sample 1: ACCGTGAGAGATACAATTGA.........
Sample 2: ACCGTGAGAGATACAATTGA.........
Sample 3: ACCGTGAGAGATACAATTGA.........
Sample 4: ACCGTTAGAGATACAATTGA.........
Sample 5: ACCGTGAGAGATACAATTGA.........
Sample 6: ACCGTTAGAGATACAATTGA.........
Sample 7: ACCGTGAGAGATACAATTGA.........
Allele frequencies comparison
Sample 1: ACCGTTAGAGATACAATTGA.........
Sample 2: ACCGTTAGAGATACAATTGA.........
Sample 3: ACCGTTAGAGATACAATTGA.........
Sample 4: ACCGTTAGAGATACAATTGA.........
Sample 5: ACCGTTAGAGATACAATTGA.........
Sample 6: ACCGTTAGAGATACAATTGA.........
Sample 7: ACCGTTAGAGATACAATTGA.........
Significant loci
−log10 p-value
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome

Figure 6. Basis for genome-wide association studies. GWAS test hundreds of thousands of genetic variant to find those with allele frequencies discrepancies between a group of cases and controls. The results are plot usually in a Manhattan plot, displaying the -log10 p-value in the y-axis and the genome in the x-axis. All the variants above the p-value threshold (in general p-value < 5x10^{-8} for genome-wide analysis) are considered statistically associated with the disease.

5.2 Statistical methods for GWAS

Statistical methods, such as regression analysis, are fundamental to the analysis and interpretation of GWAS association results. In the analysis of binary traits, such as diseases, logistic regression has traditionally been applied as the standard method as they allows for the incorporation of multiple co-variates and confounding factors.

In recent years, logistic mixed models (LMMs) have emerged as an advanced statistical approach in the context of GWAS, particularly when analyzing binary traits while accounting for population stratification and relatedness among study participants. LMMs are an extension of logistic regression, incorporating both fixed effects and random effects to model the relationship between genetic variants and binary traits. The fixed effects typically capture the main genetic effects, while the random effects account for confounding factors such as population stratification or relatedness within the study population. By incorporating both fixed and random effects, LMMs allow for a more accurate estimation of the association between genetic variants and binary traits, while also mitigating the risk of false positives due to unobserved confounding factors[110]. The application of tools that use LMMs, such as SAIGE[111], is of particular interest, especially in the analysis of biobanks, where confounding factors like relatedness can have a significant impact[110].

In the association analysis, one of the obtained results is the p-value, a statistical measure that helps to determine whether the association observed between a genetic variant and a trait is statically significant. A p-value is the probability of obtaining the observed data if the null hypothesis is true. In the case of a GWAS, the null hypothesis is typically that there is no association between the genetic variant and the trait or disease. A small p-value suggests that the association is unlikely to have occurred by chance, indicating a statistically significant finding[112].

Another result obtained in a GWAS is the effect size of each variant, which is expressed as the odds ratio (OR)[99]. An OR greater than 1 indicates that the variant is associated with an increased risk of disease or trait occurrence, while an OR less than 1 indicates that the variant is associated with a decreased risk of disease or trait occurrence[113].

One of the principal challenges inherent in the use of these statistical methods in the association analysis step is the issue of multiple testing. Given that thousands or even millions of genetic markers are examined, the probability of encountering false positive associations increases substantially. To account for this, stringent significance thresholds are imposed. However, this strict threshold may inadvertently result in false negatives, whereby true associations are overlooked[114].

One of the most popular correction methods for multiple testing is the Bonferroni correction. This method tries to keep the experimental-wise error rate (αe) at a nominal level by adjusting the point-wise error rate (αp). In the case of Bonferroni, correction gives αp = αe/N, where N is the number of independent tests[115].

In 2005, The HapMap project estimated that the number of common independent variants were 150 each 500kb[23]. Considering that the whole genome has around 3.3 gigabase pair, it suggested a total of around 100.000 independent tests resulting in a p-value threshold of $5x10^{-8}$ [116]. This value was subsequently established as the standard significance threshold for GWAS, where any variant with a p-value of less than $5x10^{-8}$ was considered to be associated with genome-wide significance[117].

5.3 The importance of sample size in GWAS: repositories and biobanks

The initial step in a GWAS is sample collection, which plays a critical role in the success of the study. The number of samples included in the study is a crucial factor to consider, as larger sample sizes offer increased statistical power for detecting true associations. This is because larger sample sizes result in smaller standard errors, making it easier to detect differences between the phenotype and genotype groups, especially when effect sizes are small. Additionally, larger sample sizes increase the chances of detecting rare genetic variants that may have a significant effect on the phenotype[118].

Once the samples have been collected, usually from blood or saliva, the DNA is extracted and genotyped to determine the specific genetic variants present in each participant, generating the input data for the study. In this context, public genetic repositories, and large-scale biomedical databases (also known as biobanks) are crucial, making a large number of samples available to researchers for further analysis.

The Human Genome Project paved the way for extensive research in human genetics, leading to the emergence of numerous large-scale projects and initiatives. The proliferation of multiple projects has led to a substantial output of discoveries and knowledge of interest, triggering the need to share data, making it publicly accessible, and thus allowing it to be used for further research[119].

Data-sharing has led to emergence of genotype repositories in recent years revolutionizing the field of genetics. This has enabled researchers to validate their results, to identify new genetic associations, and to advance our understanding of the genetic basis of complex diseases and other phenotypes.

The main repositories containing genotype information include (i) the National Center for Biotechnology Information's (NCBI) Database of Genotypes and Phenotypes (dbGaP)[24], including information on over 2,000 studies. (ii) The European Genome-phenome Archive (EGA), a centralized repository that provides access to large-scale genotype data from various studies[120]; and (iii) the GA4GH, a resource that provides access to genotypic data from multiple sources, including data from over 100 institutions[121]. These public repositories consist of data from an extensive range of studies, frequently concentrating on specific diagnoses or disease cohorts. Since these repositories compile data from a multitude of sources, there can be significant variation in study populations and recruitment methods across individual datasets. This heterogeneity within the repositories may pose challenges when attempting to compare or merge datasets, as recruitment methods and control types may not be uniformly consistent throughout the studies.

In addition to repositories, genetic biobanks have also played a crucial role in facilitating data sharing and advancing the field of genetics. These biobanks have recently emerged to link genetics and epidemiological factors to disease risk[122]. Genetic biobanks consist of large population-based cohorts with hundreds of thousands of individuals, containing genetic data associated with extensive phenotypic information, making them an invaluable resource for GWAS[123].

Among the existing biobanks, the United Kingdom (UK) Biobank is a vital biobank and one of the first established. Since its establishment in 2006, it has collected extensive biological and medical data from 500,000 UK residents aged 40-69 at recruitment. Biological samples, such as blood, urine, and saliva, were collected from participants to conduct genetic analysis. Additionally, detailed information on their medical history, lifestyle, and health was obtained through face-to-face interviews, touch screen questionnaires, and medical examinations. The UK Biobank also gathered a vast amount of phenotypic data from electronic health records and information from national health registries. The majority of this phenotypic data is based on International Classification of Disease (ICD)-10 codes, a standardized coding system utilized to categorize medical conditions and diseases.

This large-scale study includes genotype data from over 500,000 individuals[124], whole-exome sequencing data from 470,000 individuals[125], and whole-genome sequencing data from over 300,000 individuals[126], with plans to increase their sample size in the coming years.

5.4 Genotype imputation: a key breakthrough in GWAS

The number of markers analyzed in GWAS is also a crucial factor in determining the statistical power and accuracy of the study. As GWAS aims to identify genetic variants associated with complex traits or diseases, a larger number of markers allows for a more comprehensive assessment of the genetic landscape. By analyzing a higher number of markers, researchers can capture a broader range of genetic variation, which increases the likelihood of identifying associations between specific genetic loci and the trait or disease under investigation[118].

Early GWAS studies relied on information from genotyping arrays, which only included a few hundred thousand variants. The limited number of variants included in the analysis resulted in a reduced power to detect genetic variants associated with traits. In addition, commercial genotyping arrays are also biased towards common variants, hindering the study of less frequent variants in the population. In the beginning of the use of GWAS, the assumption was that common variants were the cause of common diseases following the CD/CV principle. However, other trends suggested that lower frequency variants may have greater impact on diseases based on the observation of rare variant with larger effect sizes and, not studying them would be contributing to the gap in our understanding of the genetic architecture underlying complex traits and diseases[102,104,127].

In this context, the most logical solution would be to work with whole-genome sequencing methodologies, which, apart from allowing us to increase drastically the number of variants analyzed, would also allow us to work with lower frequencies[128,129]. However, this is unaffordable due to its high cost[130]. To address this issue, genotype imputation was introduced as a method to infer genotype data based on available data from haplotype reference panels[131], a collection of genetic sequences, that are used as a reference for genomic imputation (Figure 7).

Genotype imputation is based on the principle of LD. The idea behind this process is to leverage the LD relationships observed in a reference panel to predict the missing genotypes in a target sample[132,133].

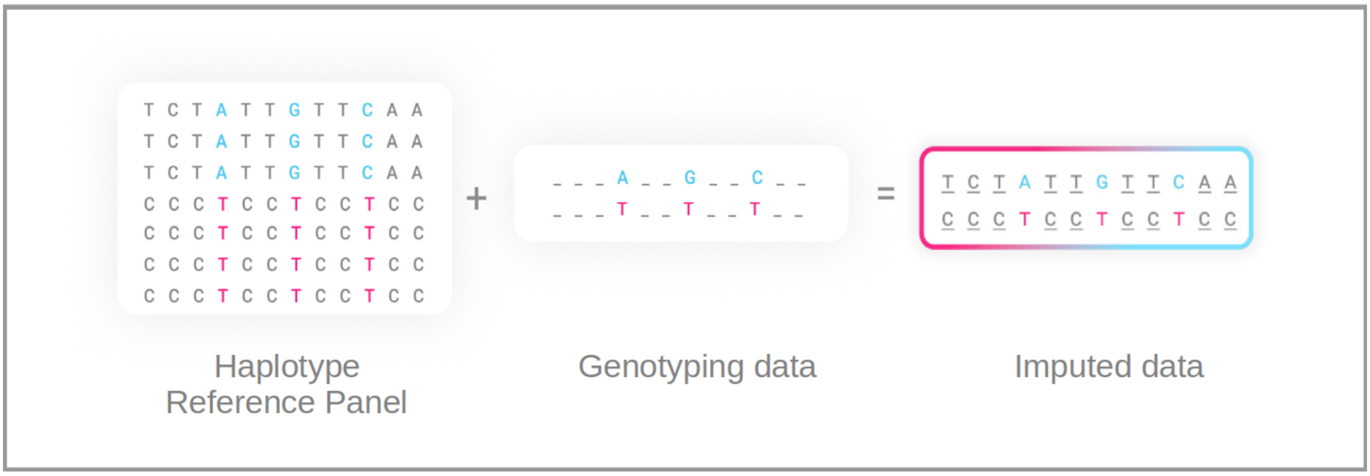


Figure 7. Scheme of imputation process. By using a reference panel of haplotypes, we are able to infer part of the variants which are not present in the genotyping data based on the principle of LD.

A variety of tools and software have been designed to perform imputation, including IMPUTE[124,132–134], BEAGLE[135] and MINIMAC[136] among the most popular. Furthermore, given the high computational demands of genotype imputation, pre-imputation methods have been developed to reduce the computational cost while maintaining high-quality results. The most notable of these is the estimation of haplotypes of individuals in the cohort prior to imputation, known as phasing[134], where the SHAPEIT tool is particularly noteworthy[137–139].

The quality of the haplotype reference panel plays a crucial role in the imputation process. A high-quality reference panel should include a large sample size, represent a diverse range of haplotypes from the target population, and have a high density of genomic markers.

Building a reference panel involves multiple steps, including sample collection, sequencing (typically via WGS), quality control, variant calling, haplotype inference, and validation The first reference panel, 1000G, sequenced 2,504 individuals from various ethnic groups, revealing a wide range of genetic variability[64]. Since then, more panels have been created, including specific population panels as the Genome of the Netherlands (GoNL)[140] and UK10K[30], or larger multi-ethnic panels, as the Haplotype Reference Consortium (HRC)[131].

However, the cost of genome sequencing has limited the quality of these panels, leading to two main problems: low sequencing coverage and small sample size. Low coverage leads to incomplete variant detection and under-representation of SVs, which play a significant role in diseases[141]. On the other hand, a small sample size results in reduced capabilities to detect rare variants, affecting the accuracy of imputation results[140].

The limitations in reference panels, such as lack of rare and SVs or poor quality, negatively affect GWAS by causing their poor imputation quality and exclusion from association studies.

5.5 Limitations of GWAS

GWAS have been very successful in identifying novel variant-trait associations. At this moment, more than 4,000 GWAS have been published[142] for diseases like cancer and its subtypes[143], type 2 diabetes mellitus[144], inflammatory bowel disease[145] or Crohn’s disease[146], among others. However, despite their significant contribution to advancements in genomics, GWAS do present certain limitations.

One of the main limitations is regarding the discovery of associations. Although the use of GWAS is widely established, there is still a gap between the heritability estimated by population-based studies and the amount of heritability that can be explained by currently known genetic variants involved in a given trait[99]. This phenomenon is commonly referred to as the missing heritability problem highlighting that, current methodologies for studying the genetic architecture of complex diseases, such as GWAS approaches, still have several limitations, not being able to discover all the genetic variants influencing the presence of a trait .

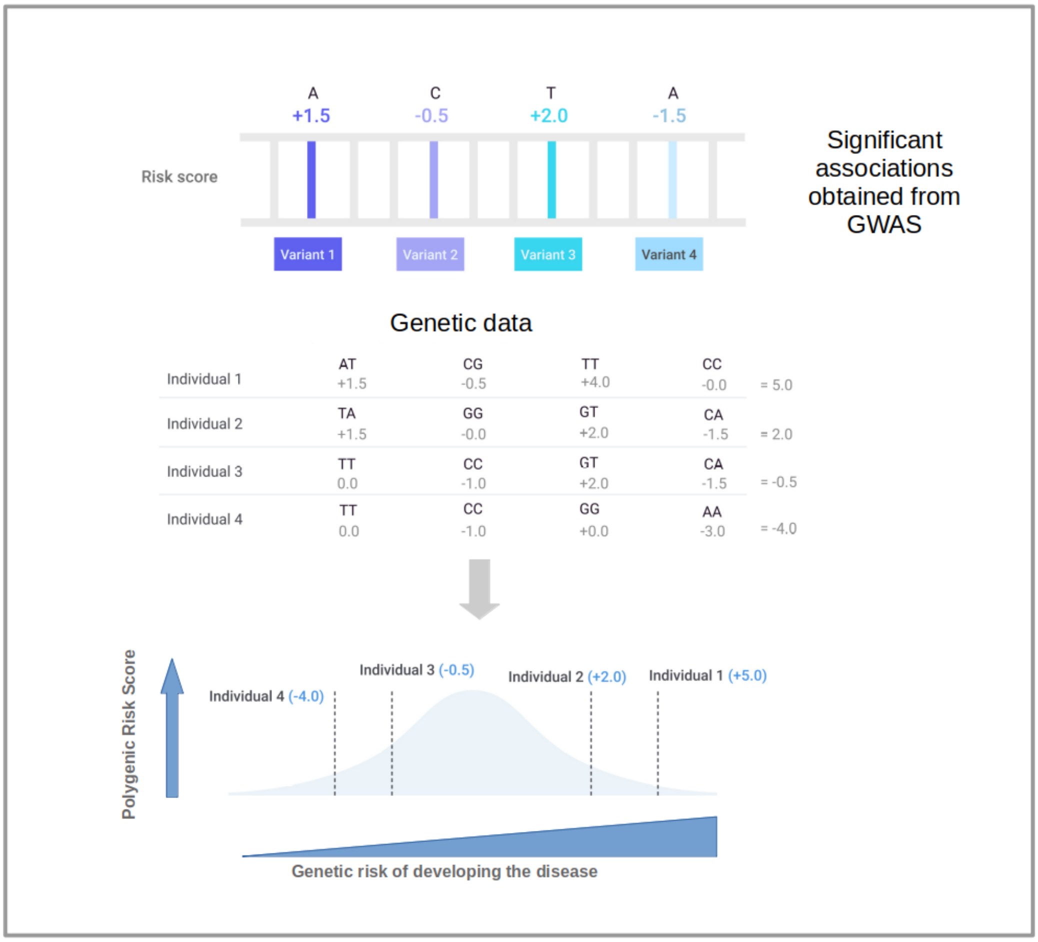
Another limitation of GWAS is that they may not determine the causal variants and causal genes. As explained before, the presence of local correlation among multiple genetic variants due to LD facilitates the initial identification of a locus but makes it difficult to discern the causal variant[147]. In addition, the majority of association signals are located in non-coding regions of the genome, making it challenging to understand their biological implications[148]. Therefore, additional steps such as fine-mapping, functional analysis, or evolutionary genetic analysis are often necessary to identify the causal variants and the target genes[149]. Nonetheless, even though additional methods attempt to explain the functional interpretation of these signals, a large number of loci remain uninterpreted.

Another widely criticized limitation of GWAS is the limited clinical predictive value. The modest explanation of heritability and small effect sizes of the SNVs identified by GWAS hinder their clinical predictive power. However, polygenic risk scores (PRSs), which are quantitative measures of risk calculated from multiple risk alleles, are starting to demonstrate potential in categorizing populations into distinct risk groups that can influence clinical and personal decision-making[150] (Figure 8). However, GWAS have historically been conducted in populations of European ancestry, potentially limiting the applicability of findings to other populations.

Concerning the discovery limitations, as previously mentioned, sample size is a critical factor as it directly influences the power of the analysis. A larger sample size enhances the capacity to detect genuine associations between genetic variants and traits, resulting in more accurate and reliable outcomes. With a smaller sample size, the study might face decreased statistical power, which increases the likelihood of both false positive and false negative results. Therefore, an adequate sample size is crucial to ensure the robustness of GWAS findings and to further our understanding of the genetic foundation of complex traits and diseases. However, assembling large sample sizes for all traits of interest can be challenging, posing a significant constraint. Although GWAS cannot account for all the heritability of complex traits, it still represents a practical way to identify genuine associations, and increasing sample size in GWAS is expected to uncover new loci. Research has shown that for each complex trait, there is a sample size threshold where the rate of locus discovery increases in GWAS[151], and so far, the discovery of risk loci has not yet plateaued for any trait[152].

Another limitation in GWAS regarding the discovery is the number of variants included in the analysis. As previously discussed, the number of variants examined plays a critical role in determining the statistical power and accuracy of the study. GWAS have typically been based on SNP arrays that are imputed with pre-existing haplotype reference panels which commonly lack of rare variants and SVs rising in difficulties for their imputation. These rare and SVs often have larger effects on complex traits[67,153]. This can lead to imputation bias, with missed associations. To address this issue, it is important to use and combine large reference panels that are representative of the population being studied[1,154] and to develop methods that are specifically tailored to handle rare and SVs. Reference panels including a higher number of high-quality genetic variants would allow for an increase in the number of variants interrogated and, therefore, enhance the power of the study.

Lastly, GWAS typically exclude the X chromosome from analysis due to methodological challenges that the analysis of this chromosome involves. The X chromosome, as explained below, contains many genes with important roles in disease, including sex-linked traits and diseases[155]. As explained in more detail in the following section, despite being a challenge, it is necessary to develop a pipeline and methods that allow the systematic integration of the X chromosome into GWAS studies, owing to its interesting characteristics.



Individual 1	AT +1.5	CG -0.5	TT +4.0	CC -0.0	= 5.0
Individual 2	TA +1.5	GG -0.0	GT +2.0	CA -1.5	= 2.0
Individual 3	TT 0.0	CC -1.0	GT +2.0	CA -1.5	= -0.5
Individual 4	TT 0.0	CC -1.0	GG +0.0	AA -3.0	= -4.0

Figure 8. Polygenic Risk Scores. Polygenic risk scores are calculated by computing the sum of risk alleles corresponding to a phenotype of interest in each individual, weighted by the effect size estimate of GWAS. People with a higher score have a greater genetic risk of presenting the trait or disease.

6 The impact of the X chromosome in human genetics

6.1 Biology of the chromosome X

The X chromosome is one of the two sex chromosomes in humans. With a length of approximately 155 million base pairs, it is the 8th largest chromosome in the genome, and it contains over 1200 genes[156]. The X chromosome is involved in determining the biological sex of an individual, with females possessing two copies of the X chromosome, while males having one X and one Y chromosome. The X chromosome is inherited to offspring, with females inheriting one X chromosome from each parent and males inheriting one X chromosome from their mother and one Y chromosome from their father.

The X chromosome can be divided into two distinct regions: the pseudoautosomal regions (PAR) and the non-pseudoautosomal region (non-PAR). The PAR regions (PAR1 and PAR2) are homologous sequences between the X and Y chromosomes that resemble sequences in autosomes and are present in both men and women in two copies. In contrast, the non-PAR region has no homologous sequence in the Y chromosome, resulting in a difference in dosage between men and women.

To balance the difference in dosage, the non-PAR region of one copy of the X chromosome in females is silenced. In males, the absence of second copy of the X chromosome, implies that they have only one copy of X-linked genes, condition known as hemizigosity. Females, on the other hand, achieve a state of functional hemizigosity for most X-linked genes through X- inactivation. This leads to the expression of only a single functional X chromosome per cell, resulting in equivalent X chromosome-linked gene expression between males and females. The X-inactivation process, also known as lyonization, therefore plays a crucial role in maintaining the balance of gene expression between the sexes and helps to ensure normal cellular function[157].

The process of X chromosome inactivation occurs during early embryonic development and is randomly and independently established in each cell. This random inactivation generates female mosaicism, meaning that females possess a mix of cells with either the maternal or paternal X chromosome inactivated. Once the inactivation has occurred, it is stably maintained during further cell division, ensuring that the same X chromosome remains inactivated in all daughter cells.[156,158]. X chromosome inactivation is accomplished through the silencing of the entire X chromosome by accumulating repressive epigenetic marks, such as DNA methylation, histone modifications, and the accumulation of non-coding RNA molecules known as XIST (X-inactive specific transcript)[159].

However, certain genes on the X chromosome can escape X-inactivation and remain active. This phenomenon is referred to as “escape from X-inactivation”[160]. Escape from X-inactivation can occur for several reasons, including differences in the location and structure of the XIST gene, the presence of cis- or trans-acting elements that modify XIST expression, and other epigenetic modifications[161]. As a result, escape from X-inactivation can lead to differences in gene expression patterns between females, which can contribute to individual differences in phenotypic traits and disease susceptibility[156,162,163].

6.2 Impact of the chromosome X in complex diseases

It has long been known that the incidence and severity of many complex diseases varies between males and females. Males and females commonly differ, for example, in disease prevalence, symptoms, and drug response. However, the mechanisms behind these differences are still poorly understood, with important clinical repercussions[164]. Both, environmental and genetic factors are implicated in sexual differences in complex diseases. From the genetic perspective, evidence is accumulating for variants having differential risk effects between the two sexes[165]. In this context, the X chromosome is particularly interesting, as, besides carrying such variants, it could further contribute to sex differences through dosage imbalance mechanisms. Indeed, while sexual dosage compensation is achieved by the random inactivation of one X chromosome in females, about 30% of X-linked genes escape this inactivation[162,166], realizing an intrinsic dosage imbalance between sexes.

There is a growing body of evidence that the X chromosome plays a significant role in complex diseases. For example, a number of studies have reported an over-representation of X-linked genes in several complex diseases, including psychiatric disorders[167], and autoimmune diseases[168]. Additionally, several studies have identified X-linked genetic variants that are associated with complex diseases, including variants in the *FOXP2* gene, which has been implicated in the development of language-related disorders[169], and variants in the *NALCN* gene, which has been linked to the development of several neurological disorders[170].

This growing body of evidence suggests that the X chromosome plays a significant role in complex diseases and has a critical impact on sex differences. However, much more research is needed to fully understand the mechanisms behind the X chromosome's influence, and how it contributes to sex-differences in disease susceptibility and severity. Future studies will likely focus on developing improved methods for X-chromosome analysis and investigating the X chromosome's role in a broader range of complex diseases.

6.3 Widespread exclusion of the X chromosome in association studies

Upon examining the SNP-trait associations reported in the National Human Genome Research Institute's (NHGRI) GWAS catalog, a clear shortfall in the representation of the X chromosome becomes apparent[171]. Only 0.5% of the known associations map on chromosome X. Assuming a correlation between chromosome size/genetic variability with the expected number of loci associated to disease[172], this 0.5% is almost 10 times less than expected from a similar chromosome size (e.g., chromosome 7 and 8, with 4.8% and 4.5% of the associations, respectively), and 8 times less than expected from a similar chromosome variability (e.g., chromosome 16, 3.8%)[173]. (Figure 9)

The under-representation of the X chromosome in GWAS is primarily due to the added methodological challenges associated with its analysis[164,172]. These challenges arise from the unique pattern of inheritance of the X chromosome, both of which results in allelic imbalances between sexes, lower power in males and possible noisy statistics during genetic association studies[174]. Furthermore, the complex processes related to X-inactivation and escape from X-inactivation in females can introduce noise into the statistical analysis. The presence of both inactivated and escape genes on the X chromosome generates a mosaic pattern of gene expression in females, which can confound the analysis of genotype-phenotype associations and may mask or obscure true genetic effects[160,175]. Further, the X chromosome is more prone to false positives, due to the higher overall rate of genotyping errors. This has resulted in many projects directly excluding the X chromosome from their genotyping data, which is visible in a large part of the genotyped cohorts present in EGA[120] or dbGaP[24].

Hence, X-chromosome analysis requires specific strategies and algorithms, including additional quality control measures and alternative association testing methods distinct from those for the autosomes, to account for male hemizygosity and female dosage compensation. For instance, while analyzing data from both sexes combined can uncover global variant-phenotype associations, separating the sexes can enhance discovery and reveal potential signals for sex-specific phenotypes, aiding in the identification of associations that might be masked when both sexes are analyzed together[176].

Gradually, the scientific community is working on new strategies and methodologies[177] to overcome the limitations of X-chromosome analysis, but there are still limited studies being conducted in this area. A deeper knowledge of the connection between complex diseases and the X-chromosome would have a major effect, not just in addressing the issue of missing heritability, but also in deciphering and explaining the differences in disease susceptibility and manifestation that are seen between males and females[178].

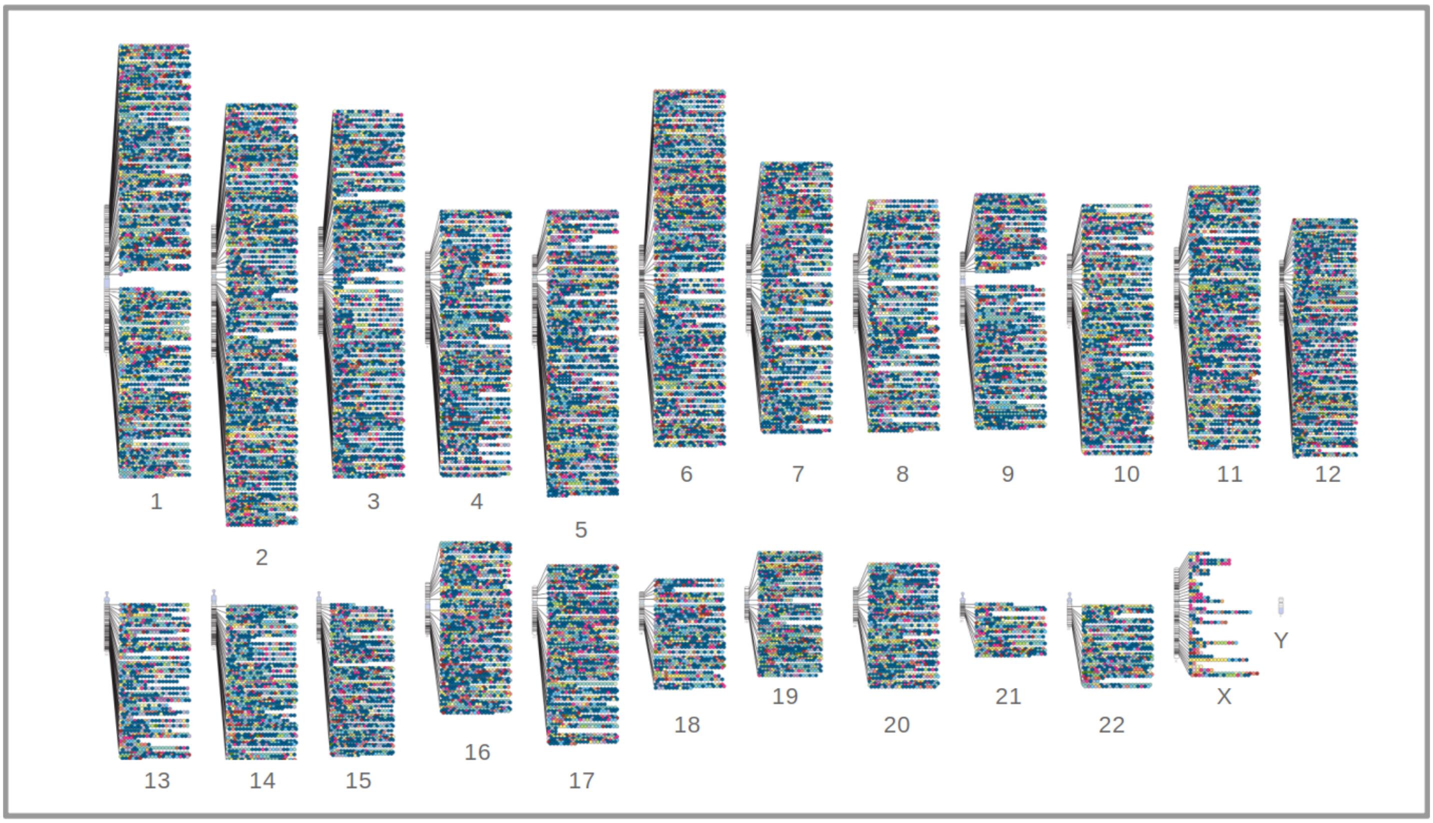


Figure 9. GWAS catalog diagram associations (p-value < 5x10^{-8}) in February 2023. Each of the dots represents an association reported in the GWAS catalog. Although the X chromosome is similar in size to chromosome 7, the number of reported associations is smaller than expected, being even smaller than chromosome 22, the smallest of the human autosomes.

Objectives

The main aim of this thesis is to contribute to the understanding of the genetic basis of complex disease, by addressing specific gaps and limitations within the field. These specific aims are focused on the generation of a genetic resource, i.e. an Haplotype map, and on the understanding and completing the study of the role of the X chromosome in complex traits.

Specific goals that have been pursued for each block:

- Development a comprehensive structural variant haplotype map from high-coverage whole-genome sequencing (GCAT|panel):
    1. Design, implementation and benchmarking of a complete variant calling pipeline for the detection and genotyping of all types of germline variation, including SNVs, INDELs, and with particular focus on SVs.
    2. Benchmarking and application of phasing strategies and validation through imputation across different populations.

- Recovery of the role of the X chromosome in complex diseases:
    1. Collection and preparation of genetic and phenotypic data from UK Biobank and dbGaP databases.
    2. Design, implementation and application of a X-Chromosome-wide association study (XWAS) strategy for the identification and classification of genetic associations with complex diseases.
    3. Identify and characterize previously unknown variants of the X-Chromosome that are associated with complex diseases. Identification of sex-specific signals

Methods

The methods section is divided into two primary sections. The first section outlines the resources and methodologies employed to create the GCAT reference panel. This section covers (1) the design and validation of the variant calling strategy; and (2) the use of the strategy to generate the GCAT reference panel and its comparison with other databases. The second section describes (3) the development of a strategy to include the X chromosome systematically in complex trait association analyses, and its application to identify new loci; (4) a fine-mapping and gene-mapping analysis of the discovered loci; and (5) a preliminary investigation of sex differences.

Development a comprehensive structural variant haplotype map from high-coverage whole-genome sequencing (GCAT|panel)

1 Selection of the best approach for variant identification and validation

1.1 Samples used for testing the variant calling strategy

We evaluated the performance of each variant caller, including their precision, recall, and computational time, using two samples with known genetic variants:

- **Artificial sample (*in silico):*** this sample was previously created[1] at the computational genomics group at the Barcelona Supercomputing Center (BSC) using ART software (ART-Illumina)[179]. This sample mimics a human genome (reference hg19) including 5,330,762 human variants described in the 1000Genomes[26] and the PanCancer[180] projects. Additionally, 3,925 artificial structural variants (SVs), covering different ranges of size and type were included[1].
- **Genome in a bottle sample (GIAB)**: sample NA12878 from the GIAB Consortium[181]. This sample was downloaded in its 30X BAM file version from NCBI, where it is publicly available (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NIST_NA12878_HG001_HiSeq_300x/). Additionally, the variant call file (VCF) containing all the variants validated by the GIAB consortium was obtained (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001 /latest/GRCh37/).

To minimize technical variability and ensure that the results obtained were as accurate and reliable as possible in the analysis of both samples, we implemented the GATK Best Practices[182]. This toolkit comprises a collection of guidelines and recommendations aimed at achieving high-quality results when processing genomic data. First, we used PICARD (version 1.108) to mark duplicated reads; then we applied two modules (VariantRecalibrator and ApplyVQSR) from the GATK4 package (version 4.0.11) to recalibrate the Base Quality Scores (BQSR) of the BAM file. The resulting files are available at http://cg.bsc.es/GCAT_BSC_iberianpanel/.

1.2 Preliminary variant caller selection

Variant callers are software tools designed to identify genetic variation in genomic data by comparing it with a reference genome. To achieve accurate results, they employ statistical models to evaluate the probability of each variant call and filter out false positives due to sequencing errors or other technical artifacts. However, different variant callers use distinct detection strategies leading to variations in their ability to detect different types and sizes of variants.

To ensure comprehensive coverage of variant types and sizes in developing the reference panel, we conducted a thorough selection process of variant callers based on the data described in their documentation. This selection was guided by their detection strategy, ability to detect specific types of variants, and accuracy and recall in different size ranges. Consequently, we selected 16 variant callers, of which five were capable of detecting single nucleotide variants (SNVs) and small insertions and deletions (INDELs) (Table 1), and 11 were able to detect SVs (Table 2), including deletions, insertions, duplications, inversions, translocations, and mobile element insertions (MEIs), based on their performance in detecting these specific types of variant and sizes.

Variant caller	Type of variant detected	Calling strategy	Genotyping method	Resources per sample
Haplotype caller[183] (version 4.0.2.0)	SNVs, INDELs and mid-size deletions and insertions (< 300bp)	Split Read Assembly	Included in the software	16 CPUs
Strelka2[62](version 2.9.2)	SNVs and INDELs	Assembly	Included in the software	48 CPUs
Deepvariant[63] (version 0.6.1)	SNVs and INDELs	Deep Neural Network	Included in the software	48 CPUs
Platypus[60] (version 0.8.1)	SNVs, INDELs, deletions (< 2kb) and insertions (< 500bp)	Assembly	Included in the software	16 CPUs
Varscan2[61] (version 2.4.3)	SNVs and INDELs	Split-Read + BAM Map Quality	Included in the software	16 CPUs

Table 1. Variant callers selected for SNVs and INDELs detection.

Variant caller	Type of variant detected	Calling strategy	Genotyping method	Resources per sample
Manta[78] (version 1.2)	Deletions Duplications Inversions Insertions Translocations	Split-Read Discordant-Read Assembly	Included in the software	24 CPUs
Delly2[77] (version 0.7.7)	Deletions Duplications Inversions Insertions Translocations	Split-Read Discordant-Read Read-depth	Included in the software	24 CPUs
Lumpy[73] (version 0.2.13)	Deletions Duplications Inversions Break-ends	Split-Read Discordant-Read Read-depth	SVtyper tool[184]	Pe-processing: 12 CPUs Calling: 24 CPUs Genotyping: 1 CPUs
Pindel[185] (version 0.2.5b9)	Deletions Duplications Inversions Insertions Translocations	Split-Read Discordant-Read	In-house script*[1]	184 CPUs (8 per chromosome)
SvABA[186] (version 7.0.2)	Breakpoints without indicating the type of variant	Split-Read Discordant-Read Assembly	Included in the software	16 CPUs
CNVnator[187] (version 0.3.3)	Deletions Duplications	Read-Depth	Included in the software	12 CPUs
Wham[74] (version v1.7)	Deletions Duplications Insertions Inversions	Split-Read Discordant-Read Machine learning classification	SVtyper tool[184]	Calling: 48 CPUs Genotyping: 1 CPUs
Popins[75] (version damp v1-51)	Insertions	Assembly	Included in the software	48 CPUs
Genome Strip[188]*[2] (version 2.0)	Deletions Duplications	Split-Read Discordant-Read Read-Depth	Included in the software	48 CPUs
Pamir[189]*[2] (version 1.2.2)	Insertions	Split-Read Discordant-Read Read-Depth One-End Anchored	Included in the software	48 CPUs
Melt[76] (version 2.1.4)	MEIs	Split-Read Discordant-Read	Included in the software	24 CPUs

Table 2. Variant callers selected for SVs detection. *[1]. This script extracts the position information from the BAM file where a variant was detected and calculates the proportion of altered reads (with a mapping quality >20). If the proportion was less than 20%, genotype was 0/0. If the proportion was between 20 and 80%, the genotype was 0/1. If the percentage > 80%, the resulting genotype was 1/1.

1.3 Variant caller execution

Once we selected the 16 variant callers, we ran them on both the *in silico* sample and the sample NA12878.

All variant callers were run on the Marenostrum 4 supercomputer at the BSC. This computer has 48 racks housing 3,456 nodes with a total of 165,888 processor cores and 390 Terabytes of main memory. Each compute node is equipped with 48 cores (96 GB of main memory 1,880 GB/core, 12x 8GB 2667Mhz DIMM (216 special nodes with high memory, 10,368 cores with 7,928 GB/core)). The processors support well-known vectorization instructions such as SSE, AVX up to AVX-512.

Although we initially selected GenomeStrip (version 2.0) and Pamir (version 1.2.2) software, we had to discard them due to their incompatibility with the LSF system of Marenostrum 4. As a result, we reduced the list of variant callers to 14.

1.3.1 Variant callers for SNVs and INDELs detection

All variant callers described in this section were used to detect SNVs and INDELs in the *in silico* sample and the NA12878 sample. The required resources and detection strategies for each tool are detailed in Table 1.

- **Haplotype caller (GATK package):** We ran Haplotype Caller using the default parameters and the following command flags: -ERC GVCF, --dbsnp dbsnp_138.b37.vcf -L chr -G Standard Annotation.
- **Strelka2:** The execution involved two main steps performed using default parameters. First, we used the configureStrelka GermlineWorkflow.py script to configure the calling process. Then, we performed the calling process by running the runWorkFlow.py script.
- **Deepvariant:** This execution involved three main steps, each executed with default parameters. First, we ran the make_examples.zip module to create a reference calling dataset. Second, we ran the call_variants module for variant calling. Finally, we filtered the output using the postprocess-variants.zip module.
- **Platypus:** We ran Platypus with default parameters and added the following flags: --assemble=1 –assembleBrokenPairs=1 mergeClusteredVariants=1.
- **Varscan2:** Varscan2 was run using default parameters in its germline variant detection mode.

1.3.2 Variant callers for SV detection

The variant callers explained in this section were executed on the *in silico* sample, for which the list of SVs was known in advance. The specific types of variants detected, resources needed for their execution, and detection strategies employed by each tool are detailed in Table 2.

- **Manta:** Manta was executed in two steps using default parameters. First, we ran the configManta.py module, which maps regions in the genome where SVs can be found. Next, we ran the runWorkflow.py module, which identifies SVs in each region and assigns a score, resulting in the final VCF file.
- **Delly2:** We executed this tool once for each type of structural variant by modifying the flag -type accordingly. Moreover, we excluded the telomere and centromere regions using the flag -x. Subsequently, all types of variants were merged into a single file using the CALL BCF and MERGE modules.
- **Pindel:** The execution of Pindel involved two steps. In the first step, variant calling was performed, resulting in an output coded in Pindel's format. In the second step, we used the pindel2vcf module to convert that format output into VCF format. Each of these steps was performed for each chromosome individually using the following parameters: -a 3 -C -k -l -I -M 8 - T 6 -x 5 -v 10 -c 1 -R hs37d5 -d Feb2009.
- **SvABA:** SvaBa was run with default settings in its germline variant detection mode.
- **CNVnator:** CNVnator was executed using a bin size parameter of read-length=150. The output of this tool was in .root format, and we converted it into VCF format using the cnvnator2VCF.pl script included in the software package.
- **Lumpy**: We first preprocessed the BAM file following the recommendations of the Lumpy developers. This involved extracting split reads with the extractSplitReads_BwaMem module and discordant reads with Samtools[46] (version 1.5) from the BAM file. We then ran the variant detection process with the Lumpyexpress module using default parameters. All detected variants were genotyped using SVTyper. We filtered out variants with quality scores < 20 and BNDs with both endpoints located in the same chromosome.
- **Wham:** The variant calling process was executed with default parameters, followed by the genotyping process using SVTyper.

- **PopINS:** To execute PopINS, we followed several steps using default parameters: 1) assembly, 2) merge, 3) contigmap, 4) place-refalign, 5) place-splitalign, 6) place-finish, and 7) genotyping. Variants labeled as NO-ANCHOR in the output file were filtered out.
- **MELT:** The execution process for MELT involved four steps, all of which were executed with default parameters: (i) pre-processing the BAM file, (ii) individual analysis, (iii) genotyping, and (iv) creating the VCF file.

1.4 Benchmarking of variant callers

Once the 14 variant callers were executed and their corresponding variant caller files were obtained, we assessed their performance by evaluating their precision, recall, and F-score.

$$Recall = \frac{TP}{TP + FN} \qquad Precision = \frac{TP}{TP + FP} \qquad F - score = 2 * \frac{Recall * Precision}{Recall + Precision}$$

To analyze the performance of the variant callers that detect SNVs and INDELs, we used both the *in silico* sample and sample NA12878 as reference. We then compared the metrics of each variant caller between the two samples to ensure consistency. For the benchmark of variant callers that detect SVs, we solely used the *in silico* sample as a reference, as it was the only sample for which the list of SVs that contains was known. Notably, we did not consider variants on the Y chromosome or mitochondrial chromosome in either of the two analyses.

To determine the number of true positives (TP), false positives (FP), and false negatives (FN), we considered the following parameters:

- Breakpoint detection: A variant caller's ability to accurately detect the exact position of a variant in the genome may vary depending on the detection strategy and the size of the variant. While base-pair resolution was expected for INDELs and SNVs, breakpoint detection becomes less precise as the size of the variant increases.

  Taking this into consideration, we conducted an assessment to determine the optimal window or margin of error for detecting the position of the variant for each variant caller. We tested different detection windows, ranging from 10bp to 300bp, being this maximum value the sample's insert size (10, 20, 50, 100, 200, and 300bp). If the variant in the *in silico* sample fell within the range reported by the variant caller ± the selected window, we considered the variant as correctly detected. Based on this criterion, we estimated the window at which the best precision and recall metrics were obtained.

- Type of variant: Some tools have difficulties in accurately identifying the type of SV detected, even though the breakpoint detection was correct. Therefore, we only considered calls as true positives if they matched the reference samples in both position and type.
- Length: The precision and recall of variant callers are influenced by the length of the variant, generally, the quality of the call decreases as the length of the variant increases.
- Genotype: Genotyping is a process that enables the determination of whether a variant affects one allele (0/1) or both alleles (1/1). While genotyping accuracy is typically high for SNVs and INDELs, some tools encounter difficulties in accurately genotyping SV.

During the benchmarking we divided the process into two primary blocks. Firstly, we evaluated the performance of the detection process, which referred to the ability to detect a variant, including its size and type. Secondly, we assessed the tool's capability to accurately genotype the detected variant. Then, based on the results obtained from the benchmarking, we made a decision on which callers to include in the study and which to discard. The objective of discarding variant callers was to reduce the computational requirements of the project and eliminate those with lower performance.

1.4.1 Variant detection evaluation

Regarding SNVs and INDELs, before assessing the quality of the variant calls, we applied a normalization process on the called INDELs in VCF files. This process consisted of aligning the indel-containing sequence to a reference sequence and then, determining the minimal representation of the INDEL in relation to the reference. The normalization was done using the pre.py tool designed by the Global Alliance for Genomics and Health (GA4GH). After the normalization, we deemed a variant as a TP if the chromosome and position reported by the callers matched the reference samples at base-pair resolution, and the alleles were identical. Any variant that did not meet these criteria was considered a FP. All the variants present in the reference samples that were not detected by the callers were classified as FN.

In the case of SVs, we classified a variant as a TP if (i) it overlapped with the position reported in the *in silico* sample within the window range used for each variant caller; (ii) the type of SV matched the *in silico* sample, and; (iii) the length of the variant reported by the caller varied by no more than 80% from the length of the variant in the *in silico* sample. Any variant that did not meet these criteria was considered a FP. All the variants present in the *in silico* sample that were not detected by the callers were classified as FN.

Using this strategy, we calculate the recall, precision, and F-score for each variant caller. Additionally, for SVs, we estimated these metrics for each SV type in different variant length intervals, including [30-50bp], [50-75bp], [75-100bp], [100-125bp], [125-150bp], [150-300bp], [300-500bp], [500-1000bp], [1000,2000bp], [2000,3000bp], and [> 3000bp].

1.4.2 Variant genotyping evaluation

We assessed the genotyping accuracy of each variant caller for SNVs, INDELs, and SVs by analyzing whether the genotype of the detected variants matched the genotype of the variants present in the reference samples. We classified a variant as TP when its genotype matched the genotype reported in the reference sample, while a variant was considered FP if the genotypes did not match.

After evaluating the detection and genotyping accuracy of the variant callers, we decided to discard Varscan2 and Platypus due to their poor performance compared to the other three variant callers detecting SNVs and INDELs, as well as their longer run time. As for SVs, all the variant callers included in the benchmarking were deemed suitable for the study. Thus, we performed the study including a total of 12 variant callers, consisting of three variant callers for the detection of SNVs and INDELs, and nine for the detection of SVs.

1.5 Development of models for combining calling results and for filtering out low-quality variants

Upon executing the three variant callers for detecting SNVs and INDELs in the *in silico* sample and in the NA12878 sample, and the nine variant callers for detecting SVs in the *in sili*co sample, the next step was to merge the outcomes of each variant caller into a final file that contained all the calls. To accomplish this, we utilized two distinct approaches depending on the type of variant.

1.5.1 Logical criterion model to combine SNVs and INDELs calls

Due to the high performance of each variant caller for SNVs and INDELs detection, we decided to use a logical criterion model to merge and filter the results. This strategy involved retaining only SNVs and INDELs called by at least two out of three callers (Haplotype caller, Deepvariant, and Strelka2) with the same position and alleles. As was done in the benchmarking process, INDELs calls were normalized in order to have a consensus between the different variant callers and to be able to combine the calls appropriately. Then, for each variant, we reported the consensus genotype among the callers that detected the variant. If a variant was detected by two variant callers, but they did not report the same genotype, the variant was considered as missing genotype “./.”.

1.5.2 Logistic Regression Models for SVs

Regarding SVs, to combine the results from the variant callers, we established a set of criteria to determine if a variant detected by different callers could be considered the same. Our strategy consisted of considering that two variants called by different tools were the same if (i) the position ± the breakpoint detection deviation window of each caller coincided; (ii) the variant type was the same; (iii) the size of the variant detected by both callers did not vary by more than 80%. In this way, one file was generated per SV type, which included all the combined calls from the variant callers that were able to detect that type of variant.

Once we combined the results from the variant callers, we design a filtering strategy to remove low-quality variants by using logistic regression models (LRMs). A total of five LRMs were created, one model per SV type excluding MEIs, which were detected by only one variant caller and there was no need to combine the calls. To filter variants, we chose LRMs strategy due to the limited number of features and the large number of variants that needed to be filtered.

We used the R packages *caret* (version 6.0-85) and *e1071* (version 1.7-3) to develop the LRMs. The function to train the LRMs was:

```
train(PASS ~ independent variables, data=database70, method="glm",
family="binomial", trControl= ctrl)
```

The models were trained using predictors including the presence or absence of detection by each variant caller, the size of the detected variant, the number of variant callers detecting the variant, the number of strategies detecting the variant, and the breakpoint resolution of each caller. The output of the model was a binary variable (PASS or NO PASS) indicating whether a variant should be filtered or kept.

To train the models, we used 10-fold cross-validation for a random set of variants (70%) from the *in silico* sample and tested them using the remaining set of variants (30%). In addition, Receiver Operating Characteristic (ROC) curves and area under the ROC curve of the LRMs were computed for the test sets of each SV type using the "ROCR" R package.

After training the models, we applied each model to the combined call file of each SV type, retaining only those variants that received a PASS output.

1.5.3 Strategy to report SVs position, length, and genotype

Once the results of the variant callers were combined and the different LRMs were applied to obtain the final set of variants detected in the *in silico* sample, consensus was necessary to determine the position, size, and genotype of each variant. To determine the position of each variant, we reported the position provided by the variant caller with the best accuracy among those that detected the variant, as determined during the benchmarking. To determine the size of each variant, we calculated the median of the sizes reported by the variant callers who detected the variant. Finally, regarding the genotype, the reporting strategy was different depending on the type of SV:

- Deletions and Insertions: We reported the most frequent genotype between all the variant callers which call the variant. If a consensus was not possible because a genotype was not called more often than the others, the variant was considered as missing genotype "./.".
- Duplications: Due to the low accuracy of individual variant callers in genotyping duplications, we developed a genotyping method based on BAM information. We calculated the total coverage of the region where the duplication was located. Then, we obtained the number of split-reads altered in that region, sequencing reads where the read is split into two or more segments that map to different regions of the genome, indicating the presence of a structural variant. To obtain that number, first we discarded Hard-clipped reads and those containing insertions or deletions in the Concise Idiosyncratic Gapped Alignment Report (CIGAR).

  Finally, we calculated the ratio of altered reads over the total coverage. If the proportion was ≤ 0.2, we discarded the variant; if it was between 0.2 and 0.8, the genotype was 0/1; and if the ratio was ≥ 0.8, the genotype was 1/1. The script for genotype duplications is at https://github.com/gcat biobank /GCAT_panel/tree/main/genotyping/Duplication_genotyping.
- Inversions: The reported genotype was determined by the variant caller that showed the highest accuracy in estimating the genotype during the benchmarking process. Based on the results obtained in the benchmarking the order of priority was:

  1. Lumpy 2. Pindel 3. Whamg 4. Delly2 5. Manta

- Translocations: We adopted a strategy similar to that used for duplication genotyping based on the BAM file information. We counted all the reads covering the breakpoint to obtain the total coverage, while the number of altered reads was determined by considering all reads with a mapping quality ≥ 20, a CIGAR label different from 151M, and by discarding hard-clipped reads. Finally, we calculated the proportion of altered reads to the total coverage, and based on this ratio, the variant was either discarded if the proportion was ≤ 0.2, genotyped as 0/1 if the proportion was between 0.2 and 0.8, or genotyped as 1/1 if the ratio was ≥ 0.8. The script for translocation genotyping can be found at https://github.com/gcatbiobank/GCAT_panel/tree/main/genotyping/Translocation_genotyping.

1.5.4 Comparison of LRMs variant filtering against other strategies

To evaluate the effectiveness of our SV filtering strategy, we compared it with the method used by the Genome of the Netherlands (GoNL) project, which aimed to sequence the whole genomes of 250 Dutch parent-offspring families and build a haplotype map including SVs. The GoNL project filtered SVs using a similar criterion to our logical criterion for filtering SNVs and INDELs (described in section 1.4.1), considering a variant valid if at least two variant callers detected it.

In addition to the comparison with the GoNL strategy, we conducted two additional studies where we varied the minimum number of callers required to consider a variant valid to three and four callers. Using our variant set generated from the variant calling of the *in silico* sample, we applied the different filtering strategies and estimated the accuracy and recall of each one comparing them with the metrics obtained by using LRMs.

2 Development of the GCAT reference panel

2.1 The GCAT cohort

The GCAT project aims to study the genetics behind complex diseases. It consists of 19,267 volunteers recruited in the region of Catalonia (Spain). It is a cohort of unrelated participants aged 40-65 years and mainly of Caucasian origin (16 % non-Caucasian)[190].

The genetic data available from this project includes Single Nucleotide Polymorphism (SNP) array data (N=5,459; 56% female) and whole genome sequencing (WGS) data (coverage= 30X, N=808, 50.6% female). 71% of the WGS samples (n=570) were also analyzed in the SNP array set.

WGS data were generated at the Germans Trias I Pujol Health Sciences Research Institute, Badalona, Spain, using paired-end synthesis (SBS) sequencing on an Illumina HiSeq 4000 sequencer (Illumina, San Diego, California, USA). These data were obtained in FASTQ format and converted to BAM file format, aligning the reads against the hs37d5 reference genome.

2.2 Quality control

2.2.1 Alignment quality control

Alignment quality control was performed applying the GATK Best Practices, a set of recommended procedures for processing high-throughput sequencing data to produce accurate and reliable variant calls. In this step, one sample was discarded from further analysis as it was not within the metrics required to pass quality control. (Table 3)

Quality Control metrics required
Fraction purified reads > 0.90
Fraction read aligned in pairs > 0.95
0.495 < Strand balance < 0.505
250bp < Mean insert size < 350bp
Standard deviation of insert size < 50bp
Fraction of duplicated reads < 0.1
27X < Mean Coverage < 37X
Fraction of paired reads mapped in the same chromosome > 0.88

Table 3. Alignment Quality control metrics required

2.2.2 Contamination analysis

In this step, we aimed to verify the absence of cross-contamination among the 807 WGS samples, which refers to the presence of genetic material from a different individual or sample in the sequenced sample. We conducted a contamination study using the VerifyBamID tool[191]. VerifyBamID can detect contamination and sample swaps using two different approaches: the sequence+array method, which compares external genotype information with sequence reads, and the sequence-only method, which estimates contamination by modeling the sequence reads as a mixture of two unknown samples using allele frequency information in the VCF file.

First, we performed variant calling with Haplotype Caller following the strategy outlined in the benchmarking (section 1.2). Using the resulting VCF file, we then ran VerifyBamID with the command:

```
VerifyBamID --best --ignoreRG --maxDepth 30 --precise run
```

We first analyzed the samples for which SNP array data was also available (n=570). For this analysis, the thresholds used to determine contamination were [CHIPMIX] ≥ 0.02 and [FREEMIX] ≥ 0.02. [CHIPMIX] is the estimated proportion of contaminated sequence reads in the BAM file, while [FREEMIX] is the estimated proportion of contaminated genotypes based on array data. If 2% or more of non-reference bases were observed in reference sites, the sample was considered contaminated. Next, we analyzed samples for which genotyping information was not available (n=237) using the sequence-only method. For this analysis, the thresholds used to determine contamination were [FREEMIX] ≥ 0.03 and [FREELK1] - [FREELK0], as recommended by the developers. None of the 807 samples showed contamination.

2.2.3 Population structure analyses

To ensure a homogeneous cohort and remove samples with a high level of genetic relatedness, we conducted several steps. First, we performed two principal component analyses (PCAs) using reference populations with known ancestry to identify potential outliers, the 1000Genomes project and the Population Reference Sample Project (POPRES). We used the variant calling results from the contamination analysis and filtered variants based on minor allele frequency (MAF) > 0.01 and linkage disequilibrium (LD) $r^2 < 0.2$ using PLINK (version 1.9). We matched the list of variants obtained with the 1000genome data and retained the variants that were present in both datasets, resulting in a final subset of 1 million variants. Then, we estimated the PCs based on known ancestries using 1000 Genomes data and projecting the GCAT data onto these principal components. This analysis led to the removal of 16 GCAT samples. After removing those 16 samples, we repeated the process using the Population Reference Sample Project (POPRES) as the reference sample. This PCA was performed using the LASER project webserver[192], resulting in the removal of an additional two GCAT samples. In total, 18 GCAT samples of non-Iberian ancestry were discarded in this step.

Next, we performed Identity by Descent (IBD) analysis to detect and discard samples with at least a 3rd degree of genetic relatedness using the PLINK software. We estimated IBD probabilities in the remaining 789 GCAT samples and identified one full-sibling pair with probabilities of sharing 0, 1, and 2 IBD alleles equal to (0.3, 0.48, 0.22) and one first-cousin relationship with probabilities of (0.78, 0.22, 0). For each related pair, we kept the sample with the lowest proportion of missing genotypes.

Finally, we conducted an additional PCA without reference populations using the 1M variant subset on the remaining 787 samples. This analysis resulted in the removal of two additional GCAT samples as they were found to be outside of the generated cluster.

In total, we removed 22 samples, resulting in a final set of 785 WGS samples used in subsequent steps of the project.

2.3 GCAT cohort variant calling

We conducted the variant calling process on the 785 WGS samples that passed the quality checks. This process was divided into two parts: calling of SNVs and INDELs, and calling of SVs.

2.3.1 SNVs and INDELs detection

Strelka2 and DeepVariant were executed following the steps outlined in section 1.3. However, for Haplotype Caller, the availability of a large number of samples in the cohort allowed us to utilize additional modules to improve the accuracy of the variant calling. Initially, we utilized the GenomiDBimport module to merge individual VCFs from each of the 785 samples into a single file, grouping variants into 1MB batches. After merging the results, we genotyped all variants using the GenotypeGVCF module. We then ran the VariantRecalibrator and ApplyVQSR modules to recalibrate the score provided by the software for each variant and regenotyped the variants using the CalculateGenotypePosteriors module. Finally, we filtered out variants with a low genome quality score (<20) using the VariantFiltration module.

2.3.2 SVs detection

We performed SV detection on the 785 samples by executing the selected 11 variant callers as explained in section 1.3. However, using a sample set instead of an individual sample enabled running the variant calling of certain tools in multi-sample mode. This approach allowed for analyzing all samples at once, thereby reducing the runtime. The software tools that supported this capability included Delly2, which permits the use of the FILTER module to discard redundant calls between samples after variant calling; Pindel; Lumpy, which also enables the discarding of redundant variants with its lsort and lmerge modules; and MELT, which involves an additional step after individual analysis, where group analysis was conducted to enhance the accuracy of the calls.

2.4 Selection of the definitive set of variants by filtering out non-reliable calls

We combined the results of variant calling for both SNVs, INDELs and SVs for each sample using the strategy outlined in section 1.5. After combining the results, as explained previously variants that were deemed low-quality were filtered out.

As explained in section 1.5.1, for SNVs and INDELs, variants were discarded if at least two variant callers did not detect the variant, or if there was no majority consensus on the genotype reported by the variant callers that detected the variant. This resulted in a VCF file containing the combined variants from the three variant callers used for SNV and INDEL detection for each sample. Finally, the individual files for each sample were merged into a final file. In order for a variant in one sample to be considered the same as a variant detected in another sample, they had to match in chromosome, base pair resolution position, and alleles.

Regarding SVs, after combining the outputs of each variant caller for each sample, we applied the different LRMs generated for each SV type to identify and discard those considered low quality as described in section 1.5.2. We retained only those variants that were labeled as "PASS" by the LRMs. Then, for each variant, we reported its position, length, and genotype, following the strategies outlined in section 1.5.3 for each SV type. Finally, we combined all samples into a single file by (i) variant type, (ii) chromosome, (iii) position, using the maximum value of the callers that detected the variant in each individual as the breakpoint error, and (iv) requiring a reciprocal overlap of the variant ≥ 80% between individuals. We then applied a second filtering process by calculating the PASS rate of each variant across all samples, based on the number of times it obtained a PASS label from the LRM. Variants with a PASS rate greater than 50% passed the filter, while those with a lower rate were discarded. For the set of valid variants generated, we reported the median length and median position of all samples presenting the variant. This information was used as size and position data for each SV, generating a final VCF. In this VCF, we also provided for each variant the allelic count (AC), the Minor Allelic count (MAC), and the Minor Allele Frequency (MAF).

Lastly, we combined the set of SVs and the set of SNVs and INDELs into a single VCF file. Subsequently, we performed a final quality control by filtering out variants that were not in Hardy-Weinberg equilibrium (Bonferroni correction p-value < $5x10^{-8}$) and variants with missing calls of ≥ 10% across the entire cohort. The GCAT catalog with all variants accepted after the filtering step is deposited at http://cg.bsc.es/ GCAT_BSC_iberianpanel/. The GCAT catalog with genotype information is deposited at https://ega-archive.org/ studies/EGAS00001003018).

2.5 Variant validation

We conducted a validation to check the reliability of our final set of variants. Initially, we performed a comparative validation using public datasets by comparing our list of variants with sets of variants from similar projects that have been previously described. Subsequently, we carried out an experimental validation for SV types that were more sensitive to false positives.

2.5.1 Experimental validation

- SNVs and INDELs

We used the GCAT SNP array to validate the accuracy of our methodology to call SNVs and Indels. To this aim, we selected the 570 samples that had both SNP array and whole-genome sequencing data available. The SNP array set included 732,978 SNPs and 1,168 indels in chromosomes 1-23. We determined the number of variants from this set that were present in our final set of SNVs and INDELs, as well as their genotype concordance.

- Inversions

We assessed the accuracy of our methodology for detecting inversions by comparing our results to the experimentally validated inversions from the InvFEST project[193] as a reference. This validated set included 64 inversions mediated by Non-Homologous (NH) mechanisms that lacked inverted repeats at their breakpoints. We compared our inversion set against the InvFEST dataset, taking into account the variant position, size, and frequency in three major continental ancestries (Africa, Europe, and East Asia).

- Copy Number Variations (CNVs)

We performed a comparative genomic hybridization (CGH) analysis to validate large deletions and duplications (size > 20kb). We randomly selected five samples from the GCAT cohort and the NA12878 sample[194] from the 1000Genomes project as a reference sample, for which the list of CNVs has been previously described (https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3658-x#Sec21).

We collaborated with QGenomics (https://qgenomics.com/) to conduct this analysis. They carried out the hybridization process on both sets of samples, with a probe added every 3kb to cover the entire genome. This enabled us to detect changes in probe intensity when the GCAT sample gained, or lost information compared to the reference sample. An increase in probe intensity corresponded to duplications while a decrease in probe intensity corresponded to deletions. Through this strategy, we validated variants that were present in both the reference and GCAT samples, which did not show a change in probe intensity, as well as GCAT-specific variants that produced a change in probe intensity.

2.5.2 Public datasets comparative analyses

- SNVs and INDELs

We compared our final set of SNVs and INDELs with the NCBI dbSNP Build 153 database, which we downloaded from https://ftp.ncbi.nlm.nih.gov/. To merge the datasets, we considered a variant to be the same between the two datasets if it shared the same chromosome, position at base-pair resolution, and alleles. By performing this comparison, we determined the number of variants in our dataset that were also present in dbSNP and the number of GCAT-specific variants.

- SVs

We compared the final set of SVs to various reference projects to determine the number of variants already reported and the number of variants identified for the first time in the GCAT cohort. The reference projects we used were: (i) The Genome Aggregation Database (gnomAD.v.2) downloaded from https://gnomad.broadinstitute.org/downloads, (ii) the Database of Genomic Variants (DGV) available at http://dgv.tcag.ca/dgv/app/downloads?ref=GRCh37/hg19, (iii) the Human Genome Structural Variation Consortium set (HGSVC) downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/working/2520181025_EEE_SV-Pop_1/VariantCalls_EEE_SV-Pop_1/), (iv) the Hall-lab dataset available at https://github.com/hall-lab/sv_paper_042020, (v) the 1000 Genomes (Phase3) dataset available at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/, and (vi) the GoNL (release 6.2) dataset.

We considered a variant from different projects to be the same as the variant reported in the GCAT cohort if they shared (i) variant type, (ii) chromosome, (iii) position with a window of ±1000 bp, and (iv) if the length of the variant reported in a project varied by no more than 80% from the length of the variant reported in the GCAT cohort. Variants found in at least one of the reference projects were considered already described, while all variants found only in the GCAT cohort were considered novel.

2.6 Phasing of the GCAT variant set

The final step to obtain a haplotype reference panel involved phasing the final VCF file containing all the variants identified for the 785 WGS samples of the GCAT project. We employed ShapeIt4[195] (version 4.1.3) and WhatsHap[196] (version 0.18) software tools to perform phasing. We used WhatsHap to extract phasing informative reads (PIRs), which improves the efficiency of the phasing process for SNVs and Indels. To speed up the computation, we ran the process by chromosome in parallel. Finally, we merged all the chromosomes into a single VCF file annotated with the information obtained from WhatsHap. We then compressed with bgzip and indexed with tabix the entire annotated VCF file. Finally, we ran Shapeit4 including the PIRs information obtaining a phased VCF by chromosome.

To convert the phased VCFs file into .hap, .legend, and .sample files, we executed the BCFtools module. We built a haplotype panel for each autosomal chromosome and X chromosome. To generate the haplotype-resolved panel of chromosome X, we separated chromosome X into pseudo-autosomal regions 1 and 2 (PAR1 and PAR2) and non-pseudo-autosomal regions (NOPAR). Then, for male samples, we coded the heterozygous genotypes in NOPAR as "./.".

(GCAT|Panel is deposited at https://ega-archive.org/studies/EGAS00001003018).

2.7 Analysis and performance of imputation

To evaluate the quality of the GCAT reference panel, we conducted several imputation analyses using SNP-genotyping arrays from two different sources: the GCAT dataset and the 1000 Genomes Project dataset. We assessed the accuracy of the imputation by measuring the concordance between the imputed genotype and the genotype obtained from the array or variant calling.

2.7.1 Imputation analysis using the GCAT data

We utilized 95 randomly selected samples from the GCAT reference panel, which had available SNP-genotyping array data. Firstly, we conducted a quality control (QC) analysis on the SNP-genotyping array data using PLINK software[197]. During the QC, we removed variants with missing calls ≥ 10%, monomorphic variants, and INDELs. We also aligned the variants in the forward strand and standardized alternate and reference alleles as A1 and A2, respectively. As a result, we obtained a dataset comprising 754,593 high-quality variants.

Subsequently, we excluded those 95 samples from the GCAT panel, creating a modified panel consisting of 690 samples. We then performed imputation on the 95 samples using IMPUTE2[198] with the modified reference panel, in batches of 5MB.

We utilized the WGS variant calling of the same 95 GCAT individuals as a reference and evaluated the genotype concordance of the imputed variants (info score ≥ 0.7) and the WGS-called genotypes.

2.7.2 Imputation analysis using 1000genomes array data

We used the 1000genomes array data from ftp://ftp.1000genomes.ebi.ac.uk /vol1/ftp/release/20130502/supporting/hd_genotype_chip/ALL.chip.omni_broad_ sanger_combined.20140818.snps.genotypes.vcf.gz. This dataset included 2,318 samples and 2,458,861 variants from 19 populations and 5 main ancestries[199].

First, we applied sample filtering removing those with unknown gender (n=41), related samples (≥2nd-degree relatedness, n=395), and unrepresented populations (Masai population, n=2). Then, we split the samples by population, and we performed an additional QC. We used PLINK software to filter variants for genotype missingness > 0,1, we discarded A-T, C-G sites, variants that were not in Hardy-Weinberg equilibrium (Bonferroni correction p-value < $5x10^{-8}$), individuals with an excess of heterozygosity (± 2 standard deviation) and with missing call rate ≥0.1, obtaining a total of 1,880 samples. For chromosome X, we separated males and females and divided the chromosome in PAR and NONPAR regions.

2.7.3 Imputation performance against different panels of genetic variability

To assess how well the GCAT reference panel performs compared to other widely used reference panels, we utilized IMPUTE2 to impute the GCAT genotyping array data (n=4,988) using various population reference panels: 1000G phase3[26], GoNL-SV[140], UK10K[30], and HRC[131]. Initially, we conducted a QC analysis to eliminate samples with a missing call rate ≥ 0.10 (n=3). We then removed samples that had available array data and were included in the GCAT reference panel (n=537). Following this QC procedure, we obtained a final set of 4,448 samples. We retained variants imputed by each reference panel that had MAF > 0.001 and info score ≥ 0.7. Then, we compared the number of unique variants imputed by the different panels. For SNVs and INDELs, we considered two calls as the same variant between panels if their chromosome, position, and alleles matched. For SVs, we consider that two variants were the same if the type coincide and they match in position using a window of ±1,000 bp.

2.8 Genome-wide Association Study using the GCAT|Panel

We performed a genome-wide association study (GWAS) using the GCAT imputed data (n=4,988). Phenotype data included only chronic conditions defined by Electronic Health Records from the cohort registry (20112-2017), considering ICD-9 codes and descriptions[200,201]. We analyzed phenotypes with more than 50 cases, resulting in 70 binary traits. Genome-wide association tests were done with PLINK 2.0 software for autosomal chromosomes, assuming an additive model for allelic effects and using as co-variates age, sex, and the first five principal components (PCs) obtained from previous QCs. In addition, gender-specific traits were analyzed only for the specific gender. Finally, we plotted a Locus Zoom for those specific regions where we found a signal below the suggestive p-value threshold ($1e^{-05}$), obtained after a Bonferroni correction.

3. Development of a strategy for the systematic inclusion of the X chromosome in complex trait association analyses

3.1 Data collection

We selected two different genetic data sources, from which we obtained SNP genotyping array data, the database of Genotypes and Phenotypes (dbGaP)[24] and the UK Biobank cohort[122]. We opted to use two types of data sources due to the differences in the sample recruitment process.

These differences are primarily related to study design and sampling strategy. Biobanks are typically designed to collect data from a more general population and encompass a broader range of phenotypes, whereas phenotype-specific cohorts, such as those included in dbGaP, concentrate on a particular disease or trait of interest, and have stricter inclusion criteria. This can result in biobanks having a larger and more diverse participant pool, while phenotype-specific cohorts may have a more homogeneous population concerning certain characteristics such as age or health status.

3.1.1 The database of Genotypes and Phenotypes

The dbGaP database was developed to archive and distribute the data and results from studies that investigate the interaction of genotype and phenotype. The cohorts were requested through different projects fulfilling the needs and requirements of dbGaP. In total 169 cohorts were applied for, including more than 90 phenotypes spread over 13 projects.

3.1.2 UK Biobank

The UK Biobank is a population-based cohort study that comprises around 500,000 individuals aged between 40-69 years in 2006-2010 from various regions of the United Kingdom. Participants provided biological samples, including blood, urine, and saliva, for genetic analysis. They also provided comprehensive information on their health, lifestyle, and medical history through touch screen questionnaires, face-to-face interviews, and medical examinations. Moreover, the UK Biobank collected additional phenotypic data, which includes electronic health record data and information from national health registries. A major part of the phenotypic data relies on International Classification of Diseases, 10th Revision (ICD-10) codes, which is a standardized coding system employed to classify medical conditions and diseases.

Under application ID 85085, we obtained and downloaded genotype data for 422,178 individuals and nearly 7,000 phenotypes from the UK Biobank.

3.2 Data selection, preliminary screening, and harmonization

3.2.1 The database of Genotypes and Phenotypes (dbGaP)

To select the cohorts for our analysis, we used specific criteria. Firstly, we only included cohorts for which phenotypic information and genotypic information on the X chromosome was available. Additionally, we solely considered binary traits for the analysis, and a phenotype was deemed eligible for the analysis only if genotypic and phenotypic data were available for at least 4,000 individuals, with a minimum of 1,000 cases across all cohorts for that particular trait. Phenotypes that did not meet these criteria were excluded from the analysis.

All the cohorts that passed the initial screening underwent a formatting process to obtain genetic information in PLINK format. Subsequently, we filtered individuals for each cohort by removing those without genetic or phenotypic data available.

3.2.2 UK Biobank

In the case of UK Biobank, data was for each phenotype obtained from different sources highlighting diagnosis data (ICD10 codes, Data field 41202) and self-reported information (Data field 20001,20002). We decided to analyze only binary traits with at least 1000 cases for each trait to ensure adequate statistical power for detecting association. This screening process resulted in a final selection of 564 different binary traits for analysis of UK Biobank.

For each phenotype, we filtered the controls by excluding individuals with traits similar to those defined in the cases. If controls have similar traits to the cases, there is a risk of misidentifying individuals as controls when they should be classified as cases, leading to a biased analysis. Therefore, by excluding individuals with similar traits to those defined in the cases, the risk of misclassification is reduced, increasing the likelihood of identifying true genetic associations with the phenotype of interest.

3.3 Data quality control

To ensure the quality of the data in our project, we implemented a rigorous sample-level and variant-level quality control processes for each of the cohorts included.

Initially, using PLINK we removed duplicated variants as our first step, then we filtered out variants based on genotype missingness (>0.05), MAF (≤0.01), and variants that were not in Hardy-Weinberg equilibrium (Bonferroni correction p-value < 1x10^{-20}). After filtering, we split the cohort into cases and controls, and applied the same filters to each subset independently. Lastly, both subsets were merged to create a single cohort.

We also performed a sample filtering step to exclude low-quality samples that may introduce bias into the dataset. Specifically, we removed samples with gender inconsistencies, discrepancies in the reported gender of an individual in the sample and their actual genetic sex as determined by their genotyping data, and samples with high levels of missingness (missing genotypes <0.02).

We then applied a quality control step to remove samples with ancestry and relatedness outliers to define a homogeneous population with similar ancestry. Firstly, we merged each cohort with 1000genomes (Phase3), keeping only variants that were common to 1000genomes and the cohort. Then, we applied a quality control step to remove variants based on genotype missingness (> 0.001), minor allele frequency (≤0.01), and variants that were not in Hardy-Weinberg equilibrium (Bonferroni correction p-value < $1x10^{-6}$). We performed LD pruning using the PLINK software, using a windows size of 1000kb, a step size of 100 variant count, and 0.2 as the r^2 threshold.

With the resulting dataset, we constructed the distance matrix to evaluate the genetic relatedness among individuals in a dataset. We used the triangle algorithm of PLINK1.9 to generate a lower-triangular matrix. We then estimated the multidimensional (MDS) scale coordinates and generated PCs to remove ancestry outliers from each cohort.

Next, we generated a genetic relationship matrix (GRM), which quantifies the genetic similarity or relatedness between individuals in a dataset based on their genotypes. We discarded individuals with a relatedness > 0.125 to remove close relatives from the dataset. Finally, we performed PCA using SmartPCA package and removed additional outliers.

Then, we performed a second variant filtering step using the same criteria as in the first on the dataset generated after sample filtering.

An additional step was implemented specifically for the X chromosome in order to ensure the quality of the data. For this step, we performed a filtering process separately for the two sexes. Variants were filtered based on minor allele frequency (MAF) (≤0.01) and genotype missingness (≥0.05%). Next, variants not in Hardy-Weinberg Equilibrium (p-value < $1x10^{-20}$) were filtered out in female non-affected samples. Finally, variants that showed differences in frequency and genotype missingness between non-affected males and females were excluded (p-value < $1x10^{-6}$).

It is worth noting that during the QC process of the UK Biobank cohort, we chose not to perform sample filtering based on relatedness and ancestry. This decision was made based on our selection of the Scalable and Accurate Implementation of Generalized mixed model software (SAIGE)[111] for the association analysis, which can effectively account for family structure within a dataset and adjust for population stratification, mitigating the potential confounding effects of ancestry in the association analysis. This decision allowed us to include a larger number of individuals in the study, which ultimately increased the statistical power of our analysis by providing a larger sample size and increasing the likelihood of identifying true genetic associations.

3.4 Phasing, imputation and association

3.4.1 Phasing, imputation, and association in dbGaP cohorts

The phasing, imputation and association steps, which comprise the chromosome X wide association analysis (XWAS), on the cohorts obtained from dbGaP were performed using GUIDANCE[154]. This tool consists of an integrated framework that is able to perform haplotype phasing, genotype imputation and association testing of large datasets in a single execution, as well in a modular way with optional user intervention. For each of the aforementioned steps, GUIDANCE allows the use of different software. Additionally, one of the major advantages of this tool is that it allows to use multiple reference panels for the phasing and imputation process and to finally obtain the combined results.

In addition to its flexibility, a major advantage of GUIDANCE is that it allows for the distribution of all steps into different modules. For our XWAS analysis, we utilized two modules:

- The first module includes the data pre-processing and phasing processes. The software selected for phasing was Shapeit2[202] (v2.r727). For this process, 4 nodes with 48 CPUs were used for those cohorts containing both male and female samples. For single-sex cohorts, we required 5 nodes with 48 CPUs due to scalability considerations in GUIDANCE.
- The second module includes the processes of imputation, post-imputation QC and association. First, to overcome potential computational limitations, we adopted a strategy of splitting chromosome X into 1MB chunks for imputation. We used the IMPUTE2 (v2.3.2) with five reference panels; 1000G phase3[26], GCAT[1], HRC [131], GoNL[140] and UK10K[30]. After imputation, we performed a post-imputation quality control step and only considered imputed variants with infoscore ≥0.7 and MAF >0.001 for the association analysis. Finally, the association process was performed using SNPtest software[133,144] (v2.5), considering an additive model, sex, and year of birth as covariates. In addition, for cohorts with both female and male individuals, imputation and association was performed separately in the two sexes. For this process, 10 nodes with 48 CPUs were used for cohorts with less than 10,000 individuals. Cohorts with a larger sample size were analyzed using 50 nodes with 48 CPUs.

3.4.2 Phasing, imputation and association in UK Biobank database

Due to the vast size of the UK Biobank cohort, implementing the analysis with GUIDANCE was infeasible because of the significant computational demands of such a large dataset, including memory issues and extensive computing time.

For the UK Biobank analysis, we initially separated the cohort by gender and then performed all subsequent steps separately for males and females. We first performed the haplotype phasing process using Shapeit4[195] (v4.2.0), a tool designed for the efficient analysis of large-scale biobanks. For imputation, first we split chromosome X into 1MB chunks, then we used IMPUTE2 and the five reference panels listed above. As in the post-imputation quality control of the dbGaP cohorts, only imputed variants with info score ≥0.7 and MAF >0.001 were considered for association. Then, we performed a merge of the chunks to obtain a final imputation file for each of the sexes. Finally, we estimated which variants were not in Hardy-Weinberg equilibrium (Bonferroni correction p-value < 1x10-6). These variants and those that gave missing call in the imputation were removed from both the male and female imputation data.

Single variant association test was performed using SAIGE (version 0.43.2). This tool uses saddle point approximation to account for case-control imbalance, a very common characteristic in large-scale datasets, and is also able to take into account relatedness by creating a genetic relationship matrix.

The SAIGE association process involved two main steps, which were executed separately for males and females. In the first step, the null logistic mix model was fit to generate a genetic relationship matrix (GRM) using genotypic data in PLINK format (.bim, .bed, .fam) and a phenotype file containing sample IDs and co-variates. The results produced a Generalized linear Mixed Model Association Test (GMMAT) file in .rda format and a variance ratio file in .txt format, which were both essential for the second step. The first step was executed using default parameters, 4 CPUs per execution, and with the leave-one-chromosome-out (LOCO) option set to false.

In the second step, single-variant association tests were conducted for each of the 564 phenotypes analyzed in our UK Biobank study. To have the necessary formats to execute the SAIGE association analysis, we converted our imputed date into .bgen format using QCTOOL software (v2.0.6). This conversion process required as input file the imputed data and a sample file with the identifiers (IDs) of each sample. Before conversion, we performed a hard-calling process to obtain genotypes from the genotype probabilities generated by the imputation step using a hard-call-threshold of 0.49 in the PLINK software. Then, since the conversion process involved very high computational resources, we split each of the imputed data files into eight chunks. The conversion was performed with QCTOOL[203] using default parameters and the flags --infer-ploidy-from sex and --asume-chromosome X, indicating that the chromosome to be converted was the X chromosome, generating .bgen files in 8-bits format.

Subsequently, we merged the eight chunks for each sex back into a single .bgen file format using cat-bgen module included in the BGEN library package[203]. Finally, each file was indexed using the bgenix module as recommended by its developer, obtaining a .bgen file and a bgen.bgi file including the imputed data for each of the sexes.

Using the step2_SPAtests.R SAIGE package, we conducted the association test separately for both sexes and 564 phenotypes, using 3 CPUs per execution. The input files for the analysis were the .bgen and .bgi files, a sample file containing phenotypic information, and the GMMAT model and variance ratio files generated in step 1. The minor allele frequency threshold was set to 0.001 and the minimum number of allele count was set in 3 in order to filter out variants. The additional --IsOutputAFinCaseCtrl=TRUE –LOCO=FALSE flags were use in the process.

3.5 Meta-analysis

We conducted a meta-analysis of summary statistics from different cohorts for the same complex trait, obtained in the previous step, using Genome-Wide Association Meta-Analysis software (GWAMA)[204] (v2.2.2). Male and female summary statistics from the same cohort were treated as independent and meta-analyzed. GWAMA was executed using the –sex flag, running gender-differentiated and gender-heterogeneity analysis and –indel_alleles, allowing to include INDELs in the analysis using 10 CPUs per task.

Two scenarios were considered in building the meta-analysis statistics: i) a given genetic variant has the same disease risk effect in both sexes; ii) a given genetic variant may have different effects in males and females, including unique sex-specific effects, effects of different magnitudes, or opposite effects. In both scenarios, we assumed that the effects were fixed across the different cohorts.

3.6 Calling of significant loci

To determine the significance of a signal found in the X chromosome, we first established a significance threshold (p-value). We estimated the p-value by correcting for the number of independent tests performed using the Bonferroni method to account for multiple testing. The number of tests was equivalent to the total number of loci or independent recombination blocks present on the X chromosome.

To calculate the number of independent tests, we performed LD pruning (r^2=0.8) on the imputed data from three cohorts in dbGaP and the imputed results from UK Biobank using PLINK software. To be as conservative as possible in the generation of our threshold, we selected the largest number of independent blocks, which was obtained by analyzing the UK Biobank data.

To validate the methodology employed in our X chromosome analysis, we carried out the same process at the whole-genome level, using the imputed array data from the GCAT project (section 2.1). We obtained the number of independent blocks by performing LD pruning (r^2=0.8) on the imputed data using PLINK software and calculated the whole-genome-wide significance threshold. We then compared this threshold to the standard threshold used in GWAS ($5x10^{-8}$) to determine if they were equal.

The chromosome X p-value threshold that we calculated was used as the overall threshold for calling significant loci in the entire project We identified four different types of loci: loci from significant variants in both sexes, loci from significant variants in males only, loci from significant variants in females only, and loci from significant variants in both sexes with different directions of effect. Loci were built taking the lead variant, the variant with the lowest p-value, and adding a window of 125,000 bp upstream and downstream from its position.

3.7 Replication of the obtained loci

The next step was to determine whether the disease-associated loci discovered (i) had been previously described and reported in the GWAS Catalog database; (ii) had been found in previous UK Biobank analyses; (iii) could be replicated between the same phenotypes from dbGaP and UK Biobank.

3.7.1 Loci discovered against the GWAS Catalog database

We used the GWAS Catalog database[205] (version 1.0, release 25-02-2021), a consistent, searchable, displayable and freely available database of SNP-trait associations.

First, we generated a list of proxies (r^2≥0.2) for each statistically significant variant present in our summary statistics using Ldlink software (https://ldlink.nci.nih.gov). Then, as GWAS Catalog is annotated on the hg38 reference genome and our analysis was done on hg19, we performed a liftover of the obtained proxies to enable the match to be feasible. The liftover process was carried out using the liftover tool available in the University of California Sant Cruz (UCSC) Genome Browser (https://genome.ucsc.edu/cgi-bin /hgLiftOver).

We then matched the proxies against the GWAS Catalog and determined whether the trait to which the variant was associated in the GWAS Catalog was the same or directly related, considering the phenotype description, to the one analyzed in our project. If the signal and the phenotype matched, we noted the locus as previously described. If not, we noted it as a new discovery.

3.7.2 UK Biobank loci vs PankUKBB project

We compared our findings in UK Biobank with the PanUKBB project[206]. This project conducted genome-wide association studies using UK Biobank on over 7,000 phenotypes, including binary and quantitative traits, generating publicly available summary statistics. The summary statistics were downloaded as explained by their developers (https://pan.ukbb.broadinstitute.org/downloads).

First, we curated the phenotypes present in the PanUKBB project's summary statistics and matched the phenotypes studied in our analysis with the binary phenotypes from PanUKBB, using their phenotype code, phenotype source (ICD10 codes, verbal interview, touchscreen surveys), and description.

To generate the list of loci from the summary statistics of the PanUKBB project, we followed the same strategy as in our project (section 3.6). Since the PanUKBB project is a whole-genome analysis project, the significance threshold correcting for multiple testing must take into account the number of tests performed, including the entire genome. Therefore, we used the standard p-value used in GWAS ($5x10^{-8}$) as the threshold. After obtaining all the significant variants from the summary statistics and creating the loci, we compared them with our findings in the UK Biobank analysis.

To compare the list of PanUKBB loci with our results, we used genomic coordinate overlap between loci to determine whether they were the same locus. This allowed us to identify the number of loci that overlapped in both projects, the number of loci found exclusively in PanUKBB, and the number of loci found only in our project. Finally, for loci present in PanUKBB but not in our discovery list, we checked whether there were differences in the number of samples analyzed in each project for that phenotype.

3.7.3 UK Biobank and dbGap discoveries replication

To determine whether there was replication between the results obtained for the same phenotype present in both the UK Biobank and dbGap cohorts, we followed several steps. Firstly, we matched the phenotypes of each database based on their description and characteristics. Next, we matched all the loci discovered in these phenotypes based on their genomic coordinates. Finally, we considered a locus to be replicated if its positional range overlapped between the two sources for the same phenotype.

4 Fine mapping and gene-mapping analyses

4.1 Fine mapping analysis

We performed a fine mapping analysis in order to obtain a credible set for each of the loci found. This process consists of generating a list of variants among which the causal variant is statistically very likely to be found. For our study, we used a threshold of a 95% probability of containing the causal variant to define the credible set of variants for each locus.

The first step consisted of creating a correlation matrix of the variants present in each of the summary statistics obtained from the meta-analysis. To do this, we used the LDstore2 software[207] (v2.0), which allows us to store and estimate linkage disequilibrium correlations. This process requires as input a master file containing (i) a file with the variants present at the locus using a 1.5 MB window downstream and upstream of the lead variant, their genomic coordinates and allele information; (ii) an imputed cohort from which to establish the correlation matrix in .bgen format and its index in .bgi format; (iii) the number of samples present in the cohort; (iv) a sample file with the identifiers of each of the samples. We used the imputed data from UK Biobank as the reference cohort for the estimation of the matrix, regardless of whether the locus came from a dbGaP analysis or from UK Biobank.

LDstore was run with –write-bcor, --read-only-bgen and –bcor-to-text commands using 48 CPUs per task. As output, a bcor file (v1.1) which consists of binary files that store SNP correlations together with information about the SNPs in the same file; and LD file, which contains SNP correlation matrices, were generated.

The second step was to create the credible sets from the files generated in the previous step for each of the loci identified using the Finemap software[207] (v1.4). This software allows us to identify causal SNPs by applying a shotgun stochastic search algorithm. The process requires as input a master file containing (i) a file with the same variants that have been used in the process of creating the correlation matrix, including the id of the variants, genomic coordinates, alleles, MAF, beta and standard error; (ii) the LD file generated in the previous step; and (iii) the number of samples. As output we obtained (i) a CONFIG file, which contained the posterior summaries for each causal configuration and each independent variant one per line ordered by the posterior probability of that configuration being the causal configuration; and (ii) a CRED file including the 95% credible sets for each causal signal in the genomic region.

4.2 Gene mapping

Based on the results obtained in the fine mapping, we carried out a gene mapping process with the purpose of linking each locus discovered to a gene. First, we filtered the different credible sets for each locus, retaining the credible set or configuration with the highest posterior probability of containing the causal variant. These configurations were ranked in the first position and could include between 1 to 5 variants.

These variants were then assigned to a gene using three different methodologies to ensure robustness, complementary, and cross-validations of the results: (i) using the g:SNPense module of gprofiler (https://biit.cs.ut.ee/gprofiler/snpense); (ii) using Variant Effect Predictor, from Ensembl (http://www.ensembl.org/Tools/VEP), with an upstream/downstream annotation distance of 50kb from the variant; (iii) checking each variant manually in Ensembl[208] with an overlap window of 50kb. In the case that the variants of one credible set mapped more than one gene, the gene that had been mapped by more variants was selected. If the number of variants was the same, the gene containing the variant with the highest individual probability of being causal was selected.

5 Preliminary sex differences analyses

During the process of identifying loci associated with the X chromosome, we found that some of them were only significant in one of the two sexes. In this analysis, we will study this phenomenon from two points of view.

5.1 Impact of heterozygous genotypes in females on association analysis.

The impact of heterozygous genotypes in females during association analysis is an important factor to consider in X chromosome studies. Unlike the autosomes, where both sexes have two copies of each chromosome, females have two copies of the X chromosome while males have only one. This is due to the fact that males inherit one X chromosome from their mother and one Y chromosome from their father, while females inherit one X chromosome from each parent. As a result, males are hemizygous for X chromosome, while females are homozygous or heterozygous depending on whether their X chromosomes carry the same or different alleles, respectively.

We investigated the impact of heterozygous genotypes in females by conducting a simulation analysis using the imputed data from the female cohort in UK Biobank. In this analysis, we replaced all heterozygous genotypes (0/1) with a missing call (./.), simulating the scenario observed in males. We then repeated the association analysis using SAIGE (as described in section

3.4.2) for each of the 564 phenotypes previously analyzed in UK Biobank. Loci were called as outlined in section 3.6, and the resulting loci were compared to those obtained in our standard analysis to evaluate the impact of heterozygous genotypes in females on the association results.

During the comparison, we evaluated (i) the disappearance of significant loci found only in females in the standard analysis; (ii) the discovery of any new loci not previously identified in the standard analysis; and (iii) the identification of any significant loci found in males but not in females in the standard analysis that were now significant in females in the new analysis.

5.2 Pathway enrichment analyses

First, we divided the genes obtained during gene mapping (section 4.2) into three lists depending on whether the locus was significant only in females, only in males or in both sexes. Next, we determined whether there were biological pathways that were enriched in these gene lists. We performed the enrichment analysis process with the g:GOST module of g:Profiler (https://biit.cs.ut.ee/gprofiler/gost). For this, we used a custom gene matrix transposed file (GMT) including only the genes present on the X chromosome and all the biological pathways present in g:Profiler database. A value of 0.05 was used as threshold and Benjamini-Hochberg false discovery rate (FDR) method was used for multiple testing correction. Biological pathways that were significant with the gene listing of one sex and not with the listing of the other, or both sexes together, were determined to be pathways that were differentially affected between sexes.
In addition, we determined if there was any relationship between the significant signals found only in one sex and the X-chromosome escape inactivation phenomenon in females. To do this, we first created a list of genes that have been reported to escape X-chromosome inactivation by different studies[162,166]. We then created an "artificial'' biological pathway in GMT format including this list of genes. Finally, we determined if this artificial pathway was enriched in this list gene more than would be expected by chance. To do this, we used g:GOST with the same parameters above mentioned.

Finally, in order to increase the power of our analysis, we decided to rerun the study including the genes obtained from the loci that met the following criteria; (i) the p-value obtained in the meta-analysis of males and females was significant (p-value $< 2.5\text{x}10^{-07}$), (ii) the p-value obtained from looking for sex differences was suggestively significant (p-value $< 1\text{x}10^{-05}$), and (iii), the p-value from the separate sex analysis was suggestively significant in one sex but not in the other (male/female p-value $< 1\text{x}10^{-05}$).

Results

1 Development of a comprehensive structural variant haplotype map from high-coverage whole-genome sequencing (GCAT|panel)

This section contains the results of the first of the two main blocks of this thesis and covers the identification and classification of germline variants (including SVs) within a population cohort of Iberian individuals, to then phase them and generate the first comprehensive haplotype reference map of the Iberian population from whole genome sequencing (WGS). This study was done in close collaboration with Dr. Valls from our group and Dr. Galván member of Rafael de Cid's group at the Germans Trias i Pujol Research Institute (IGTP), which whom I share co-first authorship in the resulting article[1].

In brief, I here describe the main results of this study, focusing on the parts where I most contributed. These parts, as explained in thesis trajectory block, include the design, benchmark, and development of a comprehensive methodology for the calling of all types of germline variants from whole genome; followed by the generation of the final genotypes; and the final construction of the haplotype reference panel. Each one of these steps has several rounds of validation and testing.

1.1 Generation of a comprehensive variant identification and filter strategies

We have designed, evaluated, and implemented a comprehensive strategy for detecting, classifying, and genotyping a broad range of germline variants from short-read Illumina sequencing data. Our approach particularly emphasizes identifying and sub-classifying SVs, including deletions, insertions, inversions, duplications, translocations, and mobile element insertions (MEIs).

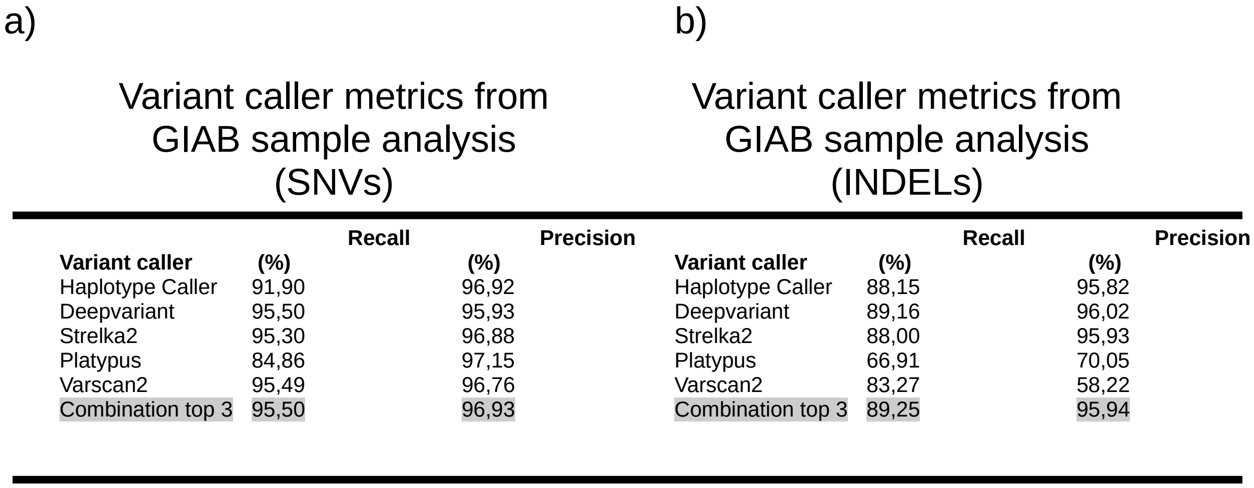
As outlined in the methods section, in order to select the most suitable variant callers for our study, we conducted several comparative analyses to assess their accuracy and recall in detecting and genotyping germline variants. We then integrated a logistic regression model (LRM) to combine the results of the variant callers, maximizing both precision and recall.

During this process we employed two reference samples to evaluate the performance of the variant callers. The first sample was an *in silico* sample containing a controlled set of variants, while the second was the GIAB Consortium sample NA12878, which has a set of validated variants.

1.1.1 Benchmark of variant callers detecting SNVs and INDELs

We carried out a thorough evaluation of five different variant calling tools (methods section 1.2.1) to identify the most reliable and efficient combinations. Our assessment was based on several metrics, including precision and sensitivity, as well as computational requirements. For this, we executed the five tools on the GIAB sample and calculated their corresponding precision and recall. After analyzing these results, we selected the three top performers, as to call efficiency and better use of computing resources. We excluded Platypus and Varscan2 software from our selection as they demonstrated the lowest performance during our benchmarking analysis.

After identifying the most appropriate variant callers to use, we devised a strategy by combining the results of all three software. This method was designed to consider a variant as valid only if it was detected by at least two of the three selected software. The application of this filter allowed us to minimize the number of false positives and enhance the precision of our variant calls (Table 4).



a)

Variant caller metrics from GIAB sample analysis (SNVs)

Variant caller	Recall (%)	Precision (%)
Haplotype Caller	91,90	96,92
Deepvariant	95,50	95,93
Strelka2	95,30	96,88
Platypus	84,86	97,15
Varscan2	95,49	96,76
Combination top 3	95,50	96,93

b)

Variant caller metrics from GIAB sample analysis (INDELs)

Variant caller	Recall (%)	Precision (%)
Haplotype Caller	88,15	95,82
Deepvariant	89,16	96,02
Strelka2	88,00	95,93
Platypus	66,91	70,05
Varscan2	83,27	58,22
Combination top 3	89,25	95,94

Table 4. SNVs and INDELs benchmark using different variant callers and combination strategies. a) Variant calling metrics in SNVs detection using Genome in a Bottle Sample (GIAB) as a gold standard. **b)** Variant calling metrics in INDELs detection using Genome in a Bottle Sample (GIAB) as a gold standard

1.1.2 Benchmark of variant callers detecting SVs

Whereas the detection of SNVs and INDELs is fairly solved, the detection of SVs proved to be a more challenging task due to the nature of the variant, the difficulty in mapping affected reads, and the limitation of callers to detect those. Our analysis showed that SVs were detected by different combinations of variant callers as shown in Figure 10, making it essential to combine the results of different tools for optimal results.

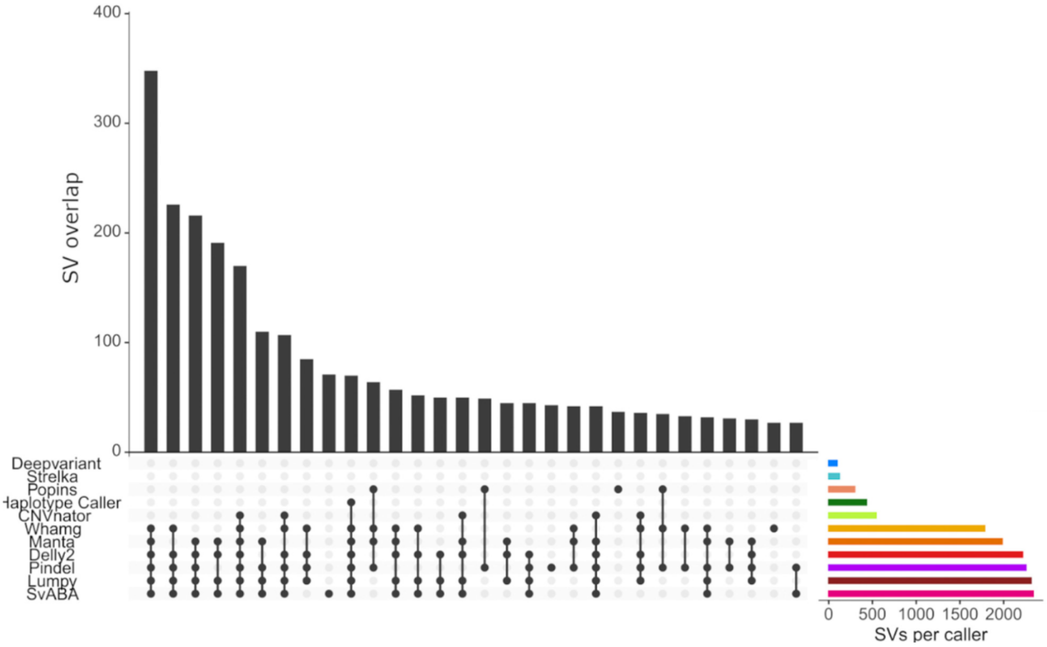


Figure 10. Structural variant detection patterns according to the programs used. Detections were estimated using the *in silico* sample as gold standard. Only the first 30 patterns with more coincident SV calls are shown. Lines and dots indicate the programs used and bars the number of overlapping calls resulting from that combination. Right colored horizontal bars indicate the total number of SVs detected by each caller.

Initially, we evaluated the performance of 11 variant callers, but two of them, Pamir and Genome Strip, were excluded due to computational difficulties during their execution. In order to maximize the detection of SVs, we designed a LRM that combined the results of the remaining nine variant callers. The LRM approach was utilized to exploit the individual strengths of each variant caller and filter out results based on a score derived from multiple parameters, including the precision and recall of each variant caller, the size and type of variant detected, and the number of software that detected the break-point. By integrating the results of multiple variant callers through this LRM, we were able to significantly improve the overall accuracy and precision of our SV detection (F-score=0.9) compared to other filtering methods which consisted of considering a variant as valid when it was detected by at least two (F-score = 0.83), three (F-score = 0.79), or four (F-score =0.72) variant callers. (Figure 11a). Additionally, we estimated the F-score for each strategy based on the variant size, observing that the LRM remained the strategy with a higher F-score in all size windows (Figure 11b).

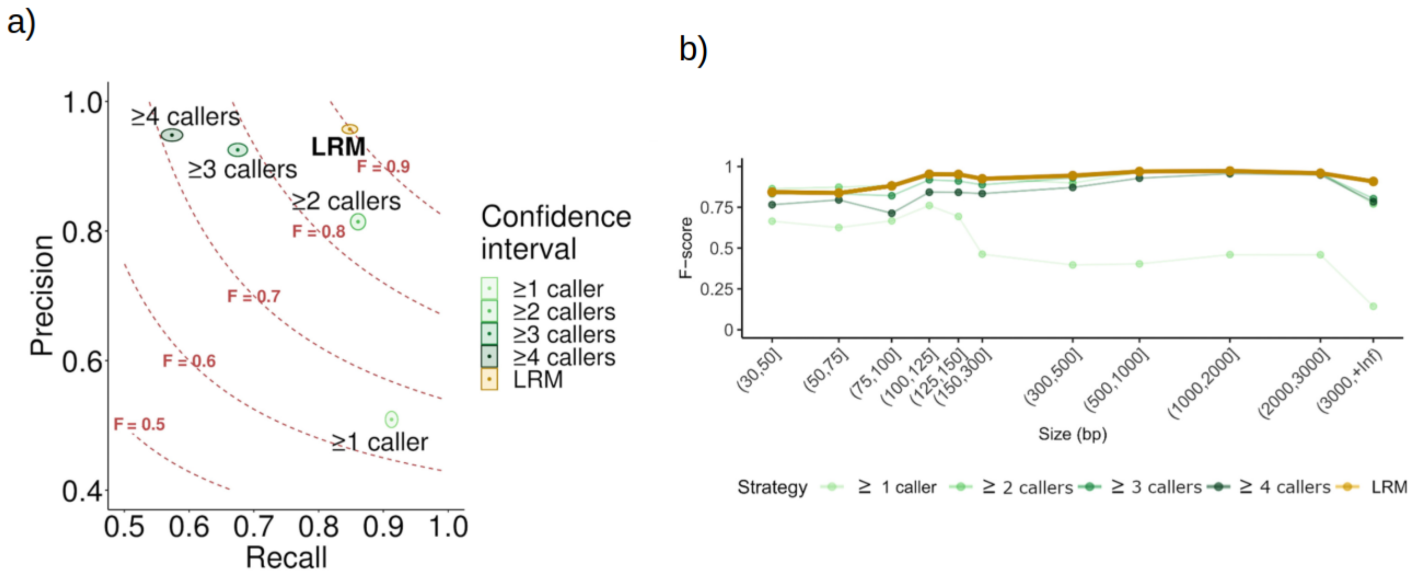


Figure 11. Overview of the detection performance of different strategies and filtering results from multiple SV callers. a) Each strategy is plotted according to the recall and precision ratios (F = F-score) using the *in silico* sample as benchmarking dataset. The LRM, with a F-score of 0.9, outperformed other commonly used strategies that are based on the number of coincident callers. The confidence interval for each case is represented by colored area of each strategy. **b)** Comparison of performances (F-score) of different merging and filtering strategies according to SV size.

1.1.3 Benchmark of the genotyping process

Having high genotyping accuracy is crucial when creating a haplotype reference panel because it directly impacts the quality and reliability of the results obtained from subsequent genetic analyses. With the aim to achieving high genotyping performance, we evaluated the genotyping accuracy of each variant caller and the LRM combination approach. To achieve this, we employed the *in silico* sample as a reference, which comprises variants with established genotypes, and contrasted the genotype calls from each software against those found in the sample. While the genotyping process for SNVs and INDELs exhibited a low error rate (≈ 5%), this was not the case for genotyping SVs (Figure 12a). In the case of SVs, once the performance of each independent variant caller in the genotyping process was established, we integrated these results into the previously designed LRM. In this way, the model prioritizes and combine variant callers with higher individual accuracy in determining the correct genotype. This approach resulted in a reduced genotyping error rate (5%) compared to other filtering methods, which involved considering a genotype as valid if it was called by at least two (error = 14%), three (error = 12%), or four (error = 11%) variant callers (Figure 12b).

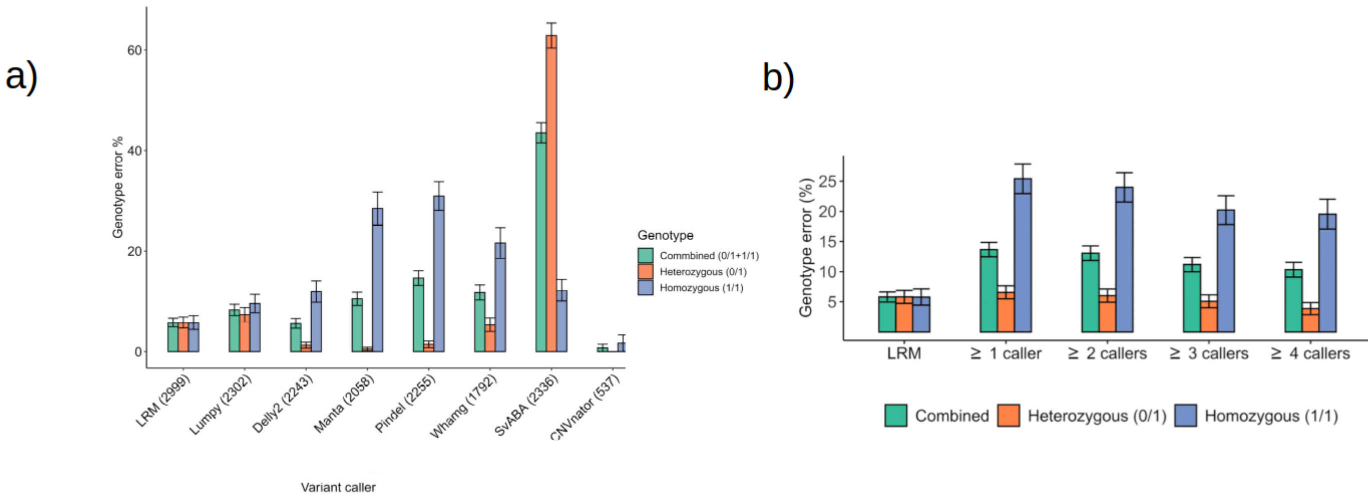


Figure 12. Comparative overview of the genotyping accuracy. a) Genotype error of each individual SV caller. **b)** Genotype error from the LRM and strategies based on the number of coincident callers. LRM showed a lower genotype error, outperforming the number of coincident callers' strategies.

1.2 Evaluation of the GCAT cohort data quality

The development of the haplotype reference panel was a collaborative effort with the GCAT-Genomes for life project. The project provided us with two types of genetic data: 5,489 genotyping samples and, 808 Illumina whole-genome sequencing (WGS) samples at high coverage (30X), from which we generated the final haplotype map. The GCAT cohort was constructed by gathering volunteers from different regions in Catalonia and ensuring a balanced representation of genders.

To obtain a high quality and homogeneous cohort, we first performed a quality control of the BAM files, resulting in the exclusion of one sample as it failed to meet the alignment quality control metrics (methods section 1.2.1). Next, we conducted a principal component analysis (PCA) to remove samples from populations genetically distant from the Iberian population. In this filtering process, we removed 18 samples identified as belonging to another population (Figure 13a and 13b). To avoid including samples with familial relatedness that could cause deviations in the subsequent analysis, we performed an Identity by Descent (IBD) test. We identified two pairs of related samples, one pair of siblings and one pair of first cousins and eliminated the sample with lowest call rate from each pair (Figure 13c). Lastly, as a final analysis to ensure a homogeneous cohort, we conducted a final PCA on the remaining samples and removed two samples using the mean ± 4sd criteria (Figure 13d).

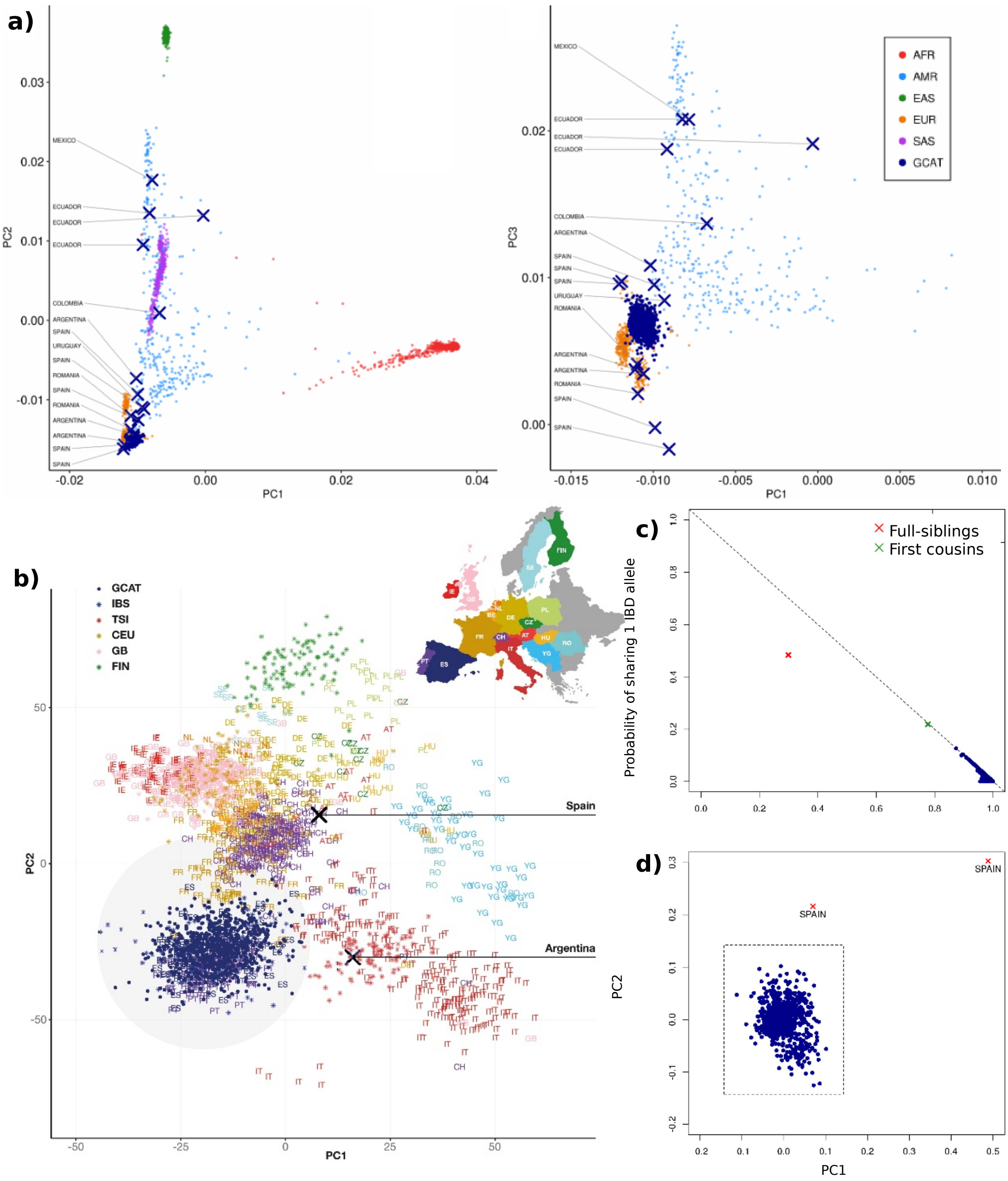


Figure 13. Evaluation of GCAT cohort data quality. Samples labeled with X were discarded samples**. a)** Principal component analysis (PC1, PC2, PC3) of the GCAT samples against 1000 Genomes. **b)** Principal component analysis (PC1, PC2) of the GCAT samples against PROPES project. **c)** Identity by Descent plot. **d)** Principal component analysis to test GCAT cohort homogeneity.

1.3 Description of the GCAT variant set

Following the application of quality control measures to the GCAT cohort, we retained for analysis 785 WGS samples, on which we applied our variant calling strategy explained above. As a reminder, this strategy encompassed a multi-variant calling process designed to enable the detection of SNVs, INDELs, and SVs, along with the use of LRMs to filter out false calls and enhance the accuracy of the variant calls while simultaneously reducing the genotyping error.

By executing this strategy, we obtained a final file containing the set of variants after applying the filtering rules of the LRM model (variants were considered positive when the score was > 0.5), along with the genotype of each sample for each variant contained in the set. (Table 5)

Variant type	RAW set (n)	Final set (n)
SNVs	58,529,907	30,325,064
INDELs	10,452,204	5,017,199
DEL	1,359,594	33,244
DUP	674,817	6,269
INS	352,939	12,782
INV	228,091	10,115
MEI	170,735	18,779
TRA	117,048	7,989

Table 5. Total count of variants for each variant type in both the raw and final sets. The raw set column includes the number of variants detected for each type without applying the filtering strategy. The final set column contains the number of variants of each type after applying the LRM to keep those variants considered as valid (score > 0.5).

Upon application of the logistic regression model filtering process, the final set was composed of a total of 35,431,441 variants, including 30,325,064 SNVs, 5,017,199 INDELs, and 89,178 SVs (Figure 14). SVs were found to exhibit a broad range of sizes, with an average size of 291 base pairs and a diverse distribution depending on the specific type of variant (Figure 15). Moreover, in terms of population frequency, these variants showed a varying distribution, with 21% of them being classified as common variants (MAF > 0.05), 10.12% as low-frequency variants (MAF > 0.01), and 69% as rare variants (MAF < 0.01). Among these rare variants, singletons (MAF < 0.0125) accounted for 42.3% of the total, while doubletons (MAF < 0.025) represented 7.88% of the total number of variants.

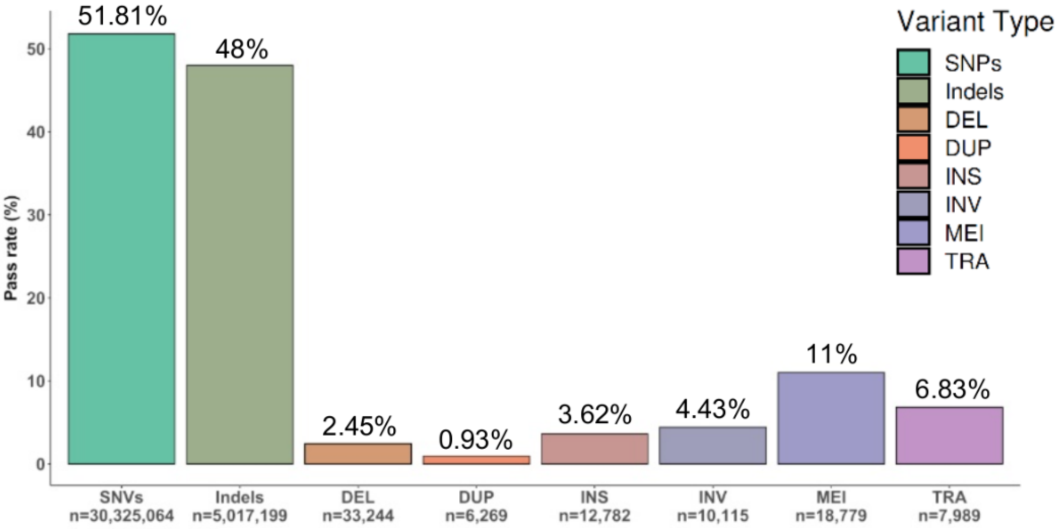


Figure 14. Pass rate of detected variants by variant type. Each bar represents a variant type. The lower part of the bar shows the total number of valid variants. The upper part of each bar shows the percentage of variants that have been accepted against the total number of variants detected (pass rate).

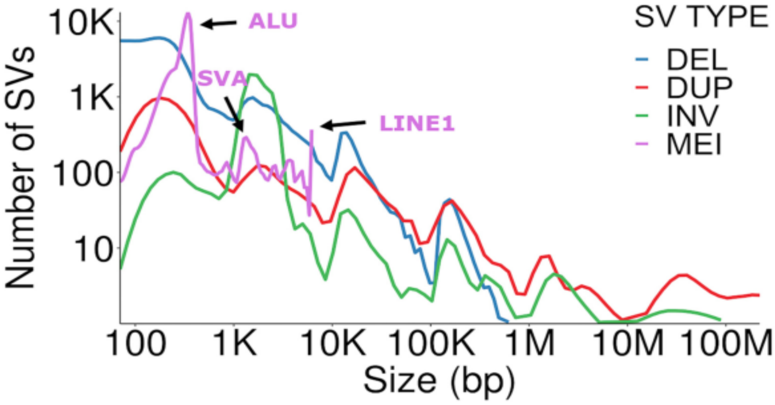


Figure 15. Distribution of SV type according to their genomic sizes. MEI are mainly found between 100bp and 10kb. Deletions are similarly distributed from 1kb to 10kb in size, with the number of variants decreasing with increasing size, reaching a maximum of almost 1Mbp.

The results of our study aligned with what has been previously reported in other large-scale whole-genome sequencing projects. Although we cannot discard that this distribution derives from the methodology used, we found that deletions were the most frequent type of SVs, accounting for 37.3% of all variants, followed by MEIs at 21.1%, insertions at 14.3%, inversions at 11.4%, translocations at 8.8%, and duplications at 7.1%. (Figure 16)

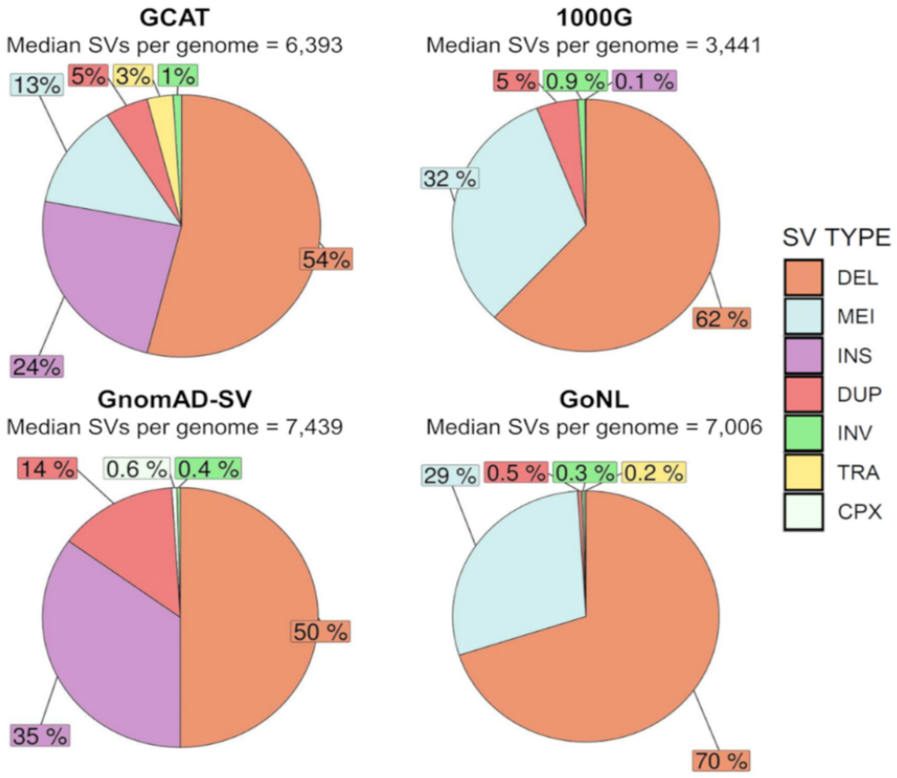


Figure 16. Comparative overview of the SV type number and distribution across the GCAT, 1000G, GnomAD and GoNL catalogs. The average number of SVs per genome is similar between the GCAT (30x coverage), GnomAD-SV (30x coverage) and GoNL (11-13x coverage) projects. This number is lower in the 1000G project (3-5x coverage). In reference to the type of variant detected, the percentage of insertions and deletions is similar to that of GnomAD-SV, while the difference in the percentage of MEI observed is due to the fact that GnomAD-SV does not include this type of variant in its analysis.

Regarding the number of variants per genome, we identified an average of 3.52M SNVs (SD=24,983), 606,336 INDELs (SD=8,060) and 6,393 SVs (SD =222) per individual, demonstrating consistency across the cohort. Regarding the distribution of allele frequency in an individual genome, the most represented were common variants at 85.28%, followed by low-frequency variants at 8.43% and rare variants at 5.21%.

1.4 Validation of the GCAT variant set

To validate the variants detected, as well as the number of new variants contributed by our study, we performed comparative and experimental approaches.

We employed two different validation strategies to ensure the accuracy and reliability of SNVs and INDELs. Firstly, we cross-checked our variant calls with those obtained in the genotyping array analysis (n= $2x10^6$ variants) of a subset of 570 samples also included in our initial 785. By comparing the results of our variant calls to the genotyping array data, we were able to confirm the presence of the variants in both datasets, as well as determine the concordance of genotypes for each variant. The comparison showed that our method had a detection concordance of 96% and 87% for SNVs and INDELs, respectively. Moreover, our method displayed high genotyping accuracy, with 97% for SNVs and 96% for INDELs.

For our second validation strategy, we conducted a comparison of the SNVs and INDELs included in the GCAT variant set (35.3M) with those present in the dbSNP database (Build 153.v). We performed this comparison by matching variants based on their chromosome, position, alternative allele, and reference allele. This comparison revealed that 19.18% (≈ 6,78M) of our SNVs and INDELs were previously unclassified in dbSNP. Among these unclassified variants, 84.32% were rare variants. Singletons and doubletons were the most frequent, representing 34.25% and 22.32%, respectively. Interestingly, we found that 12.84% of common SNVs and INDELs were not yet included in dbSNP, potentially due to population-specific variants. (Figure 17a)

For the validation of the accuracy and reliability of SVs calls, although it is particularly challenging, we applied two specific experimental validation strategies targeting deletions, duplications, and inversions. First, we employed a Comparative Genomic Hybridization Array (CGH) methodology to analyze the accuracy in detecting large deletions and duplications (>20kb), which are among the most challenging variant types to detect using short reads. We performed this analysis by randomly selecting five samples from our cohort and using the NA1287 sample as a reference sample, as its list of deletions and CNVs has been reported. This CGH analysis revealed that we were able to validate 76% of the detected deletions and 19% of the duplications in the five selected samples.

Regarding inversions, we compared our list of inversions to the recent benchmark dataset of the InvFEST[193] project, which includes 59 inversions that were experimentally validated in the European population. Our analysis showed that 84% of the inversions present in InvFEST were also present in the GCAT variant set, matching in both allele frequency and size.

In order to study in more depth the contribution of our set of variants, we conducted a comparative analysis of the GCAT SVs set against various public databases, including gnomAD-SV, 1000G, GoNL, HGSVC, DGV, and dbVar. This analysis allowed us to determine the number of previously identified variants in other databases and the number of new variants provided by the GCAT, based on variant type and population frequency.

Our results showed that 61% (49,333) of the SVs in the GCAT set were new. Deletions (n=32,697) had the lowest proportion of novel variants, with a total of 13,057 (39.9%) previously unclassified variants. On the other hand, inversions (n=10,116) had the highest proportion of novel variants, with a total of 9,652 (88.7%) new variants. (Figure 17b)

In terms of population frequency, the vast majority (88.3%) of the novel variants were rare (MAF < 0.01). As expected, the percentage of novel variants with common frequency (MAF > 0.05) was the lowest in the comparison. (Figure 17c)

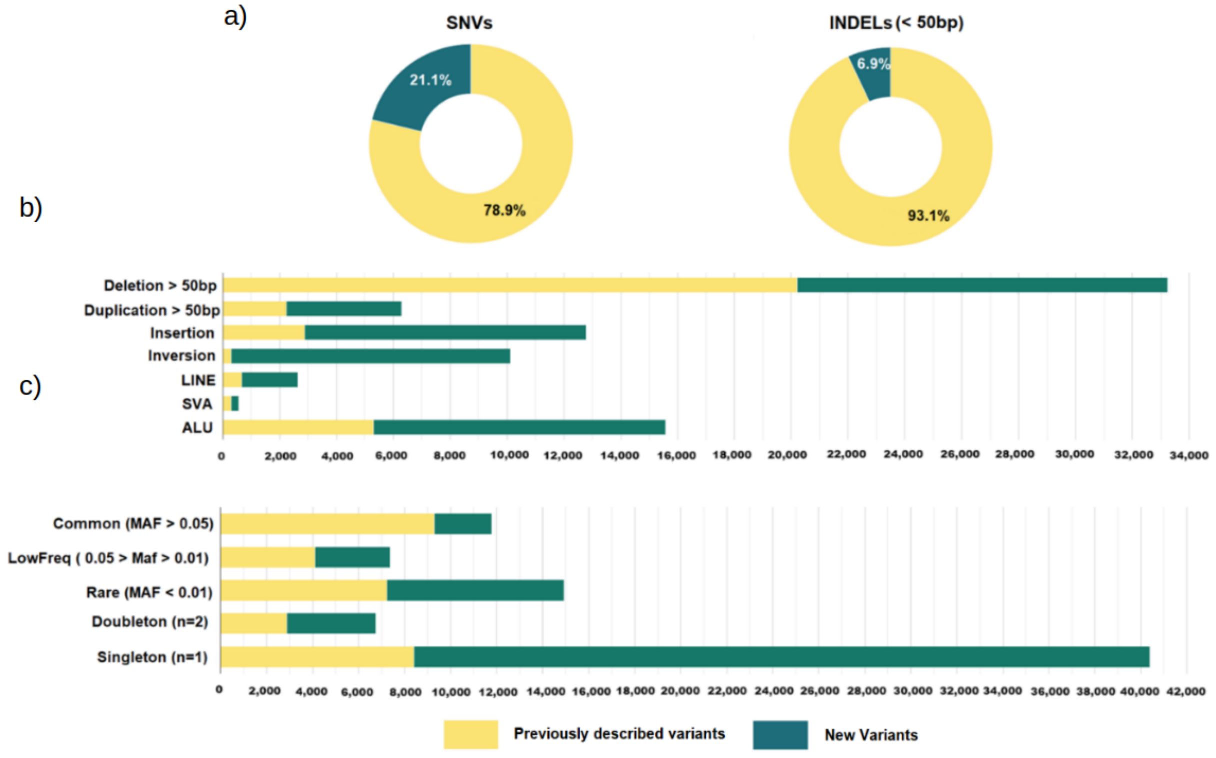


Figure 17. GCAT variant contribution in comparison to popular datasets. **a)** Proportion of new and described SNVs and INDELs against dbSNP (v.153). Variants were matched considering both genomic coordinates and alleles. **b)** Proportion of new and previously described SVs against popular datasets distributed by SV type. Variants were matched by genomic coordinates considering a breakpoint error of ±1000 bp, SVs type and if the size was available, 80% of reciprocal overlap. **c)** Proportion of new and previously described SVs compared to popular datasets distributed by variant frequency.

Finally, in order to determine the potential to provide new variants in the imputation process, and therefore the value of the GCAT SV set as a reference panel, we compared the list of SVs in the GCAT set against the 1000genomes and GoNL reference panels. This comparison showed that 6,523 common variants (MAF > 0.05), 5,913 low-frequency variants (0.01 > MAF > 0.05), and 63,722 rare variants (MAF < 0.01) included in the GCAT set had not been previously described by any of the reference panels used in the comparison. This represents a contribution of 76,158 new SVs to reference panels ready for imputation.

1.5 Development of phasing and imputation strategies

With all the final genotypes generated for all the 785 individuals, we next defined haplotypes by phasing all the variants. To determine the most effective phasing strategy, we established a cross-validation framework that relied on imputation results as the evaluation metric, by focusing on identifying the phasing strategy that would generate a haplotype panel capable of imputing the highest number of variants with a high imputation quality (info score ≥ 0.7).

In order to identify the optimal phasing strategy for the Iberian-GCAT catalog, we generated several test reference panels of chromosome 22, including SNVs, INDELs, and SVs. To achieve this, we applied different phasing strategies, including Shapeit2, Shapeit2+MVNcall, Shapeit2+PIRs+MVNcall, Shapeit4, Shapeit4+MVNcall, and Shapeit4+WhatsHap.

After evaluating all the mentioned strategies, we determined that the combination of Shapeit4 and WhatHap, which employs phase informative reads (PIRs), was the optimal approach for phasing SVs. This resulted in a high-quality haplotype panel that enabled the imputation of 98% of common SNVs, 92% of INDELs, and 90% of SVs observed in the variant calling of the same samples imputed.

1.6 Imputation performance and comparison against main reference panels

Using the phasing and imputation approaches mentioned, we were able to generate a complete and operational panel of Iberian haplotypes that includes all variants detected in our 785 individuals.

To assess the performance and benefits of the GCAT haplotype reference panel in enhancing genetic association studies, particularly in the case of SVs, we conducted a preliminary study. First, we selected 95 samples from the reference panel that also had genotyping data available. Then, we removed these 95 samples from the GCAT haplotype reference panel, obtaining a haplotype reference panel with 690 samples. Next, we imputed the genotyping data of the 95 removed samples using the 690-sample panel. Finally, we compared the imputation results of the 95 samples with the variant calling data previously obtained for the same samples.

With this strategy, the median number of imputed SVs was 5,120 (SD = 50), from a maximum of 6,393 SVs estimated per individual in the variant calling, resulting in a 80% of imputation rate (variants imputed / variants called). Our analysis showed a high level of genotyping concordance for common imputed variants, with 99% (SD = 0.4) for SNVs, 97% (SD = 0.6) for INDELs, and 98% (SD = 1.2) for SVs. (Figure 18a).

Additionally, to evaluate the imputation capability of the whole GCAT reference panel in the context of other similar resources, we compared it against several other reference panels, including 1000G, GoNL, HRC, and UK10K. To achieve this, we imputed the GCAT genotyping array with the different haplotype reference panels and compared the results of the imputation.

Our analysis revealed that the GCAT panel was able to impute a total of 14,383,907 high-quality variants (info score ≥ 0.7) and MAF > 0.001. As shown in figure 18b, in general, all the reference panels performed well in imputing SNVs and INDELs, with the GCAT panel demonstrating a particularly high ability to impute rare INDELs. However, for SNVs, 1000Genomes and HRC demonstrated the highest performance, imputing the greatest number of variants, including rare SNVs.

Regarding SVs, the GCAT panel was able to impute a total of 23,179 high-quality variants. Our comparison against other reference panels containing SVs, 1000G and GoNL, showed that the GCAT panel outperformed both panels, resulting in a 1.6, 2.7, and 1.3-fold increase in the number of high-quality imputed SVs compared to 1000g, GoNL, and both panels combined, respectively (Figure 18c).

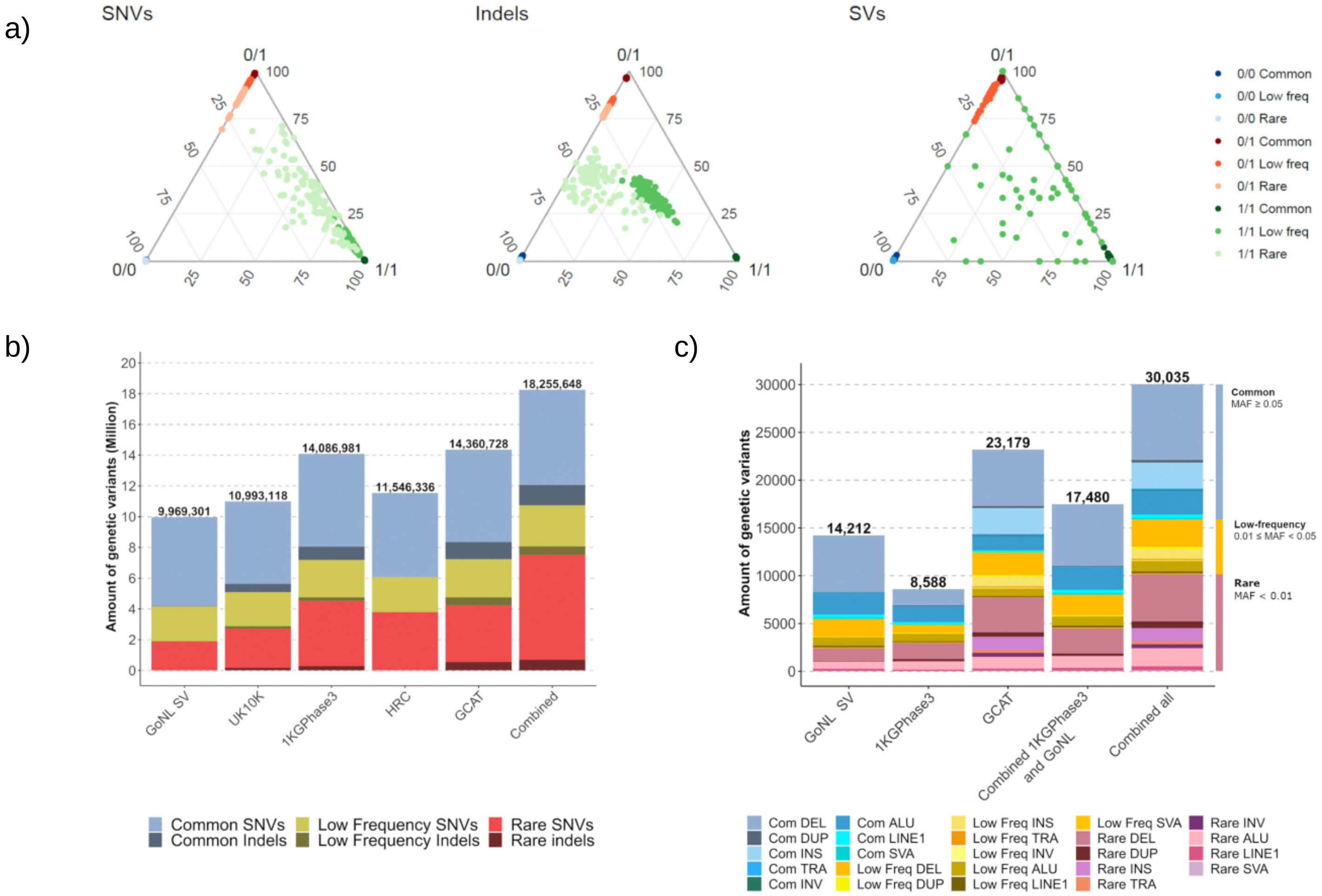


Figure 18. Imputation performance of the GCAT reference panel. a) Ternary diagram showing the accuracy of genotype imputation by variant type and frequency, using genotype calling as the reference. Each genotype state per sample is evaluated using three dots. Samples with high concordances between genotype imputation and genotype calling are located at the vertices of the diagram. **b)** Bar graph showing the number and frequency of imputed SNVs and INDELs (info score ≥ 0.7) using different reference panels and combined approaches. **c)** Bar graph showing the number and frequency of imputed SVs (info score ≥ 0.7) using different reference panels and combined approaches.

We further assessed the efficacy of the GCAT reference panel in imputing SVs in various ethnic groups. To achieve this, we employed the publicly available 1000G genotyping array and selected 1880 individuals from 19 distinct ethnic groups, grouped into four unique regions as defined in the 1000G data: Europe, Asia, Latin America, and Africa.

Our results showed that while the European population had the highest number of imputed SVs, as expected, the GCAT panel was also capable of imputing a significant fraction of SVs in other ethnicities (Figure 19). Notably, the Latin American population had the highest number of imputed variants compared to Asia and Africa, largely comprised of low-frequency variants (MAF < 0.05) due to its mixed origins. In contrast, most of the imputed variants in the other populations were common (MAF > 0.05). The African population had the highest number of imputed common variants, while the Asian population had the lowest number of imputed variants. In general, the results demonstrate that although the GCAT haplotype reference panel is composed of European samples, it exhibits a high imputation capability in more distant populations, making it a very valuable tool in imputation studies, regardless of the population origin of the cohort.

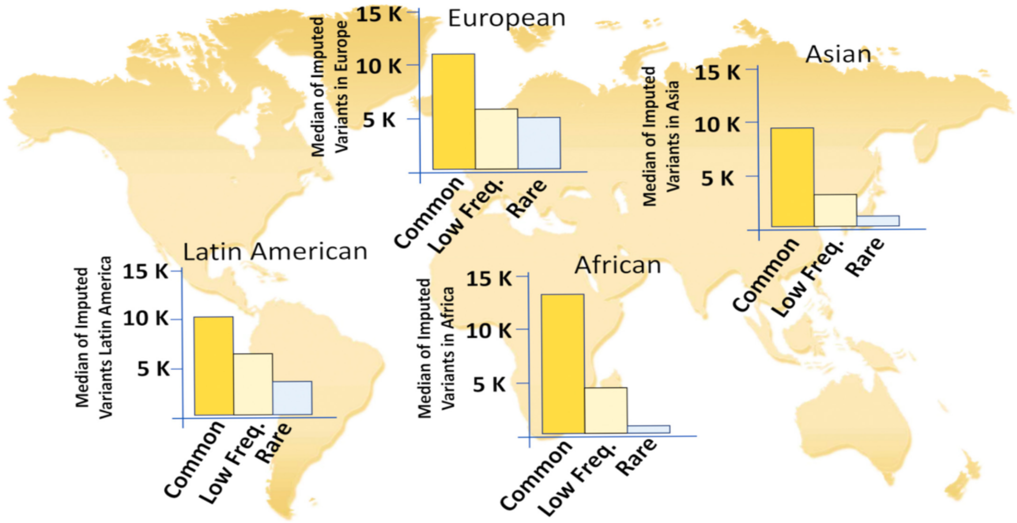


Figure 19. SVs imputation performance of the GCAT reference panel across different ethnic groups. Consistent with expectations, the European population had the greatest number of imputed variants. The Latin American population had the second-highest number of imputed variants, with a significant proportion of low-frequency and rare variants. The African population showed the highest number of imputed common variants. In contrast, the Asian population had the lowest number of imputed variants, reflecting its greater genetic distance from the Iberian population.

1.7 Preliminary association tests

Finally, using the GCAT array (n=4,448), we collaborated with Natalia Blay, a member of Rafael de Cid's group at the Germans Trias i Pujol Research Institute (IGTP), to conduct a preliminary genome-wide association study (GWAS) across 70 chronic conditions (number of cases > 50). First, we performed the imputation process with the generated reference panel and kept variants with an imputation quality info score > 0.7. Then we performed the association step obtaining 46 SV loci with a suggestive association after Bonferroni correction (p-value < 1 x 10^{-6}). The most remarkable finding was a rare AluYa5-element in chr3 (g.49494276 49494600ins (hs37d5), MAF = 0.0013), located near the dystroglycan gene (DAG1) and associated (P-value = 9.84×10^{-7}) with Mononeuritis of lower limb (ICD-9 355) (Figure 20a). This variant was imputed only with the GCAT reference panel (info score = 0,98) and experimentally confirmed by Natalia Blay in all carrier individuals through PCR analysis (Figure 20b).

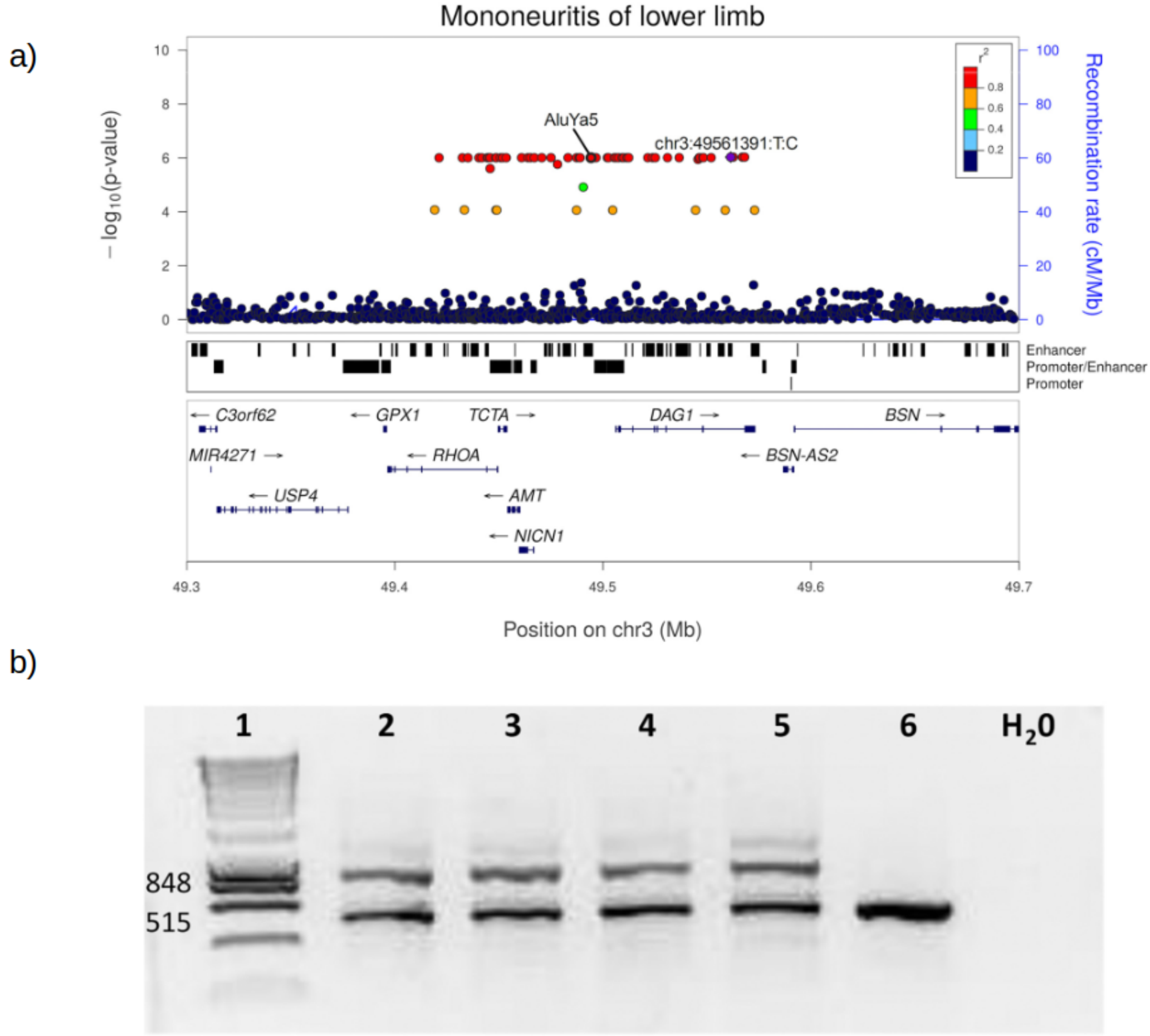


Figure 20. Genome-wide association analysis using GCAT reference panel and experimental validation of the AluYa5-element. a) Locus zoom plot of the signal associated with mononeuritis of lower limb (ICD-9 355) (p-value = 9.84×10^{-7}), showing the lead variant in purple. **b)** Experimental validation of the AluYa5-element, agarose e-gel electrophoresis of PCR products after amplification of Alu-insertion-specific DNA fragments from blood DNA. Column 1: 100 bp DNA ladder marker (Life Technologies). Columns 2–5: Alu carriers (EGA 04200, EGA 01901, EGA 13378, EGA 03940). Column 6: control sample (EGA 01399). Electrophoresis analysis of Alu carriers show two-band amplicons (515 bp and 848 bp) detected in Alu carriers (lanes 2–5) and one-band amplicon (515 bp) in control non-Alu allele individuals (lane 6).

2 Recovering the role of the X chromosome in complex diseases

This section corresponds to the second part of my thesis, where we aim to recover and study the role of the X chromosome in complex diseases, which has been mainly neglected due to methodological issues. This study was conducted alongside Dra. Salvoro, with whom I am collaborating on the preparation of an article that includes the results explained in this part.

In summary, this section will present the main findings, which include data selection and screening for the study, the development of a strategy for chromosome X-wide association analysis (XWAS), and the comparison of results against public databases and previous analyses. In this XWAS strategy, we will use, among others, the haplotype map generated above in order to capture a wider range of variability, mostly through SVs and INDELs. Additionally, this section features a preliminary functional analysis through fine-mapping and gene-mapping, enabling us to conduct a pathway enrichment analysis with the goal of identifying differences between sexes.

2.1 Data collection and preliminary cohort screening

The data collection and preparation processes are essential steps in an association study. Although the size of the cohort used is an important factor that can affect the ability to detect associations, there are other characteristics, such as the recruitment strategy of the samples analyzed, that can also play a crucial role. This is why we have made a deliberate effort to incorporate cohorts from different sources and different recruitment strategies in our project to avoid such biases.

The data used in this study were obtained from two primary sources: dbGaP (https://www.ncbi.nlm.nih.gov/gap/) and UK Biobank (https://www.ukbiobank.ac.uk/). As explained in the methods section, dbGaP contains genomic and phenotypic data from disease-specific studies that have been diagnosed, while UK Biobank is a large-scale population-based study that includes more than 500,000 participants and obtains phenotypic information through surveys, diagnoses, or self-reporting by individuals.

By incorporating a range of cohorts with varying recruitment methods, we aimed to ensure that our results are robust and reliable and guarantee that are not affected by inherent biases in recruitment strategies.

2.1.1 dbGaP cohorts

In order to encompass a wide range of phenotypes, we initially chose 169 dbGaP cohorts, covering 90 distinct binary traits. However, as cohorts were produced by different research centers and could contain varying information, we retained only those cohorts that fulfilled specific criteria. For example, we ensured that the genetic data contained the X chromosome, and that the phenotypic data were available. Cohorts that did not meet these criteria were excluded.

In addition, because a large sample size is crucial for obtaining reliable results in the association process, we determined that each phenotype should have a combined sample size of over 4,000 individuals, including around 1,000 cases, from all cohorts covering that phenotype.

After carefully scrutinizing all the cohorts, we selected a total of 46 dbGaP cohorts covering 36 complex traits that met these requirements and were deemed suitable for our subsequent analysis (Table 6).

Phenotype	Cohorts	n Cases	n Controls	n total
Alcohol dependency	phs000091; phs000125; phs000425	3,980	2,710	6,690
Allergic Rhinitis	phs000788	13,828	38,080	51,908
Alzheimer	phs000168; phs000219; phs000372; phs000496	5,009	3,982	8,991
Amyotrophic Lateral Sclerosis	phs000788	9,143	1,939	11,082
Asthma	phs000788	9,192	38,080	47,272
Bladder Cancer	phs000366	5,682	4,268	9,950
Breast Cancer	phs000147; phs000812; phs000929; phs000975; phs000305; phs001265	40,169	38,157	78,326
Cardiovascular disease	phs000788	14,997	10,457	25,454
Colorectal cancer	phs001315; phs001415; phs001856	22,834	21,328	44,162
Depression	phs000788	7,244	43,817	51,061
Dermatophytosis	phs000788	7,656	32,255	39,911
Type 2 Diabetes	phs000788	6,957	16,565	23,522
Dyslipidemia	phs000788	30,204	13,770	43,974
Gastrointestinal cancer	phs000361	3,316	1,948	5,264
Glioma	phs000652; phs001319	5,261	6,455	11,716
Hemorrhoids	phs000788	9,076	47,303	56,379
Hypertension	phs000788	28,357	10,941	39,298
Hernia	phs000788	6,274	28,425	34,699
Insomnia	phs000788	3,959	40,330	44,289
Iron deficiency	phs000788	2,444	53,935	56,379
Irritable bowl	phs000788	3,066	43,732	46,798
Lung cancer	phs00093; phs000336; phs000629; phs000716; phs000753; phs001210	15,316	15,459	30,775
Lymphoma	phs000801	3,627	1,485	5,112
Macular degeneration	phs000788; phs001039	20,778	35,216	55,994
Osteoarthritis	phs000788	20,206	9,809	30,015
Osteoporosis	phs000788	5,404	33,555	38,959
Pancreatic cancer	phs000206; phs000648	8,562	7,021	15,583
Parkinson	phs000126; phs000196; phs000918; phs001172	8,963	9,188	18,151
Pharynx cancer	phs001202	5,534	3,578	9,112
Prostate cancer	phs000207; phs000306; phs000487; phs00733;	23,426	21,599	45,025

	phs000812; phs000838; phs000882; phs001120			
Renall cell carcinoma	phs000351	1,166	3,219	4,385
Peripheral vascular disease	phs000788	4,321	11,054	15,375
Peptic ulcers	phs000788	908	55,471	56,379
Psychiatric diseases	phs000788	8,583	41,644	50,227
Stress	phs000788	4,276	38,969	43,245
Varicose Veins	phs000788	2,476	53,903	56,379

Table 6. List of phenotypes included in the selected cohorts in dbGaP after quality control. Table includes ID of the cohorts covering each phenotype, number of cases, number of controls and total number of samples.

2.1.2 UK Biobank

UK Biobank initially consisted of a total of 7,000 phenotypes which were obtained from various sources including electronic health record data from national health registries, self-reported annotations through touch screen questionnaires or face-to-face interviews and medical examinations.

First, we decided to limit our study to binary case-control phenotypes, removing all continuous phenotypes from the dataset. Next, with the aim to reduce biases due to unbalanced case-control ratios, as well as trying to avoid including phenotypes for which we do not have enough statistical power to detect a signal, we filtered the remaining list of binary traits by excluding those with fewer than 1,000 cases.

Then, to ensure the least possible contamination of risk alleles within controls, we meticulously selected cases and controls within the UK Biobank cohort. We carefully cleaned controls eliminating samples that were cases for other diseases associated with the case definition. This meticulous selection aimed to prevent the inclusion of individuals categorized as controls but presenting a similar trait to the cases, which can lead to masking of signals.

After this thorough cleaning and screening process, a total of 564 binary traits distributed in 19 disease groups were deemed suitable for inclusion in our analysis. (Figure 21)

Lastly, we performed a comparative check of the phenotypes from UK Biobank to determine the total number of unique phenotypes that would continue for downstream analyses. In this way, we determined if there were duplicate or overlapping phenotypes within the cohort based on the phenotype description. This task was key and posed a challenge due to the differences in data collection protocols, phenotype definitions, and study inclusion criteria.

This observation led us to estimate that our study collectively, including dbGaP and UK Biobank, encompassed a total of 477 unique binary traits. Regarding overlapping phenotypes between dbGaP and the UK Biobank, we chose to treat them as independent during the subsequent analyses given the distinct nature of their recruitment processes. This approach ensures that potential discrepancies between cohorts do not compromise the overall validity and consistency of our findings.

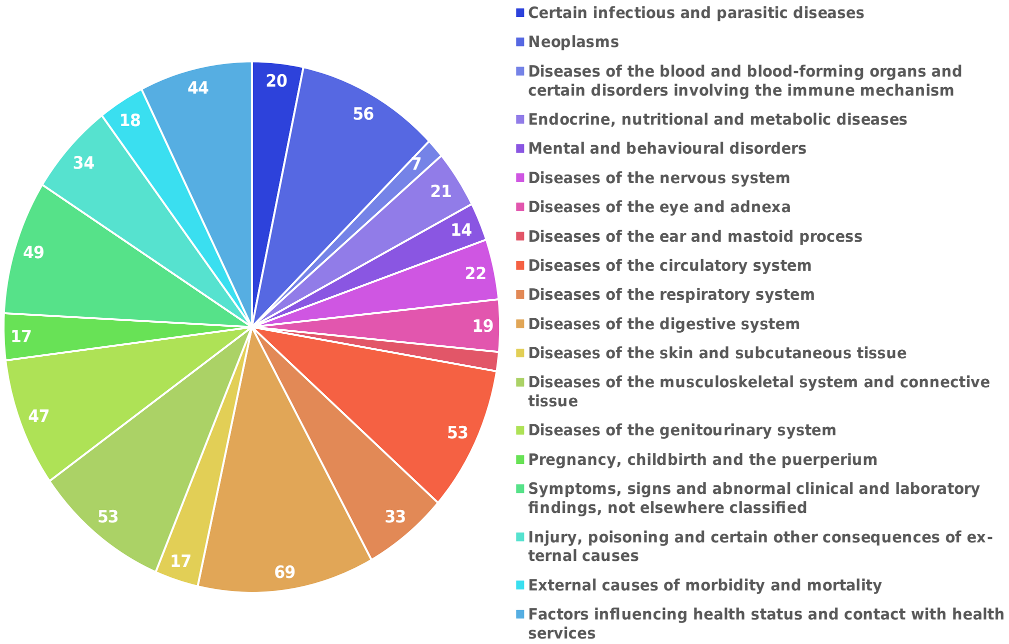


Figure 21. Trait groups and their distribution in terms of number of phenotypes in each group. Each color represents a group of traits whose total number of phenotypes per group can be seen in a pie chart.

2.2 Development of the XWAS strategy and loci identification

In order to identify loci on the X chromosome associated with complex diseases, we designed an XWAS strategy. This included a tailored data quality control for the X chromosome (as described in Methods section 3.3), followed by phasing, imputation, and association procedures outlined in Methods section 3.4. The techniques employed for these stages varied based on whether the analysis was conducted on a dbGaP cohort, primarily using GUIDANCE[154], which encompassed all steps in a single run (methods section 3.4.1), or the UK Biobank, where phasing, imputation, and association were executed as independent steps (methods section 3.4.2). (Figure 22)

With the intention to be able to study potential sex differences, the XWAS strategy was executed separately for males and females, treating them as independent cohorts, and obtaining summary statistics for each phenotype analyzed for each sex. Subsequently, to combine and synthesize the results from different cohorts for the same phenotype (in dbGaP) and merge the results from each sex, we conducted a meta-analysis of the summary statistics allowing us to increase the statistical power (methods section 3.5).

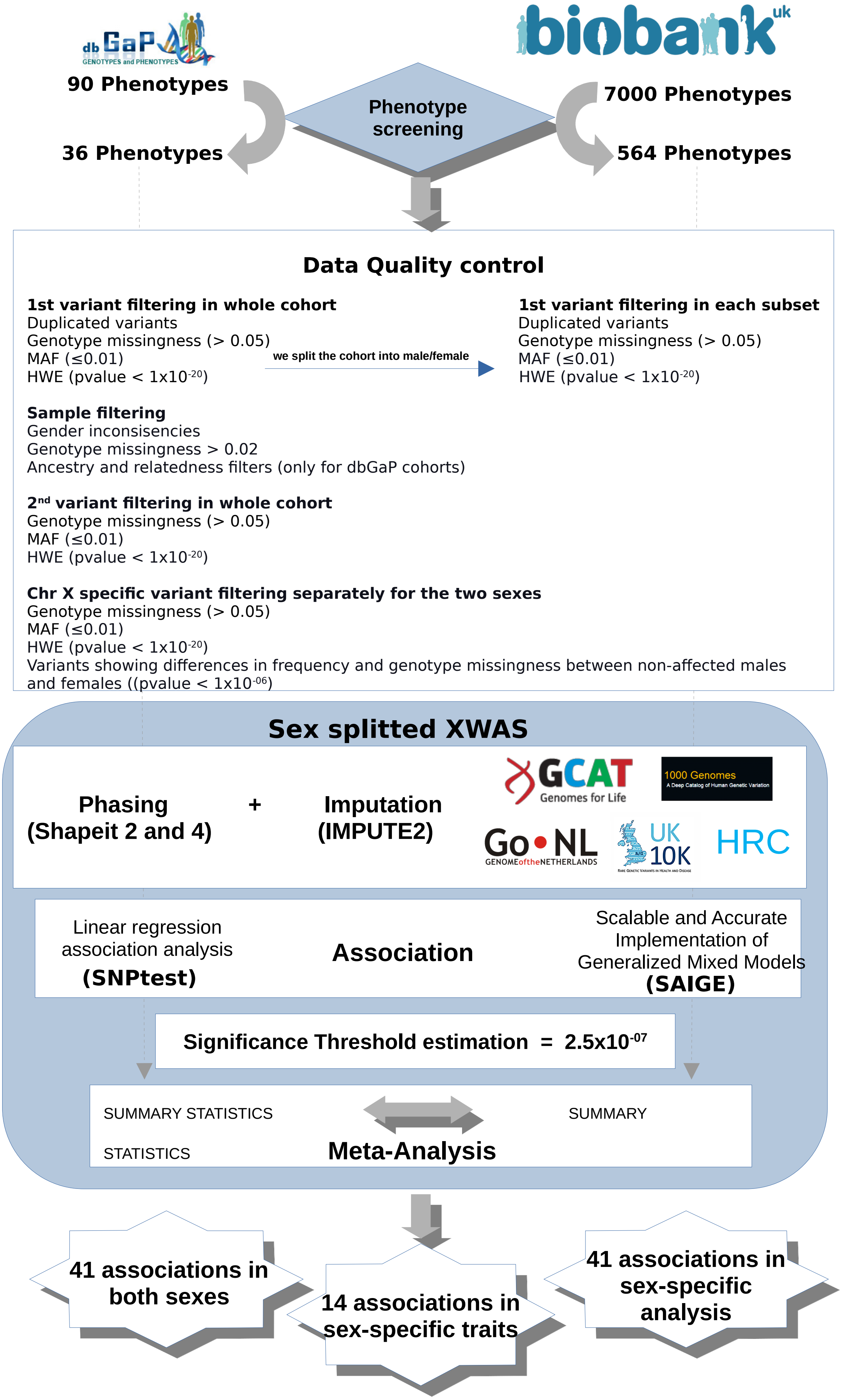

dbGaP GENOTYPES and PHENOTYPES
biobank uk
90 Phenotypes
36 Phenotypes
Phenotype screening
7000 Phenotypes
564 Phenotypes
Data Quality control
1st variant filtering in whole cohort
Duplicated variants
Genotype missingness (> 0.05)
MAF (≤0.01)
HWE (pvalue < 1x10^{-20})
we split the cohort into male/female
1st variant filtering in each subset
Duplicated variants
Genotype missingness (> 0.05)
MAF (≤0.01)
HWE (pvalue < 1x10^{-20})
Sample filtering
Gender inconsisencies
Genotype missingness > 0.02
Ancestry and relatedness filters (only for dbGaP cohorts)
2nd variant filtering in whole cohort
Genotype missingness (> 0.05)
MAF (≤0.01)
HWE (pvalue < 1x10^{-20})
Chr X specific variant filtering separately for the two sexes
Genotype missingness (> 0.05)
MAF (≤0.01)
HWE (pvalue < 1x10^{-20})
Variants showing differences in frequency and genotype missingness between non-affected males and females ((pvalue < 1x10^{-06})
Sex splitted XWAS
Phasing (Shapeit 2 and 4)
+
Imputation (IMPUTE2)
GCAT Genomes for Life
1000 Genomes A Deep Catalog of Human Genetic Variation
Go•NL GENOMEoftheNETHERLANDS
UK 10K
HRC
Linear regression association analysis (SNPtest)
Association
Scalable and Accurate Implementation of Generalized Mixed Models (SAIGE)
Significance Threshold estimation = 2.5x10^{-07}
SUMMARY STATISTICS
SUMMARY
STATISTICS
Meta-Analysis
41 associations in both sexes
14 associations in sex-specific traits
41 associations in sex-specific analysis

Figure 22. XWAS strategy. Utilizing this approach with both dbGaP cohorts and the UK Biobank, we successfully identified 96 significant associations. Among these, 41 were significant in both sexes after conducting a meta-analysis of the summary statistics; 41 were significant exclusively in one of the two sexes; and 14 were significant in sex-specific traits.

With the aim to ensure the reliability of the associations obtained from our XWAS strategy, we first established a threshold for correcting for multiple testing. While the standard p-value used in association studies is $5x10^{-8}$, this threshold is typically applied in genome-wide association studies where the number of tests is much higher. However, since our XWAS analysis was limited to the X chromosome only, we needed to estimate a specific threshold that would be appropriate for this analysis. As explained in methods section 3.6, we estimated the p-value by correcting for the number of independent tests performed using the Bonferroni method to account for multiple testing.

Given that our data came from different sources and dbGaP included a large number of cohorts, to ensure a representative sample of cohorts, we randomly selected three cohorts from dbGaP and the UK Biobank cohort. The number of LD blocks and the significance threshold to be applied in the analysis for each cohort was calculated for each of the four cohorts (Table 7).

Cohort	**Number of variants (chr X)**	**Nº of LD blocks**	**p-value threshold**
dbGaP phs000336	198,740	22,801	$2.19x10^{-6}$
dbGaP phs000346	252,360	35,246	$1.42x10^{-6}$
dbGaP GERA	328,346	64,878	$7.71x10^{-7}$
UK Biobank	445,498	169,370	**$2.5x10^{-7}$**

Table 7. P-value threshold estimation values. UK Biobank cohort showed the highest number of variants imputed in the X chromosome, leading in a higher number of LD blocks (association tests) and, therefore a more stringent p-value threshold.

In order to be conservative and determine a single p-value threshold for the entire project, we chose the most restrictive threshold, which was obtained from the number of recombination blocks estimated in the UK Biobank cohort. Based on this information, we determined that a p-value threshold of $2.5x10^{-7}$ was necessary to identify statistically significant findings.

Finally, to ensure that our strategy was correctly applied, we replicated the process at the whole genome level. We observed that the p-value obtained for a genome-wide association study was approximately $5x10^{-8}$ when only common variants were included and $3x10^{-9}$ when lower frequency variants were also taken into account, obtaining the orders of magnitude expected for this type of analysis[209].

In order to determine significant loci, we applied the significance threshold calculated in the meta-analyzed summary statistics generated for each of the phenotypes studied. After applying the corrected p-value threshold that we had established, we identified a total of 96 associations across 77 complex traits. These associations were further categorized into three groups: 41 associations were detected in both sexes, 14 were detected in sex-specific traits, and 41 were detected through sex-specific analysis (Figure 23).

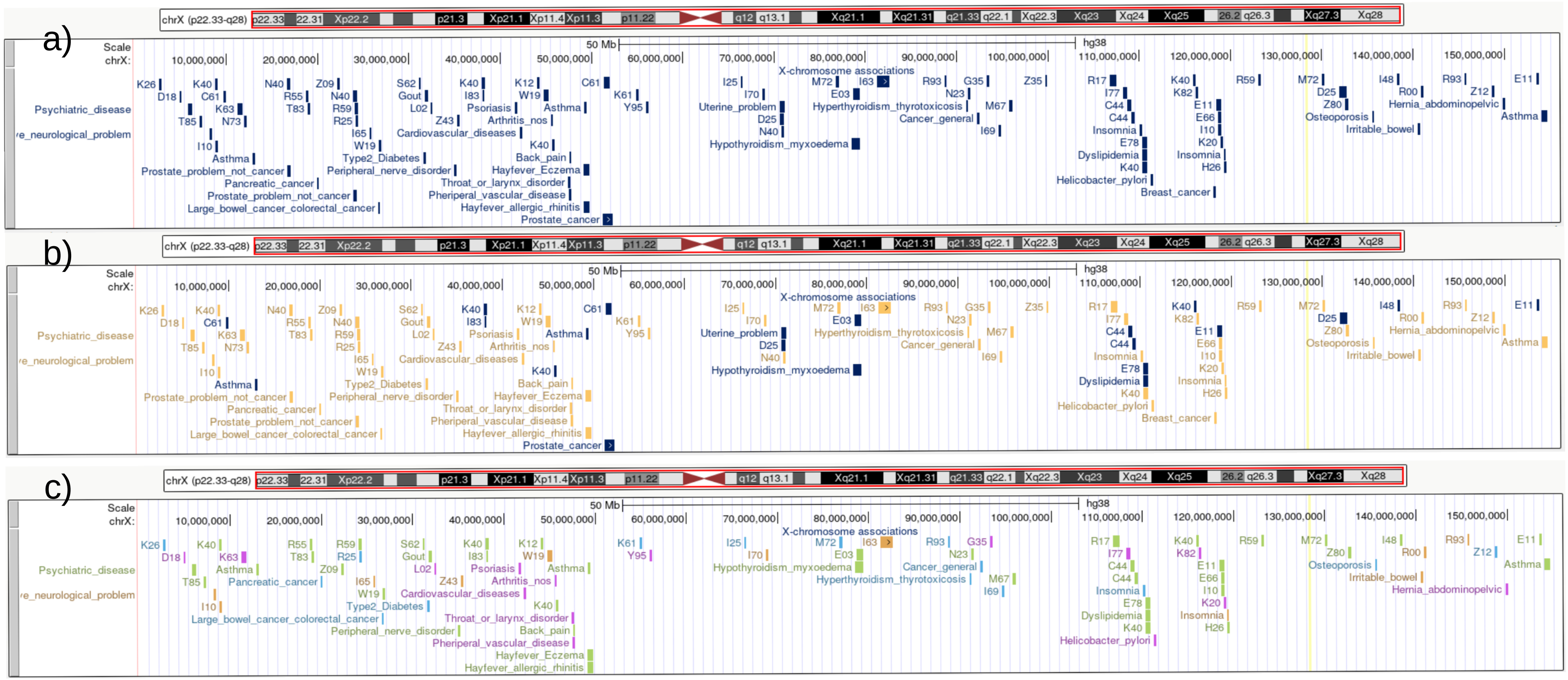


Figure 23. Genome Browser Loci visualization. a) Total of associations (n=96) detected in the project and the corresponding phenotype. **b)** Total of associations (n=96) detected in the project and the corresponding phenotype. Dark blue denotes associations previously reported in GWAS Catalog (version 1, 2019). Yellow represents associations found for the first time in the chromosome X project. **c)** Total of associations (n=82) obtained from traits analyzed in both sexes and the corresponding phenotype. Green denotes associations significant in both sexes (n=41); blue denotes associations significant only in males (n=15); magenta denotes associations significant only in females (n=15); yellow denotes associations that showed significant differences between sexes (n=11).

2.3 Validation of the loci obtained and comparison against public data

2.3.1 Comparison of our findings with the GWAS catalog

Upon identifying 96 associations, our first step was to determine which of these had been previously reported. To achieve this, we compared our results with the associations documented in the GWAS catalog[205]. Through this analysis, we found that 22 out of the 96 associations (23%) had been previously reported (Table 8, Figure 23b).

Phenotype	Chromosome	Position	rsID	Ref allele	Alt allele	p-value
C61	X	9818715	rs5933768	A	G	2,31E-08
Asthma	X	13023741	rs850637	A	G	1,97E-08
I83	X	38009121	rs35318931	A	G	1,67E-07
K40	X	38009121	rs35318931	A	G	4,76E-15
K40	X	45634577	rs56976399	A	C	6,28E-11
Asthma	X	49129023	rs4824747	T	G	5,21E-08
Prostate cancer	X	51202466	rs5945609	A	C	2,20E-13
C61	X	51277134	rs6614433	T	C	1,84E-12
D25	X	70117012	rs771835697	ATT	A	9,33E-09
Uterine problem	X	70148590	rs5936604	C	T	1,31E-10
E03	X	78212886	rs5912197	C	G	8,84E-11
Hypothyroidism myxoedema	X	78441167	rs2152772	C	A	1,34E-14
C44	X	108194346	rs1531061	A	C	2,26E-08
C44	X	108480095	rs79370791	T	C	3,65E-09
Dyslipidemia	X	109693274	rs67648651	C	T	2,47E-10
E78	X	109765216	rs5942642	T	C	9,99E-19
K40	X	115180215	rs140303061	T	A	5,01E-11
E66	X	117929190	rs10126587	G	A	1,08E-08
E11	X	117934682	rs5910417	T	C	3,62E-09
D25	X	131251326	rs5933079	T	C	4,85E-15
I48	X	137418967	rs1891095	A	C	2,55E-08
E11	X	152902768	rs4898432	T	G	3,12E-17

Table 8. Associations from the project already described in GWAS catalog. For each loci the table comprises the phenotype, genomic coordinates, rID, alleles, and the lowest p-value obtained following the meta-analysis of the association process.

In contrast, 74 associations (77%) were identified for the first time in our analysis. Among these, we noticed that some of them overlapped with known findings for traits related or similar to those in which we found the association. Notable examples include the rs5933688 variant, which was associated with Primary Essential Hypertension in our study and has been predicted to be linked to increased creatinine levels, suggesting kidney damage as a consequence of high blood pressure[210]. Another instance is the variants rs79869612 and rs7065158, both associated with Hyperplasia of Prostate and recognized for their relation to prostate-specific antigen levels[211]. Finally, the variant rs181497961, associated with Jaundice, has been connected to Alcoholic Chronic Pancreatitis[212], a condition where jaundice is a typical manifestation.

2.3.2 Analysis of our study’s nobel contribution in UK Biobank

As one way of validating our findings, we compared the results from our UK Biobank analysis to those obtained in the PanUKBB project. After curating the phenotypes present in the PanUKBB project's summary statistics and matching them with those defined in our analysis, we found that 530 phenotypes were common to both studies.

We applied our strategy in order to identify loci on the PanUKBB project's summary statistics. As this project was conducted at the whole-genome level, we used a threshold of $5x10^{-8}$ to consider an association significant following standard procedure. Using this threshold, we identified 37 significant loci, of which 30 were present in phenotypes analyzed in both sexes and seven were associated with sex-specific phenotypes. In contrast, for those 530 phenotypes in our UK Biobank analysis, we found 84 loci: 35 in traits common to both sexes, 13 in sex-specific phenotypes, and 36 sex-specific loci, which were only significant in one of the sexes.

Of the 35 loci detected in phenotypes common to both sexes, 20 (58%) were also found by the PanUKBB project. One significant difference lies in the sex-specific loci identified in our project; the PanUKBB project did not perform an analysis splitting sexes, meaning that the potential benefits in the results that this type of analysis can generate are not present in the PanUKBB project. Consequently, of the 36 sex-specific loci identified in project X, only three of them matched with loci found in the PanUKBB summary statistics, which in this project were shown to be significant in both sexes. Of the 37 loci identified in the PanUKBB project, 29 were also discovered in our study (Table 9). The remaining eight were associated with traits with different sample sizes for the same phenotype between projects, highlighting the impact of the sample size on the association process.

	Total Loci (n)	Loci detected in both sexes (n)	Single Sex Loci (n)	Sex Specific Loci (n)
Chr X project	84	35	13	36
PanUKBB project	37	30	7	Not analyzed
Chr X project loci found in PanUKBB	29	20	6	3
Chr X project loci not found in PanUKBB	55	15	7	33

Table 9. Comparison of loci detected in UK Biobank using 530 binary traits on the X Chromosome between the X Chromosome Project and the PanUKBB Project.

The results show that our approach to analyzing the X chromosome yields higher loci count than that obtained with the PanUKBB data. In our analysis, we placed emphasis on cleaning the controls to ensure the least possible contamination of risk alleles within controls, applying a specific QC tailored for the X chromosome, and conducting a stratified analysis by analyzing males and females separately, processes not carried out in the UKBB analysis by PanUKBB. The results suggest that the application of this approach produced a loci count more consistent with expectations, similar to observations for chromosome 7, which has a size comparable to the X chromosome. Conversely, the results of the PanUKBB analysis, which were obtained without splitting the sexes and analyzing the entire cohort, display a loci count below expectations.

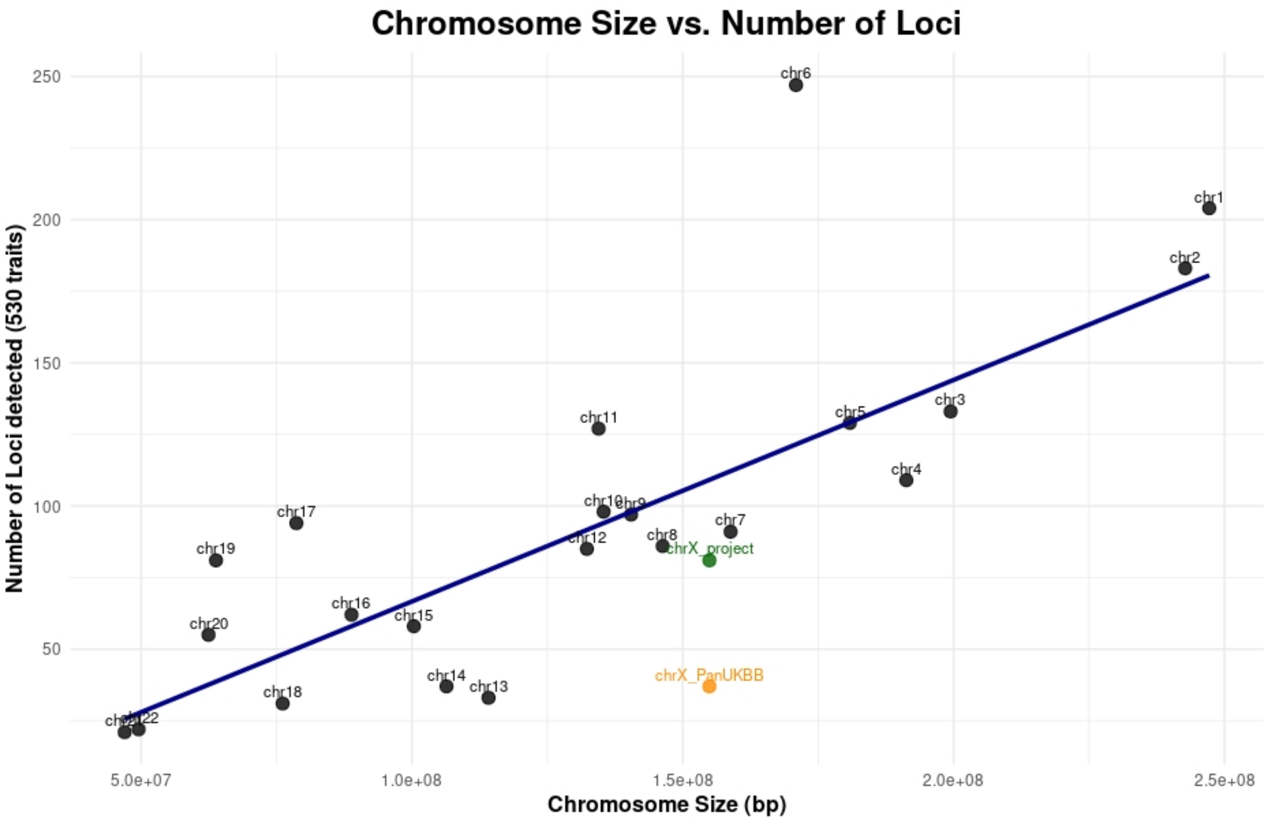


Figure 24. Trait groups and their distribution in terms of number of phenotypes in each group. Scatter-plot showing the relationship between chromosome size (bp) and the number of loci detected (in 530 traits) for each chromosome. The points are labeled with their respective chromosome names. Points representing "chrX_project" and "chrX_PanUKBB" are highlighted in dark green and dark orange and represent the number of loci found in each project respectively. A linear regression line (dark blue) is added to the plot to visualize the overall trend in the data.

2.3.3 Replication between dbGaP and UK Biobank

Finally, we assessed the replication between the results obtained for the same or a related phenotype present in the two sources examined in this project, UK Biobank and dbGaP. We found that 24 binary traits, encompassing 22 significant loci from the UK Biobank, corresponded with 14 dbGaP phenotypes, which included nine significant loci. The correspondence involved either the direct comparison of the same phenotype (e.g., Asthma, referred to as such in both sources), or the examination of directly related phenotypes (e.g., Abdominopelvic Hernia in dbGaP, equivalent to ICD10 codes K40, K42, K43, and K44 in UK Biobank, representing various types of hernias in the abdominopelvic region).

Out of the 31 loci present in traits that matched in both sources, 4 of them demonstrated replication between the two analyses. Firstly, rs5945609 and rs6614433 were significant for Prostate Cancer in dbGaP and Malignant neoplasm of prostate (C61) in the UK Biobank, respectively. Secondly, rs5942642 and rs67648651 were significant for Dyslipidemia in dbGaP and Disorders of lipoprotein metabolism and other lipidaemias (E78) in the UK Biobank.

2.4 Evaluation of sex-specificity thorough allelic selection

One of the most striking findings in our exploration of the X chromosome was the discovery of significant associations in only one sex but not the other, which could be identified as sex differences. As seen in the comparison with data from the PanUKBB project (Section 2.3.2), that project did not conduct separate analyses for each sex, leading to a much smaller number of significant associations compared to our analysis of the UK Biobank. In our case, adding this sex-specific analysis allowed us to find a number of associations much closer to the expected number for a chromosome the size of the X chromosome (Figure 24).

These results opened up a series of questions about whether these associations were truly sex-specific, or if they could be attributed to a lack of statistical power in one sex or to how the methodology and tools used in the association process deal with the biological differences that exist between men and women in the X chromosome.

The primary distinction between males and females regarding the X chromosome, which is absent in autosomes, is the existence of two copies in females and a single copy in males. This results in the presence of heterozygous genotypes (0/1) in females, which could potentially impact the outcomes of association studies. In fact, as explained earlier, although one of the two copies in females is silenced, there are genes that escape inactivation, which could also have consequences in the differences observed between both sexes.

To explore the influence of female heterozygous genotypes (0/1) on association analyses, we replaced them with missing calls (./.) in the UK Biobank imputed data. In this way we were mimicking the male-specific situation and conducting the association test again. Subsequently, we identified loci using the summary statistics produced after performing the association analysis with female heterozygotes removed.

Eliminating heterozygotes restricted the number of variants available for analysis due to the new frequency distribution resulting from their removal. Consequently, a significant number of variants failed to surpass the MAF threshold necessary for their inclusion in the association test (MAF > 0.0001). In the standard analysis, approximately 450,000 variants were evaluated in females for each phenotype; however, this value decreased to an average of 245,000 variants in the analysis without heterozygotes. As a result, the findings were affected not only by a reduced sample size but also by a decreased number of variants.

In the standard analysis, 70 significant loci were identified in the analyzed phenotypes for both genders. However, due to the reduced number of variants examined in the analysis without heterozygous genotypes, only 52 of these loci could be evaluated. After excluding female heterozygotes and conducting the analysis again, 37 loci were re-identified from the initial 52. Notably, 11 of these 37 loci exhibited differences in the analyses that deemed them significant; for example, rs5912197 was significant in both, the independent female analysis and the male-female meta-analysis in the standard evaluation, but only after the meta-analysis in the updated analysis. In contrast, rs147238402, initially significant solely in males, achieved significance after a meta-analysis with females without heterozygotes, suggesting that the presence of these genotypes added noise to the analysis, hindering the detection of this association in females as well.

The most noteworthy result of this analysis lies in the identification of four novel loci that were not deemed significant in the standard analysis. This discovery holds considerable importance as three of these loci were determined to be significant in the meta-analysis of males and females, while they were only suggestively significant (p-value < $1x10^{-5}$) in the standard analysis. The fourth detected locus (rs945030728) was entirely new, displaying a p-value of 6.91 $x10^{-08}$ in the female analysis, in contrast to its original p-value of 0.1 (Table 10).

		Analysis without heterozygous females					Standard analysis				
phenotype	**variant**	**All p-value**	**Male p-value**	**Female p-value**	**Genderdif p-value**	**Genderhet p-value**	**All p-value**	**Male p-value**	**Female p-value**	**Genderdif p-value**	**Genderhet p-value**
R07	X:111628623_AT_A	$1.76x10^{-7}$	$1.00x10^{-6}$	0.065	$1.15x10^{-6}$	0.872	$4.28x10^{-5}$	$1.00x10^{-6}$	0.823	$6.13x10^{-6}$	0.007
E83	X:22600410_G_GGTGT	0.154	0.8131	$6.91x10^{-8}$	$4.59x10^{-7}$	$1.91x10^{-7}$	0.278	0.813	0.102	0.26	0.214
I83	X:43822590_T_C	$9.21x10^{-8}$	$6.3x10^{-5}$	0.0004	$5.72x10^{-7}$	0.689	$2.67x10^{-7}$	$6.03x10^{-5}$	0.001	$1.63x10^{-6}$	0.700
N39	X:69951298_G_A	$1.84x10^{-7}$	0.001	$9.14x10^{-7}$	$9.14x10^{-7}$	0.446	$4.44x10^{-7}$	0.0013	$7.80x10^{-5}$	$2.33x10^{-6}$	0.524

Table 10. New loci discovered in the analysis without heterozygous females compared to the standard analysis. All p-value refers to the value obtained after meta-analyzing males and females; male and female p-value is the value obtained in the analysis of males and females separately, respectively; genderdiff p-value refers to significant signals considering different effects between sexes; genderhet p-value refers to significant differences between the analysis of both sexes. Four loci were found as new in the analysis without heterozygotes, of which three were found to be suggestively significant (p-value < $1x10^{-5}$) in the standard analysis, while one of them is completely new and far from being significant in the standard analysis.

To further investigate these critical findings, we employed positional mapping to identify genes overlapping with these four loci, obtaining a relation of one gene per locus. With the aim of determining if any of the identified genes were known as genes that escape the process of X chromosome inactivation, we compared them with previous literature that has studied which genes escape this inactivation process [162,166]. Through this comparison, we determined that none of the four genes showed an overlap with escape genes, indicating that these associations discovered by removing heterozygous genotypes in females do not have a direct relationship with the phenomenon of escape from inactivation. In fact, this analysis and these results could even also point to other still unknown genes that might escape inactivation.

2.5 Functional interpretation

Finally, to delve deeper into potential functional sex differences, we decided to conduct a functional interpretation of the associations identified in the project. This functional interpretation involved an initial fine-mapping of the detected associations, gene-mapping of the obtained credible sets to establish a relationship of one gene per detected loci in the project, and ultimately a pathway enrichment analysis. In this way, our goal was to determine if there were metabolic pathways enriched in one sex but not in the other, which could be interpreted as sex differences in the X chromosome.

2.5.1 Fine mapping and gene mapping of the identified loci

We performed fine mapping analysis to obtain a credible set for each of the loci identified in the project. The credible set consisted of a set comprising between one and five variants with a high statistical likelihood of including the causal variant (likelihood threshold = 0.95).

Using the results of the fine mapping, we performed gene mapping for each credible set found. This analysis enabled us to identify potential candidate genes associated with the binary traits being investigated. We employed three different strategies (see Methods section 4.2) to map the credible sets. If a credible set mapped to multiple genes, we chose the gene with the most variants mapped to it. If the number of variants was equal, we selected the gene derived from the variant with the highest likelihood of being causal. This was done to obtain a single gene for each credible set and, thus, each locus identified in our study.

The genes obtained were categorized based on their origin: whether they came from significant loci in both sexes, significant loci in only one sex (known as sex-specific loci), or loci obtained from phenotypes studied exclusively in one sex (known as single-sex traits). (Table 11).

Phenotype	Locus origin	Credible set	Consensus gene
K26	Only male	X:3151814_T_C | X:3090684_T_C | X:2700608_C_T	ENSG00000124343
D18	Only female	X:4004245_C_T | X:5534797_A_T |X:5912489_T_C X:4737290_T_G | X:4587308_G_A	ENSG00000183943
Psychiatric disease	Both sexes	X:5989782_A_G	ENSG00000146938
T85	Both sexes	X:7184763_A_G | X:7025770_G_A | X:6772737_GA_G X:7559466_TA_T | X:8438454_A_G	ENSG00000101846
Chronic degenerative neurological problem	Only male	X:8243524_G_A | X:7373907_G_C | X:8445173_A_T X:6864400_C_G | X:8865654_G_T	ENSG00000205642
I10	Only male	X:8880680_A_G	ENSG00000285896
K40	Both sexes	X:8888591_T_C	ENSG00000285896
C61	Single sex trait	X:9818715_G_A	ENSG00000146950
K63	Only female	X:11538045_G_A	ENSG00000047648
N73	Single sex trait	X:12073195_A_G	ENSG00000169933
Asthma	Both sexes	X:13023741_G_A	ENSG00000205542
Prostate problem no cancer	Single sex trait	X:16792837_G_GAAAG	ENSG00000169895
N40	Single sex trait	X:16874665_A_AAAC	ENSG00000102054
R55	Both sexes	X:18543509_G_A | X:18869524_C_T | X:18399062_T_C	ENSG00000008086
T83	Both sexes	1 X:20459167_T_A | X:18141412_A_G | X:20505404_A_G X:18762364_C_T | X:17633301_AT_A	ENSG00000086717
Pancreatic cancer	Only male	X:21229743_G_A | X:21255192_AG_A | X:21162050_T_G X:21146076_T_C | X:19958394_AT_A	ENSG00000173681
Z09	Both sexes	X:21472903_C_T | X:21153502_G_T | X:21744914_A_T X:21240504_C_G | X:21407839_A_G	ENSG00000149970
Prostate problem no cancer	Single sex trait	X:24098069_A_G	ENSG00000130741
N40	Single sex trait	X:24109619_A_G	ENSG00000130741
R59	Both sexes	X:23840016_A_C | X:23382189_G_T | X:23415657_G_A X:23444605_T_C | X:24041906_T_C	ENSG00000174010
R25	Only male	X:24197650_G_A | X:24495753_C_A | X:25118033_C_A X:23016084_A_C | X:23106033_G_A	ENSG00000005889
I65	Only female	X:25715948_C_T | X:25433783_T_C | X:26322537_C_T X:24445073_C_G | X:24321825_G_T	ENSG00000223611
Large bowel cancer and colorectal cancer	Only male	X:25953128_C_A | X:26990881_T_C X:27738438_CAAAGTT_C | X:27227481_C_T X:28070796_C_T	ENSG00000176774
W19	Both sexes	X:27095594_A_C | X:27480000_G_A | X:26832792_G_C X:28264085_G_A	ENSG00000224960
S62	Both sexes	X:29950884_CTCT_C | X:30040824_TTAC_T X:30365927_C_G | X:30606616_C_T | X:32462963_A_AT	ENSG00000169297
Type2Diabetes	Only male	X:32766396_G_A | X:33092319_G_A | X:32653928_T_C X:31723735_T_C | X:32059752_C_A	ENSG00000198947
	Both sexes	X:31888516_G_T | X:32461239_T_C	ENSG00000198947

Gout			
L02	Only female	X:31278576_G_A | X:31079346_C_A | X:32807911_T_G X:31304172_G_A | X:32797117_C_A	ENSG00000198947
Peripheral nerve disorder	Both sexes	X:34846466_A_T | X:34076543_T_C | X:34819477_G_A X:34424882_TTTA_T | X:35065534_T_C	ENSG00000189132
Z43	Gender differentiated	X:34439541_G_A | X:34711995_C_T | X:34285504_T_C X:35673003_C_T | X:35500698_G_C	ENSG00000147027
K40	Both sexes	X:37405039_T_C | X:37252435_C_G | X:39242085_T_C X:39229907_A_G | X:39358150_T_A	ENSG00000236747
I83	Both sexes	X:38009121_G_A	ENSG00000101955
Psoriasis	Only female	X:37757650_C_G | X:37508035_A_G | X:39026070_G_A X:38554463_C_CAT | X:39030419_T_C	ENSG00000250349
Cardiovascular	Only female	X:42111359_G_T | X:42016573_T_C | X:41973293_T_C	ENSG00000147044
K12	Both sexes	X:44003701_T_C | X:44175541_C_G | X:44565588_C_T X:43872746_C_G | X:42907496_G_T	ENSG00000183690
W19	Only male	X:44970788_G_A | X:44785770_T_C | X:45556805_A_G X:44892922_C_T	ENSG00000147050
Arthritis	Only female	X:45522820_A_G	Unmapped
K40	Both sexes	X:47108144_T_G | X:47106374_C_T | X:46945765_T_A X:44912294_G_T | X:44846113_T_C	ENSG00000147050
Throat or larynx disorder	Only female	X:47350035_T_G | X:45933214_T_A | X:46839591_G_A X:47265649_G_A | X:48743290_GC_G	ENSG00000102221
Peripheral vascular disease	Only female	X:47441089_G_A | X:47158487_C_T | X:46557595_G_A | X:47430688_C_T | X:48732666_A_C	ENSG00000008056
Back pain	Both sexes	X:47486217_C_T | X:48754724_G_A | X:47333344_T_C X:48624003_A_G | X:47584496_TC_T	ENSG00000126759
Asthma	Both sexes	X:49129023_G_T | X:48595301_C_A	ENSG00000049769
Hay-fever	Both sexes	X:49129023_G_T | X:49366437_C_T	ENSG00000049769
Hay-fever and allergic rhinitis	Both sexes	X:49200938_G_A	ENSG00000275113
Prostate cancer	Single sex trait	X:49753723_G_A | X:49944745_A_T | X:49726526_G_A X:50343251_T_G | X:49899450_C_G	ENSG00000268668
C61	Single sex trait	X:51277134_C_T	ENSG00000196368
K61	Only male	X:55023615_T_C | X:54805480_A_C | X:53774406_C_T X:55593706_CA_C | X:54000969_G_A	ENSG00000158571
Y95	Only female	X:55120069_T_C | X:56511875_G_T | X:56964420_C_T X:55897218_C_T | X:55827184_T_C	ENSG00000204271
I25	Only male	X:65673693_G_A	ENSG00000235892
I70	Gender differentiated	X:69389240_C_G | X:69318842_G_T | X:69454709_C_T X:67644715_A_G | X:67453430_T_G	ENSG00000079482
D25	Single sex trait	X:70117012_A_ATT	ENSG00000120498
N40	Single sex trait	X:70129087_G_A	ENSG00000120498
Uterine problem	Single sex trait	X:70148590_T_C	ENSG00000165349
M72	Only male	X:76156017_G_A	ENSG00000280870
E03	Both sexes	X:78212886_G_C | X:78112989_C_A	ENSG00000078589
Hypothyroidism myxedema	Both sexes	X:78479522_A_G | X:78234986_T_G	ENSG00000147138
I63	Gender differentiated	X:80218437_G_A | X:80955776_G_C | X:80683069_T_C X:80561709_A_G | X:80177203_G_A	ENSG00000131171
R93	Only male	X:87737812_G_A | X:88018635_CAA_C | X:87069659_T_G X:87794187_G_T | X:88855130_C_A	ENSG00000147183
Hyperthyroidism thyrotoxicosis	Only male	X:90389321_G_C | X:91757984_A_G | X:89356042_G_A X:90649144_T_C | X:91636075_A_G	ENSG00000102290
N23	Both sexes	X:90667995_C_T | X:90695584_A_G | X:90883302_G_A X:90669708_T_C | X:89911752_C_T	ENSG00000174740
Cancer general	Only male	X:91582544_G_A | X:92890167_G_C | X:90762608_CAAAGACT_C	ENSG00000102290
G35	Only female	X:92082667_C_T | X:92060444_G_A | X:93639734_C_T X:92371828_T_C | X:92703048_A_T	Unmapped
I69	Only male	X:93947495_T_C | X:94006137_T_C | X:94499609_A_G X:93994502_G_A | X:94714030_A_C	Unmapped
M67	Both sexes	X:93865619_G_A | X:93809585_G_A | X:94140670_G_T X:94474126_T_G | X:93902156_C_T	Unmapped
Z35	Single sex trait	X:99062092_G_A	Unmapped
R17	Both sexes	X:106782380_C_A | X:106929203_T_C | X:106465807_C_T X:106441705_C_T | X:106493019_T_C	ENSG00000198088
I77	Only female	X:106549312_C_T | X:108304371_C_T | X:109013924_G_T X:109234075_T_C | X:106701873_A_T	ENSG00000157600
C44	Both sexes	X:107959392_A_G | X:106769495_G_A | X:108480095_C_T	ENSG00000147234
C44	Both sexes	X:108480095_C_T | X:108736347_T_A	Unmapped
Insomnia	Only male	X:108051036_A_G | X:108533517_G_A | X:109447142_C_T X:108402912_C_T | X:108302592_G_A	ENSG00000101935
Dyslipidemia	Both sexes	X:108348068_C_T | X:108381092_A_C | X:109693274_T_C	ENSG00000243978
E78	Both sexes	X:109765216_C_T | X:109826248_C_T | X:109823107_A_C	ENSG00000225366
K40	Both sexes	X:111051010_C_G | X:111251842_A_T | X:109867437_A_T X:110606543_C_T | X:110871440_G_A	ENSG00000077279
Helicobacter pylori	Only female	X:110354055_C_T | X:110328485_G_A | X:110117185_C_T X:111381402_A_G | X:110538777_C_T	ENSG00000077279
K40	Both sexes	X:115180215_A_T	ENSG00000231371
K82	Only female	X:115457230_C_T | X:114638688_T_G | X:115569541_T_C X:115460721_G_T | X:116225768_C_T	ENSG00000268104
Breast cancer	Single sex trait	X:117331280_A_G	ENSG00000003096
I10	Both sexes	X:117837501_CT_C | X:118091385_G_A	ENSG00000175556
E66	Both sexes	X:117097235_T_C | X:117162121_T_G | X:118680017_T_C	ENSG00000003096

E66		X:117929190_A_G	
E11	Both sexes	X:116817434_G_A | X:116484814_C_A | X:116637036_A_G X:116598031_T_C | X:117934682_C_T	ENSG00000174460
K20	Only female	X:118169700_C_T | X:118929766_C_T | X:119069900_G_T X:117428599_G_A | X:118626504_C_T	ENSG00000101882
Insomnia	Only female	X:119384771_G_C | X:119042496_T_G | X:117022912_G_A X:119402013_G_T | X:117850088_G_C	ENSG00000177485
H26	Both sexes	X:117634121_C_T | X:117551627_C_T | X:118496486_C_T	ENSG00000147251
R59	Both sexes	X:121474259_G_A | X:121383804_CATT_C X:122548724_A_G | X:121927598_T_A | X:122541353_A_C	ENSG00000125675
M72	Both sexes	X:127898758_T_G | X:128031104_T_C | X:130145554_T_G | X:128077995_T_C | X:130629000_G_A	ENSG00000228659
D25	Single sex trait	X:131251326_C_T	ENSG00000165694
Z80	Both sexes	X:131919257_C_T | X:131492415_AC_A	ENSG00000171004
Osteoporosis	Only male	X:135938553_C_T | X:135768806_C_T | X:135811787_G_A X:135768800_TAC_T | X:135853380_C_T	ENSG00000129675
I48	Both sexes	X:137418967_C_A	Unmapped
Irritable bowel	Only male	X:138398721_C_T | X:138386026_T_TAC | X:140393565_A_G X:138443366_T_G | X:140560506_A_G	ENSG00000198021
R00	Only male	X:140067636_G_A	ENSG00000227234
R93	Only female	X:144817981_A_G | X:144724916_C_A | X:145333346_G_A X:145305194_G_A | X:146304165_C_T	ENSG00000185985
Z12	Only male	X:147861213_G_A | X:146463678_G_A	ENSG00000155966
Hernia abdominopelvic	Only female	X:148975998_A_G | X:147629091_A_G | X:149125169_C_T X:147855237_A_G | X:149115193_A_G	ENSG00000155966
E11	Both sexes	X:152902768_G_T | X:152891709_T_C | X:153402410_A_AC X:153561609_G_A	ENSG00000130829
Asthma	Both sexes	X:153744507_C_G	ENSG00000071889

Table 11. Fine-mapping and gene-mapping results for each locus. The first column lists the name or ICD10 code of the phenotype where the locus has been identified. The second column indicates the origin of the locus, including both sexes (significant in both sexes), only females or only males, analyzed in both sexes but significant only in females or males, respectively; and single sex trait, phenotypes only present in one of the sexes. The third column displays the credible sets for each locus, which contain between one and five variants with their position and alleles. The last column reports the consensus gene identified using the three strategies. If no gene has been mapped according to the criteria explained in the methods section 4.2, the label "Unmapped" is assigned.

2.5.2 Pathway enrichment analysis

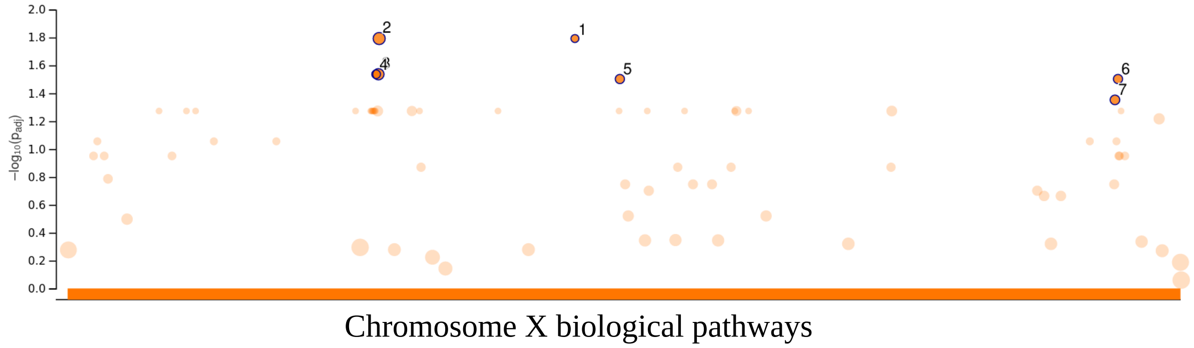
Once we got a list of genes from our association in the X chromosome, we compared the genes obtained from the gene mapping step specific to males and females with the aim to identify possible significant differences in the biological pathways that those genes are affecting. Our objective was to determine if certain pathways were enriched by genes in one sex but not the other, resulting in possible functional differences between the sexes.

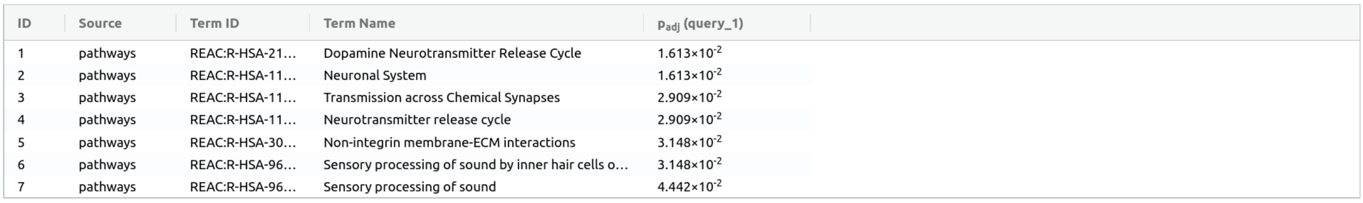
The gene mapping process linked each locus to a single gene, resulting in a list of 19 genes coming from a signal only significant in males and 16 genes coming from a signal only significant in females (Table 11). In order to determine whether the observed differences were specific to either males or females, we used the genes obtained from significant associations found in both sexes as a control group.

We performed the enrichment analysis using the g:GOST module of the g:Profiler tool (methods section 5.2), including only the genes present on the X chromosome and the biological pathways in the g:Profiler database. To ensure the reliability of our results, we corrected the enrichment outcomes using the

Benjamini-Hochberg FDR correction method with a significance threshold value of 0.05.

Our analysis revealed that seven pathways were specifically enriched with female-mapped genes, while no pathways were enriched with male-mapped genes. Furthermore, none of the seven female-enriched pathways were detected as pathways enriched using the control group of genes (Figure 25), which means that the enrichment of these pathways is exclusive to females, thereby indicating potential gender differences.





ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	pathways	REAC:R-HSA-21...	Dopamine Neurotransmitter Release Cycle	1.613×10^{-2}
2	pathways	REAC:R-HSA-11...	Neuronal System	1.613×10^{-2}
3	pathways	REAC:R-HSA-11...	Transmission across Chemical Synapses	2.909×10^{-2}
4	pathways	REAC:R-HSA-11...	Neurotransmitter release cycle	2.909×10^{-2}
5	pathways	REAC:R-HSA-30...	Non-integrin membrane-ECM interactions	3.148×10^{-2}
6	pathways	REAC:R-HSA-96...	Sensory processing of sound by inner hair cells o...	3.148×10^{-2}
7	pathways	REAC:R-HSA-96...	Sensory processing of sound	4.442×10^{-2}

Figure 25. Scatter plot representation of pathway enrichment results using genes from female-specific loci. Each dot on the scatter plot symbolizes a biological pathway influenced by genes located on the X chromosome. The seven highlighted and numbered points depict the pathways that have achieved significant enrichment (using the Benjamini-Hochberg FDR correction with a 0.05 threshold). The accompanying table provides details about each of these seven pathways, including their names and the adjusted p-values obtained.

Some notable female-enriched pathways include "Sensory Processing of Sound by Inner Hair Cells of the Cochlea" and "Sensory Processing Sound Pathways," which have been previously documented as gender-based differences in sound processing[213]. Additionally, other enriched pathways that have been reported to display sex-based differences[214] are the "Neuronal System", "Dopamine Neurotransmitter Release Cycle", "Neurotransmitter Release Cycle", and "Transmission Across Chemical Synapses."

Additionally, we explored the potential relationship between sex-specific significant signals and the phenomenon of X-chromosome escape inactivation in females. To achieve this, we compiled an extensive list of genes (n=337) previously reported to evade X-chromosome inactivation[162,166] and created an "artificial" biological pathway comprising these genes. Our analysis indicated that the pathway formed from this gene list was not recognized as a significantly enriched pathway in either male or female significant signals.

3 Impact of the GCAT reference panel on X-chromosome analysis

In addition to the two projects outlined in this thesis and to conclude the results section, we aimed to study the relationship between the projects developed and explained herein by examining the benefits of incorporating the GCAT haplotype reference panel, created in the first block of the thesis, into the analysis of the X chromosome, corresponding to the second block.

To accomplish this, we conducted an analysis to determine the impact of the GCAT panel on the imputation process and, subsequently, on the number of associations discovered in the X chromosome. This analysis involved identifying the number of imputed variants that would be lost if we did not use the GCAT and relied solely on the other four reference panels (GoNL, 1000 Genomes Phase 3, HRC, and UK10K). These panels were generated with samples of lower coverage than the GCAT, which is an important factor for obtaining high-quality INDELs and SVs in the panel and enabling their imputation.

We randomly selected five cohorts from our study and calculated the total number of imputed variants (info score > 0.7) by combining all reference panels. We then estimated this number again without including the GCAT reference panel.

Our findings revealed that removing the GCAT as a reference panel would lead to an average reduction of 7.1% (SD=0.4) in the number of variants imputed. Regarding the type of variant, SNVs would be least affected, losing an average of 3.8% (SD =0.5) of them. In contrast, we observed a more significant loss for INDELs and SVs, with 28.7% (SD=1.22) and 81% (SD=1.5) of variants lost, respectively. (Table 12)

Cohort	Total variants	SNVs	INDELs	SVs	Total variant lost	SNVs lost	INDELs lost	SVs lost
phs001315	238,362	205,619	32,390	353	17,517 (7.3%)	7,827 (3.8%)	9,403 (29.0%)	287 (81.3%)
phs000788	476,459	409,740	65,963	756	36,042 (7.6%)	14,780 (3.6%)	20,6551 (31.3%)	611 (80.8%)
phs001202	266,581	235,019	13,798	340	19,785 (7.4%)	10,499 (4.5%)	9,003 (28.8%)	283 (83.2%)
phs001039	377,177	324,566	22,178	583	24,486 (6.5%)	9,886 (3.0%)	14,139 (27.2%)	461 (79.1%)
phs001319	395,942	340,656	23,701	591	26,916 (6.8%)	11,082 (3.3%)	15,364 (28.1%)	470 (79.5%)

Table 12. Imputed variant lost removing the GCAT reference panel from the study. The first four columns display the total number of imputed variants using GUIDANCE, which includes five reference panels (GCAT, 1000genomes Phase3, GoNL, UK10K, and HRC). Subsequent columns show the number of variants lost without GCAT and the percentage of the total that represents.

These findings suggest that while the GCAT reference panel does not significantly affect SNVs imputation in comparison to other reference panels, it has a considerable impact on the imputation of INDELs and SVs. Without using the GCAT reference panel, a big proportion of INDELs and the majority of SVs would be lost, highlighting its significance compared to other publicly available panels in terms of including these types of variants in GWAS.

The impact of GCAT in our study is also evident when examining the significant loci obtained. Out of the 96 loci identified, two of them would not have been detected without GCAT. Both loci present a single significant variant, which, in both cases, are INDELs that were imputed only with the GCAT reference panel. These signals are rs766284533, associated with Multiple sclerosis (ICD10 code G35) in women (p-value = $2.2\text{x}10^{-7}$), and rs201807380, associated with Localized enlarged lymph nodes (ICD10 code R59) in both sexes (p-value = $1.8\text{x}10^{-7}$).

Discusion

In recent years, the genetics field has increasingly emphasized the identification and characterization of genetic factors involved in complex diseases, alongside environmental factors. In the effort to study the genetic and molecular basis of these diseases and identify risk predictors for clinical prevention, GWASs are employed as a primary tool, analyzing extensive genetic and phenotypic data from numerous individuals to detect genetic variations associated with specific traits. In this thesis, I present the results of two studies addressing current limitations in GWAS, contributing to the understanding of the genetic basis of complex diseases.

Over the past years, GWASs have led to the identification of thousands of loci linked to a wide range of complex traits. Despite the significant progress, there is still much to uncover regarding the genetic basis of these diseases, leading to a large portion of heritability (variance explained by genetic factors) remaining unexplained. One of the main reasons is that GWAS have historically relied on the use of genotyping array data, which analyze a limited number of variants, mainly SNVs, and with a common frequency in the population. This results in low frequency, rare variants, or larger genetic variants and their possible associations remaining elusive.

Ideally, the logical move to gain discovery power in GWAS would be to evolve from genotyping array technology to whole-genome next-generation sequencing, which would allow for the inclusion of these types of variants in GWAS effectively. However, considering the sample size normally used in GWAS (in the order of thousands or hundreds of thousands of individuals), the economic cost of sequencing and the extensive downstream analysis required is still prohibitive and not realistic. This makes the transition difficult to achieve without sacrificing sample size, and therefore statistical power, at the expense of having a larger number of variants to analyze. In any case, we do expect that the use of whole-genome sequencing for GWAS will become standardized and regularized, enabling the identification of rarer risk variants that are likely to have a greater impact on diseases than common variants.

Hence, in this context, generating haplotype reference maps remains essential, as they enable the expansion and enrichment of discovery possibilities by imputing (predicting) a large number of variants from genotyping data. During the last decade, large projects have emerged with the goal of generating haplotype reference panels from hundreds to thousands of whole-genome sequenced samples representative of one or more populations. As demonstrated in this thesis, creating these panels has facilitated the identification of variations (specifically, SVs, INDELs, and SNVs) associated with diseases that could not have been detected solely from genotyping data. Thus, when WGS is not yet an option, it is crucial to generate comprehensive and precise reference panels for studying complex diseases.

One of the major limitations of existing haplotype reference is the range and types of variants considered, which is directly translated into a limited discovery power. When looking at the associations described to date, for example in catalogs like the GWAS catalog, two observations become evident. First, almost all associations derive from SNVs and INDELs, although it is known that SVs can have a significant impact on the development of complex diseases. The lack of SV-driven associations primarily stems from their exclusion in genotyping arrays and haplotype panels, nearly eliminating the study of their role in complex diseases. Currently, available haplotype reference panels only contain a testimonial amount of structural variants, resulting in these not being imputed and therefore not being analyzed in the association process.

Another significant limitation of GWAS studies corresponds to the systematic exclusion of the X chromosome, leading to a situation where our current knowledge represents only about 10% of what we would expect if this chromosome had been considered. The exclusion of the X chromosome from GWAS is not based on scientific grounds, but rather on practical decisions to simplify the study. While the small number of associations on the Y chromosome is understandable due to its low gene content, the absence of associations on the X chromosome is expected to have a greater impact given its size and genetic content. In fact, this thesis demonstrates that there is still potential for new discoveries within the X chromosome, even when using previously analyzed data.

1 Development of a comprehensive structural variant haplotype map from high-coverage whole-genome sequencing (GCAT|panel)

The generation of the GCAT haplotype panel described in this thesis, not only has provided a valuable tool and resource for a wide range of GWAS and eQTL studies, for example, but has also highlighted many relevant aspects about the limitations and the opportunities in the area of genome analysis and its application to disease. I below highlight the most relevant points.

1.1 The use of *in silico* samples to evaluate the performance of variant callers

A primary challenge in creating haplotype reference panels is the accurate identification of variants. Currently, there are more than 150 variant callers available, which are computational tools designed to detect genetic variants from NGS data[52]. Each tool has distinct features and detection algorithms, resulting in varying performance levels, which makes selecting the optimal variant calling pipeline a challenge.

The considerable heterogeneity of solutions presents a challenge for the community in designing and implementing pipelines for variant identification. Choosing tools that maximize precision and recall in our pipelines necessitates extensive benchmarking and tuning of available options. Part of my work in this thesis addresses the implementation, benchmarking, and tuning of a comprehensive pipeline for variant identification from high-coverage WGS data, which was used to generate the GCAT haplotype panel.

One of the major issues for the benchmarking of variant callers is the overall lack of datasets with validated variants, particularly of SVs, which limits the possibilities of assessing the performance of each caller. While there are a few samples available for the benchmarking of SNVs and INDELS (e.g. NA12787, Methods section 1.1), the situation is more critical with SVs. In our study, we partially solved this problem by generating a *in silico* sample (a complete BAM) where we can insert and control all the variation we want (work developed by Dr. Valls).

However, while *in silico* samples offer valuable advantages for evaluating the performance of variant callers, they also have some limitations. One of the main drawbacks of *in silico* samples is that they may not fully capture the complexity and diversity present in real biological samples. Simulated data sets are generated based on known models and assumptions, which may not always accurately represent the true biology or sequencing errors present in real data. Consequently, the performance of a variant caller on *in silico* samples may not always translate to the same level of performance on real samples. Furthermore, *in silico* samples may not be able to account for all the intricacies of the sequencing process, such as library preparation, biases, and batch effects, which can have a significant impact on variant calling outcomes[215]. Nevertheless, despite their limitations, based on our experience and the orthogonal validations we generated, we demonstrate that *in silico* genomes are indeed a valuable and useful alternative to real data.

In conclusion, while *in silico* samples offer a valuable tool for benchmarking and enhancing variant callers, it is essential to supplement these analyses with real biological samples to ensure that findings can be generalized to real-world situations. This underlines the importance of creating a real sample containing validated SNVs, INDELs, and SVs for optimal benchmarking of variant callers.

1.2 Factors limiting the identification of variants

The accurate identification of variants is essential in genomic research and in medicine, as it usually provides the initial layer of information from which the scope and accuracy of the rest of the study and application will depend. While this is relatively critical for research, precise variant calling will become increasingly essential when applied to clinical decisions, such as selecting a treatment based on the presence or absence of a mutation. However, the variant calling process and the genotyping of variant calls remain challenging and can be influenced by various factors.

Among these factors, as shown in this thesis, the coverage at which genomes are sequenced strongly determines the accuracy and reliability of variant calling. Having a high sequencing coverage for the generation of the haplotype reference panel in our case has allowed us to identify up to three times more SVs than previous studies (e.g. 1000G) that relied on lower sequencing coverage. Higher coverage improves the chances of accurately identifying true genetic variants while minimizing the risk of false positives[216]. For instance, it allows for better identification and filtering of sequencing errors, which might otherwise be misclassified as variants. Sequencing at high coverage also provides more reads covering targeted variant breakpoints during SV calling, aiding in their resolution. These regions are particularly challenging, as SV breakpoints often occur in genomic regions with a higher content of repetitive sequences. Furthermore, these breakpoints typically have additional sequence deletions or additions, making the mapping of these reads onto the reference genome (without the variation) very challenging.

Another crucial factor that determines the design and implementation of variant calling pipelines, and ultimately, their overall accuracy, is the significant heterogeneity that exists among various variant callers, even for those targeting the same variant type. Moreover, there is considerable discrepancy in defining the coordinates and the nature of the variation. From our experience, detecting SNVs and INDELs is relatively consistent across the different solutions available. However, identifying and classifying SVs proved to be more challenging, as we observed discrepancies in structural variant detection when comparing results from different variant callers for the same sample. Due to these differences, and because each variant caller has its own optimal detection range, the community agrees that variant calling pipelines should incorporate multiple variant calling programs. The final results can then be derived from the analysis and filtering of the combined calls produced. In this case, the disagreement of these callers regarding the variant position introduces additional issues. Furthermore, the integration of calls from multiple variant callers to generate a final set of rules and lists of variants also depends on how each caller classifies the variant.

Lastly, another commonly observed difference is the discrepancies encountered when inferring the genotype for SVs from the sequencing data. Although two variant callers may find the variant in the same position and of the same type, we have observed discrepancies in the genotyping process, where one tool considers the variant homozygous and another heterozygous. This is a key factor, as accurate genotyping is essential in creating a reference panel and its use in further analysis.

The challenges in variant detection and genotyping, especially for SVs, as previously explained, primarily stem from different callers utilizing unique algorithms and statistical models to identify genetic variations. This results in variations among the detected variants, making it difficult to determine the most accurate set of variants.

To address this issue, we proposed a merging and acceptance model of variants based on an LRM. This model enabled us to determine the genomic position, variant type, variant size and genotype most likely to be accurate, which were reported in the final variant set. By leveraging the strengths of each variant caller used in the project, we obtained a set with the highest possible quality of variants. With this approach, we reduced the initial number of variants obtained from individual variant callers, from 58,529,907 variants (including 2,903,224 SVs), where we expect a high number of false positives, to a final set of 3,325,064 variants (including 89,178 SVs) with minimized number of false positives and maximized recall. The effectiveness of this strategy was demonstrated by achieving much higher F-score values (F-score = 0.9) compared to using the software individually (average F-score = 0.65) and a higher genotyping accuracy, with a genotype error rate of 5% as in contrast to the individual variant caller's average of 15%.

Thus, in this project, we demonstrated that by using a large number of variant callers and by applying robust statistical frames to merge and filter out their results, can generate a high-quality set of variants, reducing and mitigating the previously explained limitations.

1.3 Comparison of results against other databases and validation limitations

Another significant challenge I faced during my thesis was the difficulty in validating and comparing the results with other existing sources of information. When validating our set of variants, we opted for two strategies. The first one consisted of a comparative approach against public databases to determine how many of our variants had already been described and how many were new. In this way, we found that 9.18% (6.78 million) of our SNVs and INDELs were previously unclassified using dbSNP (v.153) as a reference, and 61% of the SVs we reported (49,333) had not been previously described, using gnomAD-SV, 1000G, GoNL, HGSVC, DGV, and dbVar as references.

In the case of SNVs and INDELs, as explained when discussing the limitations of variant calling, this type of comparisons are more straightforward. We compared the chromosomal coordinates and reference and alternative alleles to determine if a variant was new or had already been described.

For SVs, this comparison was more complex. As explained, some of the main limitations we encountered in the SV variant calling process are the breakpoint resolution where the variant is located, as well as the type of variant reported, both influenced by the software and coverage used. Each of the databases used for comparison followed a different strategy for detecting SVs, which requires us to make various decisions on when to consider two variants as the same. This, together with the different coverages used by these other studies, also challenged our overall comparison and validation.

Because of these limitations, we also opted for complementary validation strategies to complement our validation process, which consisted of validating different sets of variants from our catalog using experimental methodologies.

In the case of SNVs and INDELs, we used genotyping array data from 570 of the 785 samples used[197]. We decided to use this methodology because its cost is affordable, and it has a low error rate (≈0.8%)[80], as well as allowing us to validate both the presence of the variant and its genotype. In the case of SVs, the methodologies for validating these types of variants are less accurate and more economically costly. The main one is validation by PCR; however, we considered that validating a large number of SVs with this strategy is not viable from an economic standpoint. In our case, we limited our experimental validation of SVs to three of the most prevalent types in our variant set: deletions, duplications, and inversions (Results section 1.4). However, these strategies were very limiting due to the small number of variants that could be tested for each of the types.

In conclusion, we made a significant effort to validate our variant set in the best possible way to ensure that our variant set was of high quality. However, due to the lack of techniques that allow validating a large number of SVs quickly and economically feasible, this process only allowed us to determine that our detection approach was well-focused and that most of our variants, primarily those with high frequency and of greater value for a reference panel, were correct, as they also had already been reported in European populations or validated by other projects.

1.4 Including SVs in reference panels, is software ready?

In recent years, there has been a significant push to develop haplotype reference panels for use in genetic data analysis, particularly in the context of GWAS and eQTLs, utilizing them for phasing and imputation processes. Reference panels can be diverse, including those that encompass multiple populations, such as the 1000 Genomes Project (1000G), HRC, and TOPMed[217], or those that are specific to a particular population like GoNL, UK10K or the GCAT reference panelpresented in this thesis.

As previously mentioned, a common objective in these efforts is the focus on creating a high-quality catalog of SNVs and INDELs, while SVs are often deprioritized or even excluded. The identification of SVs has significantly improved in recent years, and the increased coverage of newer projects allows for the generation of high-quality variant sets. However, developing a reference panel requires more than just producing a list of high-quality variants. As detailed in this thesis, phasing must be applied to assign each variant to one of the two alleles and construct the haplotypes that will later be used in imputation to predict new variants based on linkage disequilibrium patterns.

During the development of the GCAT reference panel, we tested various phasing strategies in order to achieve the best possible imputation results, which entails imputing a larger number of high-quality variants (r^2>0.7). In our case, the most effective strategy combined Shapeit4 with WhatHap, which uses Phase Informative Reads (PIRs) to enhance phasing quality. With this approach, our panel was able to impute 98% of SNVs, 92% of INDELs, and 90% of common SVs observed in the variant calling for the same samples. This deliberate effort in carrying out efficient SV variant calling, as well as selecting a strategy that allows us to obtain the most comprehensive phasing and imputation results, enabled us to outperform both the 1000G and GoNL, two panels that include SVs. As a result, we achieved a 1.6, 2.7, and 1.3-fold increase in the number of high-quality imputed SVs compared to 1000G, GoNL, and both panels combined, respectively.

However, we observed that our phasing strategy treated SVs as point mutations, meaning they were treated like SNVs regardless of their actual type and size. Consequently, we identified some inconsistencies, such as imputing homozygous deletions (1/1) at a specific position while also imputing SNVs at positions that would be within the deletion due to its size. This led us to question the ability of phasing and imputation software to properly handle SVs. To date, we have not discovered any tools that can both methodologically perform phasing and imputation of SVs correctly and interpret the biology behind SVs, which is essential if we aim to study and systematically include these types of variants in reference panels and association studies.

1.5 The present value of haplotype reference panels

As WGS becomes increasingly standardized and accessible, the value of haplotype reference panels is being called into question for GWAS and eQTL analysis. Traditionally, haplotype reference panels have been used for imputing missing genotype data, an essential step in genetic data analysis. However, with WGS providing more comprehensive genomic information, some argue that haplotype reference panels will soon be obsolete. Therefore, it is important to critically assess their current value.

It is true that the cost of genomic sequencing has decreased significantly over the past decade, making WGS a more attractive option for researchers and clinicians alike[218]. However, it is important to recognize that WGS is still relatively expensive compared to genotyping arrays, which are often used in conjunction with haplotype reference panels. For large-scale studies or those with limited budgets, these panels can provide a cost-effective alternative to WGS, allowing for the analysis of genetic data without incurring prohibitive costs.

Even if sequencing any sample becomes economically affordable, there is an issue that is often overlooked when claiming that WGS is the present: the computational resources and the complexity usually associated with the variant calling [219]. WGS generates a large amount of data, which requires significant computational resources for the identification of variants. For example, in the GCAT project we used ~3.5M CPU/hours in the variant calling process for 785 WGS samples, which equates to €820,000 in electricity and computational costs. In comparison, using haplotype reference panels with genotyping arrays involves less data, smaller size, and lower computational needs for analysis, directly impacting the economic cost of this type of analysis. This makes the combination of genotyping arrays and haplotype reference panels a more feasible option for researchers with limited computational resources.

On the other hand, there are methodological criticisms of imputation. Critics argue that imputation accuracy can be affected by factors such as the reference panel's representativeness of the population under study and the quality of the genotyping data[220]. As a result, there is potential for biases and inaccuracies in imputed data. However, in our experience, imputation has been shown to work effectively. Using the GCAT haplotype reference panel, we were able to impute a total of 14,383,907 high-quality variants (info score ≥ 0.7) and MAF > 0.001. Moreover, our analyses have demonstrated that the GCAT reference panel allows for high-quality imputation of variants in distant populations, even if they are not represented in the panel. In this way, we found that although the European population had the highest number of imputed SVs, the GCAT panel was also capable of imputing a significant fraction of SVs in other ethnicities, including African, Latin American, and Asian populations.

In addition, critics contend that WGS-based GWASs provide a more comprehensive view of the genome, including structural variations and rare variants that may not be captured by haplotype reference panels and genotyping arrays, both of which play an important role in diseases and other phenotypes[64]. While it is true that recovering low-frequency variants using reference panels is limited, in the GCAT project, we made an effort to include SVs, imputing an average of 5,120 SVs (SD=50) per sample out of an average of 6,393 SVs (SD=62) obtained from variant calling, resulting in an imputation rate of 80%. This demonstrates that this reference panel can consistently include SVs, making the advantage of WGS in this area not as significant.

In conclusion, while the world is moving towards the standardization of WGS, haplotype reference panels still hold value for many types of genetic research. Their lower economic cost, reduced computational resource requirements, and high imputation accuracy make them a viable option for researchers with budgetary or computational constraints. However, it is essential to recognize the limitations of imputation and the potential benefits of WGS in uncovering the full spectrum of genomic variation. As technology advances and the costs of WGS continue to decline, the balance between the use of haplotype reference panels and WGS may shift, but for now, haplotype reference panels remain a valuable tool.

1.6 The future of SVs analysis with long-reads integration

In the development of the GCAT haplotype reference panel, we have focused our efforts on achieving effective and comprehensive characterization of SVs. To do this, as mentioned, we have combined different variant callers with various detection algorithms and an LRM for filtering the variant set to generate a high-quality set, capitalizing on the strengths of each tool used. Finally, we have followed a phasing strategy that, although it does not consider the type of SVs and treats them as variants affecting a single nucleotide, allows us to impute them correctly, thereby creating a high-quality haplotype reference panel.

Our entire strategy has been applied to samples obtained from short-read sequencing, a technology that, although it allows for the detection of SVs, is limiting due to the read size[48]. In recent years, sequencing technology has rapidly evolved, with long-read sequencing gaining traction as a promising alternative to short-read sequencing[221]. This technology, as explained, is an NGS method that generates significantly longer DNA reads compared to short-read sequencing technologies.

The advantages of long-read sequencing include improved detection of SVs, as long-read sequencing can span larger genomic regions, allowing for more accurate identification of SVs, which are often missed or incorrectly characterized by short-read sequencing[11]. Long-read sequencing also provides enhanced resolution of repetitive regions, which are difficult to resolve using short-read sequencing due to their repetitive nature. Long-read sequencing can span these regions, enabling a more accurate assembly and characterization of these complex genomic regions[11]. These advantages can be observed when looking at the results of variant calling obtained from this technology. For example, in a recent study, the analysis of nine long-read sequencing samples identified and validated 93,852 SVs[70], compared to the 89,178 total SVs detected in the GCAT project using 785 short-read sequencing samples.

Despite these advantages, long-read sequencing also has some limitations in terms of cost and the type of results they offer. Economically, long-read sequencing technologies, such as Pacific Biosciences and Oxford Nanopore, are still expensive compared to short-read sequencing platforms, making their systematic inclusion in studies less affordable.

In terms of performance, long-read sequencing technologies tend to have a higher error rate compared to short-read sequencing, particularly in detecting SNVs and INDELs[48], which are the basis of haplotype reference panels. This means that, despite their high accuracy in detecting SVs, this technology is limited when it comes to detecting smaller variants.

Therefore, in my opinion, combining long-read technology, which allows for accurate detection of SVs, and short-read technology, for the precise detection of SNVs and INDELs, would enable the creation of the most comprehensive reference panel to date. This panel would significantly impact evolutionary and biomedical studies at different levels, directly increasing the chances of variant discovery, as well as their functional interpretations.

2 Recovering the role of the X chromosome in complex diseases

Personalized medicine is an expanding field that aims to customize healthcare interventions based on each individual's distinct genetic, environmental, and lifestyle factors. A critical aspect of personalized medicine is understanding the genetic foundations of complex diseases to ensure its proper application. However, our understanding of the genetic architecture underlying complex diseases remains incomplete, with a considerable knowledge gap regarding the role of the X chromosome.

The X chromosome, being the 8th largest human chromosome and containing almost 5% of the genome, holds significant importance not only for its size and genetic content but also for its unique biological characteristics[156]. While males have one copy of this chromosome, females have two, but one of the copies is theoretically inactivated to balance the difference in dosage[158]; however, the phenomenon of escape from inactivation may occur, potentially leading to differences in gene expression patterns between females[160]. These distinctive features make the X chromosome a prime candidate for investigating not only its importance in complex diseases but also its impact in potential sex differences in genetics and disease. The importance of these sex differences in personalized medicine cannot be overstated, as accounting for these differences is crucial for providing more accurate, tailored, and effective healthcare interventions[222].

Despite its importance, the X chromosome has been systematically excluded from GWAS studies primarily due to the extra methodological costs associated with its analysis[172]. Aiming to fill up this gap, we developed a specific strategy to be able to properly analyze this chromosome in GWAS studies. The strategy and results presented in this thesis not only show the identification of new genetic markers associated with various complex diseases on the X chromosome, but also offer the scientific community methodological insights and guidance to encourage the inclusion of this chromosome's analysis in ongoing and future genetic studies.

2.1 Development of a strategy to analyze the X chromosome

Genetic associations between diseases and markers are commonly identified using GWAS, which rely on case-control study designs. These designs compare genotype frequency distributions between cases and controls, determining if a variant is more prevalent in one group than in another and thereby associated with the disease. While GWAS is well-established for autosomes, the techniques used for autosomal genotype data are not directly applicable to X-chromosome analysis due to potential statistical inaccuracies. This has led to the X chromosome's systematic exclusion from GWAS.

This specific behavior of the X chromosome in GWAS studies derives from the unique biological features of the X chromosome, which includes males having one copy and females having two alleles, but only one active. This is due to the X-chromosome inactivation (XCI)[223]. In addition, it is also known that some regions of the inactivated copy escape this process and become functional. Therefore, whereas in males there is genetically and functionally one single copy, the situation in females is more complex, as although they are genetically diploid for the X chromosome, they are actually also functionally haploid, except for those regions that partially escape the inactivation.

These factors necessitate accounting for genetic imbalances between the sexes and expected XCI during X chromosome analysis[224–226]. Despite this, the scientific community has not prioritized developing methodologies or strategies for proper X chromosome study. This is, in my opinion, likely because it has been easier to obtain significant publications or generate interesting scientific results without the need to study or incorporate the X chromosome in the analysis[227]. However, recent projects have emphasized the X chromosome's importance and suggested strategies specific to X chromosome genetic analysis to address these unique challenges[177].

In this thesis, I have presented a strategy that identifies and tests associations between a disease and X chromosome genetic markers by analyzing genetic data in a sex-stratified manner. We analyzed each sex separately including all project steps, from initial quality control to association, and ultimately meta-analyzing the results obtained. In this way, this approach is also particularly relevant for analyzing sex differences, where we observed signals exclusive to one sex or signals with different effects between males and females. Furthermore, by using this approach, we did not have to deal with problems arising from genotype coding, as the X chromosome is analyzed separately in males and females without requiring assumptions about XCI[228].

Applying this strategy has allowed us to discover previously not described significant associations and find new associations using public cohorts previously analyzed with other approaches. Among the most noteworthy results is the comparison of our UK Biobank analysis with that of the PanUKBB project. When comparing our results to the prior UK Biobank analysis by the PanUKBB project, we found similar numbers of associations from the "standard analysis" (males and females analyzed together) between projects. The differences lay in the sex-specific associations resulting from the sex-stratified analysis. Considering sex-specific associations, we obtained twice the number of total associations as the PanUKBB project, yielding a figure much closer to the expected number assuming a direct relationship between the number of associations on a chromosome and its size, using autosomes as reference.

This demonstrates that our strategy enables the accurate study of the X chromosome and equates it to autosomes in terms of associations discovered per chromosome size. Additionally, this strategy has helped identify genetic variants associated with diseases in only one sex and variants with different effect between sexes, leading to the discovery of potential sex differences.

However, this strategy also presents some limitations. For example, sex stratification reduces the sample size and, consequently, the statistical power to detect associations[228]. Additionally, the differences observed in the signals obtained from sex-specific analyses do not provide enough information to determine if a signal is unique to one sex and is, therefore, a sex difference, or if it has not been found in the other sex due to lack of power. This invites alternative analyses to determine when a result truly represents a sex difference.

2.2 Constraints when exploring gender specificity

As previously mentioned, one of the advantages of conducting a sex-stratified analysis is the ability to subsequently combine the summary statistics of each sex to look for sex differences, either due to the presence of a signal in one sex but not the other, or because the direction of the signal's effect is different between sexes, resulting, for example, in a risk variant in males but a protective variant in females.

However, with the obtained results, one of the main issues we have observed is determining whether these associations were genuinely sex-specific or if they could be attributed to a lack of statistical power in one sex or to how the methodology and tools used in the association process deal with the biological differences that exist between males and females in the X chromosome.

In this context, as mentioned before, it is important to note that while males are hemizygous, females have two copies of the chromosome, leading to the presence of heterozygous genotypes. Theoretically, one of the two chromosomes is silenced to achieve dosage compensation, but the inactivation process is not captured by the genotypes. Moreover, there is the phenomenon of escape from XCI, which adds another layer of complexity to the analysis.

In our case, to further explore signals showing sex differences and the impact of heterozygous genotypes in females on the discovery of these signals, we have proposed a simulation in which we remove all heterozygous genotypes in females, mimicking what we observe at the biological level in males.

Our hypothesis was to determine if signals found only in females remained significant once heterozygous genotypes were removed, or if we found any new significant signals, determining if heterozygous genotypes had any effect on the signal or added noise to the analysis. In other words, to see if we could recover some new signals that were masked by noise, assuming random inactivation, or to determine that the different effects we were seeing between sexes can be due to inactivation escape.

However, applying this strategy does not allow us to efficiently compare the initial results with those obtained by removing heterozygotes, mainly due to a decrease in both the number of interrogated variants and the sample size, by removing heterozygous samples for each variant. Despite this, it did give us indications of the impact of heterozygous genotypes, finding that 72% of loci that could be compared were rediscovered after removing heterozygotes. Notably, four new loci were found in this analysis, showing that the presence of heterozygotes does have an impact, which should be studied in more depth.

Currently, various projects are emerging proposing different models to study the X chromosome and possible sex differences, as well as dealing with inactivation phenomena in different cases, assuming complete escape from inactivation, assuming random inactivation, or assuming skewed inactivation[177]. Therefore, it would be interesting to expand the study presented here using these strategies, complementing the results presented from the sex-stratified analysis.

2.3 Dealing with multiple testing correction in GWAS

One of the main challenges inherent in the use of statistical methods in the association analysis step is the issue of multiple testing. Given that thousands or even millions of genetic markers are examined, the probability of encountering false positive associations increases substantially. To account for this, stringent significance thresholds are imposed. Nonetheless, this strict threshold may inadvertently result in false negatives, whereby true associations are overlooked.

Currently, the most widely accepted and applied threshold in the scientific community is the p-value threshold of $5x10^{-8}$. This value was estimated in 2005 by estimating the number of common independent variants in the genome and applying the Bonferroni correction[116]. However, as sequencing technologies and genetic data analysis have evolved, GWAS have incorporated variants of lower frequency, leading the scientific community to question whether the use of this threshold is still appropriate today[209].

In our study of the X chromosome, we decided to determine our own significance threshold based on the number of independent tests analyzed, which we obtained from the imputed data of the X chromosome in different cohorts aiming to use a threshold more suitable for the type of analysis and data we were using. We estimated our significance threshold considering that our study was conducted solely on the X chromosome, rather than using the whole genome, which implies a smaller number of tests and therefore a better adjusted threshold. As a result, we were able to identify associations that would not have reached the level of significance using a p-value threshold of $5x10^{-8}$ and would have been otherwise overlooked.

With this, I considered that the systematic application of this strategy in GWAS studies, obtaining a specific threshold for the project based on the type of analysis or the type of analyzed data, would help avoid potential biases resulting from the use of a standardized threshold calculated nearly 20 years ago, in which for example, low-frequency variants, or variants such as INDELs or SVs, now included in most studies, were not considered.

2.4 The value of replication in current GWAS studies

One of our main decisions in the analysis of the X chromosome was to work with genetic data obtained from large-scale biobanks and disease-specific cohorts. This way, we could increase the number of phenotypes analyzed but also analyze some phenotypes in both types of cohorts, taking advantage of the benefits of each. In addition, analyzing the same phenotype in the two sources we could determine if we were able to replicate the significant associations found, i.e., validating the association discovered from one data source independently in the other. However, discrepancies in the findings may arise not only due to differences in recruitment and classification protocols or in sample sizes between biobanks and disease-specific cohorts, but also due to differences in genotyping platforms[38] and statistical methods[218] used in the analysis of the two data sources.

This leads us to question whether the lack of replication of an association in the same phenotype between the two data sources sheds light on the validity of the association. In fact, in our study, we determined how many significant associations present in phenotypes found in both the UK Biobank and dbGaP could be replicated. In this way, only 13% (4/31) of associations present in traits that matched in both sources demonstrated replication between the two analyses. This would imply that, if we were extremists, we should not believe the other 27 associations since they were not replicated in another cohort with the same phenotype.

I agree that the replication of GWAS findings helps establish robustness and validity for those genetic associations that are replicated[229]. However, I do not think that associations that fail to be replicated in another cohort should be dismissed or lose credibility because the replication process is plagued by challenges[229]. To address these issues, I believe that the scientific community should question the current importance of replication studies, adopt standardized methodologies that take into account the differences between the cohorts analyzed, or not consider replication as a fundamental criterion for determining whether an association is true or not, but rather use replication as one more point when assessing the validity of an association.

2.5 From GWAS to clinical application

In this thesis, by analyzing the X chromosome, we have made significant contributions to our understanding of the genetic underpinnings of complex traits and diseases, adding a chromosome that until now has been mostly neglected. However, it is essential to critically assess the real value of GWAS findings in terms of advancing human health and personalized medicine.

One of the primary goals of GWAS is to improve our understanding of disease mechanisms and identify potential therapeutic targets. Our discovery of genetic associations in the X chromosome, along with our fine-mapping and pathway enrichment analysis process, can shed light on the biological pathways and molecular mechanisms underlying the diseases studied, thereby providing valuable information for developing treatment strategies[230]. Furthermore, the identification of sex differences can also add significant value in early disease prediction and prevention, allowing for targeted interventions and lifestyle modifications to mitigate disease risk according to the sex of the individual[231].

Another promising application of GWAS findings lies in predicting an individual's genetic risk of disease development. By integrating individual genetic data with GWAS results, polygenic risk scores (PRS) can be generated. These scores estimate an individual's genetic predisposition to specific diseases or traits, using summary statistics from GWAS. PRS have shown promise in predicting risks for various conditions, such as coronary artery disease[232], breast cancer[233], and psychiatric disorders[234]. However, the clinical utility of PRS is still a topic of debate, as their predictive accuracy varies across populations and sexes[235]. Here, performing a sex-stratified analysis, not only for the X chromosome as we have done but also for autosomes, would enable the generation of data for each sex independently. This approach could help address some of the limitations associated with using PRS in risk prediction.

In conclusion, the real value of GWAS lies in their ability to advance our understanding of complex traits and diseases, identify potential therapeutic targets, and inform personalized medicine. However, translating GWAS findings into practical applications is a complex process fraught with challenges. To fully realize the potential of GWAS, it is imperative to address these limitations, conduct more inclusive studies like the one presented in this thesis, and develop rigorous methodologies to ensure the robustness and clinical relevance of genetic discoveries.

Conclusions

1) We have designed and implemented a comprehensive strategy for the identification of germline variants from Whole Genome Sequencing Data. This strategy is particularly efficient in the discovery and classification of Structural Variants.

2) The application of this strategy to 785 high-coverage short read whole-genome sequencing samples from the GCAT cohort allowed us to identify an average of about 4 million variants per individual, including 6,393 structural variants.

3) From these variants and from the phasing of their derived genotypes, we also generated a comprehensive haplotype map for the Iberian population. With this map, we are able to impute up to 14,3 Million SNVs and INDELs and 23 thousand SVs, even in genetically distant populations. This represents a 2.7-fold increase for SVs, compared with commonly used genetic variability panels.

4) We have also designed and implemented a comprehensive strategy for the efficient analysis of the X-Chromosome within association studies. We applied this strategy to 600 disease across 800 thousand individuals, allowing us to identify up to 74 new significant genetic associations in this chromosome.

5) The separate analysis of males and females within our strategy has provided 36 new associations when compared to the results obtained by other projects without applying this approach. It proves the importance of adapting genetic strategies to particular genetic and biological scenarios.

6) Preliminary functional analyses of these sex specific associated variants show different female-enriched pathways that agree with already reported sex differences in phenotypes.

References

1. Valls-Margarit, J. *et al.* GCAT|Panel, a comprehensive structural variant haplotype map of the Iberian population from high-coverage whole-genome sequencing. *Nucleic Acids Research* **50**, 2464–2479 (2022).
2. Cobb, M. Heredity before genetics: a history. *Nat Rev Genet* **7**, 953–958 (2006).
3. Liu, Y. Like father like son. A fresh review of the inheritance of acquired characteristics. *EMBO Rep* **8**, 798–803 (2007).
4. Charlesworth, B. & Charlesworth, D. Darwin and genetics. *Genetics* **183**, 757–766 (2009).
5. CR, D. The Origin of Species. *London*.
6. CR, D. Variation of Animals and Plants Under Domestication. *London*.
7. Haynes, R. H. Heritable variation and mutagenesis at early International Congresses of Genetics. *Genetics* **148**, 1419–1431 (1998).
8. Dahm, R. Friedrich Miescher and the discovery of DNA. *Developmental Biology* **278**, 274–288 (2005).
9. Griffiths, A. J. F. An introduction to genetic analysis. *New York* (2000).
10. Lobo, I. & Shaw, K. Discovery and Types of Genetic Linkage. *Nature Education* **1**, 139 (2008).
11. Fisher, R. A. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Trans. R. Soc. Edinb.* **52**, 399–433 (1919).
12. Fisher, R. A. *The genetical theory of natural selection*. (Clarendon Press, 1930). doi:10.5962/bhl.title.27468.
13. Miescher, F. Ueber die chemische Zusammensetzung der Eiterzellen. *Med-Chem Unters* **4**, 441–460.
14. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (1953).
15. Nirenberg, M. W. & Matthaei, J. H. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. U.S.A.* **47**, 1588–1602 (1961).
16. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends in Genetics* **17**, 502–510 (2001).

17. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463–5467 (1977).

18. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science* **291**, 1304–1351 (2001).

19. Wang, D. G. *et al.* Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. *Science* **280**, 1077–1082 (1998).

20. †The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).

21. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

22. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics* **13**, 278–289 (2015).

23. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).

24. Tryka, K. A. *et al.* NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucl. Acids Res.* **42**, D975–D979 (2014).

25. Ku, C. S., Loy, E. Y., Salim, A., Pawitan, Y. & Chia, K. S. The discovery of human genetic variations and their use as disease markers: past, present and future. *J Hum Genet* **55**, 403–415 (2010).

26. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

27. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

28. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

29. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

30. The UK10K Consortium *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).

31. Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nat Rev Genet* **10**, 241–251 (2009).

32. Slatkin, M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* **9**, 477–485 (2008).

33. Crawford, D. C. & Nickerson, D. A. Definition and Clinical Importance of Haplotypes. *Annu. Rev. Med.* **56**, 303–320 (2005).

34. Hofker, M. H., Fu, J. & Wijmenga, C. The genome revolution and its role in understanding complex diseases. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1842**, 1889–1895 (2014).

35. Ragoussis, J. Genotyping technologies for all. *Drug Discovery Today: Technologies* **3**, 115–122 (2006).

36. Gunderson, K. L., Steemers, F. J., Lee, G., Mendoza, L. G. & Chee, M. S. A genome-wide scalable SNP genotyping assay using microarray technology. *Nat Genet* **37**, 549–554 (2005).

37. Ragoussis, J. & Elvidge, G. Affymetrix GeneChip® system: moving from research to the clinic. *Expert Review of Molecular Diagnostics* **6**, 145–152 (2006).

38. Verlouw, J. A. M. *et al.* A comparison of genotyping arrays. *Eur J Hum Genet* **29**, 1611–1624 (2021).

39. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**, 363–376 (2011).

40. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).

41. Metzker, M. L. Sequencing technologies — the next generation. *Nat Rev Genet* **11**, 31–46 (2010).

42. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat Biotechnol* **26**, 1135–1145 (2008).

43. Xiao, T. & Zhou, W. The third generation sequencing: the advanced approach to genetic diseases. *Transl Pediatr* **9**, 163–173 (2020).

44. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol* **17**, 239 (2016).

45. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

46. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

47. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

48. Mahmoud, M. *et al.* Structural variant calling: the long and the short of it. *Genome Biol* **20**, 246 (2019).

49. Webster, T. H. *et al.* Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data. *Gigascience* **8**, giz074 (2019).

50. Balachandran, P. & Beck, C. R. Structural variant identification and characterization. *Chromosome Res* **28**, 31–47 (2020).

51. Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol* **20**, 117 (2019).

52. Geoffroy, V. *et al.* AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics* **34**, 3572–3574 (2018).

53. GTEx Consortium *et al.* The impact of structural variation on human gene expression. *Nat Genet* **49**, 692–699 (2017).

54. Kishikawa, T. *et al.* Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data. *Sci Rep* **9**, 1784 (2019).

55. Rahman, N. Realizing the promise of cancer predisposition genes. *Nature* **505**, 302–308 (2014).

56. Kumaran, M., Subramanian, U. & Devarajan, B. Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data. *BMC Bioinformatics* **20**, 342 (2019).

57. the Global Alliance for Genomics and Health Benchmarking Team *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol* **37**, 555–560 (2019).

58. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

59. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. (2012) doi:10.48550/ARXIV.1207.3907.

60. WGS500 Consortium *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet* **46**, 912–918 (2014).

61. Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).

62. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**, 591–594 (2018).

63. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* **36**, 983–987 (2018).

64. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).

65. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* **15**, 121–132 (2014).

66. NHGRI Centers for Common Disease Genomics *et al.* Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**, 83–89 (2020).

67. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).

68. MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucl. Acids Res.* **42**, D986–D992 (2014).

69. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10**, 1784 (2019).

70. Audano, P. A. *et al.* Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**, 663-675.e19 (2019).

71. Lin, K., Smit, S., Bonnema, G., Sanchez-Perez, G. & de Ridder, D. Making the difference: integrating structural variation detection tools. *Brief Bioinform* **16**, 852–864 (2015).

72. Guan, P. & Sung, W.-K. Structural variation detection using next-generation sequencing data. *Methods* **102**, 36–49 (2016).

73. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**, R84 (2014).

74. Kronenberg, Z. N. *et al.* Wham: Identifying Structural Variants of Biological Consequence. *PLoS Comput Biol* **11**, e1004572 (2015).

75. Kehr, B., Melsted, P. & Halldórsson, B. V. PopIns: population-scale detection of novel sequence insertions. *Bioinformatics* **32**, 961–967 (2016).

76. Gardner, E. J. *et al.* The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res.* **27**, 1916–1929 (2017).

77. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).

78. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).

79. Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* **12**, 443–451 (2011).

80. Pompanon, F., Bonin, A., Bellemain, E. & Taberlet, P. Genotyping errors: causes, consequences and solutions. *Nat Rev Genet* **6**, 847–859 (2005).

81. Maruki, T. & Lynch, M. Genotype Calling from Population-Genomic Sequencing Data. *G3 Genes|Genomes|Genetics* **7**, 1393–1404 (2017).

82. Eggertsson, H. P. *et al.* GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat Commun* **10**, 5402 (2019).

83. The Danish Pan-Genome Consortium, Sibbesen, J. A., Maretty, L. & Krogh, A. Accurate genotyping across variant classes and lengths using variant graphs. *Nat Genet* **50**, 1054–1059 (2018).

84. Lecompte, L., Peterlongo, P., Lavenier, D. & Lemaitre, C. SVJedi: genotyping structural variations with long reads. *Bioinformatics* **36**, 4568–4575 (2020).

85. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489–494 (2009).

86. Exome Aggregation Consortium *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

87. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era — concepts and misconceptions. *Nat Rev Genet* **9**, 255–266 (2008).

88. Tenesa, A. & Haley, C. S. The heritability of human disease: estimation, uses and abuses. *Nat Rev Genet* **14**, 139–149 (2013).

89. Vinkhuyzen, A. A. E., Wray, N. R., Yang, J., Goddard, M. E. & Visscher, P. M. Estimation and partition of heritability in human populations using whole-genome analysis methods. *Annu Rev Genet* **47**, 75–95 (2013).

90. Zaitlen, N. *et al.* Leveraging population admixture to characterize the heritability of complex traits. *Nat Genet* **46**, 1356–1362 (2014).

91. Haworth, C. M. A. *et al.* The heritability of general cognitive ability increases linearly from childhood to young adulthood. *Mol Psychiatry* **15**, 1112–1120 (2010).

92. Stringer, S., Polderman, T. J. C. & Posthuma, D. Majority of human traits do not show evidence for sex-specific genetic and environmental effects. *Sci Rep* **7**, 8688 (2017).

93. Yong, S. Y., Raben, T. G., Lello, L. & Hsu, S. D. H. Genetic architecture of complex traits and disease risk predictors. *Sci Rep* **10**, 12055 (2020).

94. Janssens, A. C. J. W. & van Duijn, C. M. Genome-based prediction of common diseases: advances and prospects. *Human Molecular Genetics* **17**, R166–R173 (2008).

95. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378**, 789–792 (1995).

96. Gusella, J. F. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234–238 (1983).

97. Macdonald, M. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971–983 (1993).

98. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *The American Journal of Human Genetics* **97**, 199–215 (2015).

99. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

100. Lowe, W. L. & Reddy, T. E. Genomic approaches for understanding the genetics of complex disease. *Genome Res* **25**, 1432–1441 (2015).

101. GoT2D Consortium, Agarwala, V., Flannick, J., Sunyaev, S. & Altshuler, D. Evaluating empirical bounds on complex disease genetic architecture. *Nat Genet* **45**, 1418–1427 (2013).

102. Pritchard, J. K. Are Rare Variants Responsible for Susceptibility to Complex Diseases? *The American Journal of Human Genetics* **69**, 124–137 (2001).

103. Schork, N. J. Genetics of Complex Disease: Approaches, Problems, and Solutions. *Am J Respir Crit Care Med* **156**, S103–S109 (1997).

104. Gibson, G. Rare and common variants: twenty arguments. *Nat Rev Genet* **13**, 135–145 (2012).

105. Divo, M. J., Martinez, C. H. & Mannino, D. M. Ageing and the epidemiology of multimorbidity. *Eur Respir J* **44**, 1055–1068 (2014).

106. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* **33**, 228–237 (2003).

107. Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. & Hirschhorn, J. N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* **33**, 177–182 (2003).

108. Klein, R. J. *et al.* Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science* **308**, 385–389 (2005).

109. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. *The American Journal of Human Genetics* **90**, 7–24 (2012).

110. Gurinovich, A. *et al.* Evaluation of GENESIS, SAIGE, REGENIE and fastGWA-GLMM for genome-wide association studies of binary traits in correlated data. *Front. Genet.* **13**, 897210 (2022).

111. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* **50**, 1335–1341 (2018).

112. Alonso, L., Morán, I., Salvoro, C. & Torrents, D. In Search of Complex Disease Risk through Genome Wide Association Studies. *Mathematics* **9**, 3083 (2021).

113. Hunter, D. J. *et al.* A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* **39**, 870–874 (2007).

114. Statistical analysis for genome-wide association study. *J Biomed Res* (2015) doi:10.7555/JBR.29.20140007.

115. Gao, X., Becker, L. C., Becker, D. M., Starmer, J. D. & Province, M. A. Avoiding the high Bonferroni penalty in genome-wide association studies. *Genet Epidemiol* **34**, 100–105 (2010).

116. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* **32**, 381–385 (2008).

117. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucl. Acids Res.* **42**, D1001–D1006 (2014).

118. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics* **101**, 5–22 (2017).

119. Hood, L. & Rowen, L. The human genome project: big science transforms biology and medicine. *Genome Med* **5**, 79 (2013).

120. Lappalainen, I. *et al.* The European Genome-phenome Archive of human data consented for biomedical research. *Nat Genet* **47**, 692–695 (2015).

121. Rehm, H. L. *et al.* GA4GH: International policies and standards for data sharing across genomic research and healthcare. *Cell Genomics* **1**, 100029 (2021).

122. Bahcall, O. G. UK Biobank — a new era in genomic medicine. *Nat Rev Genet* **19**, 737–737 (2018).

123. Price, A. L., Spencer, C. C. A. & Donnelly, P. Progress and promise in understanding the genetic basis of common diseases. *Proc. R. Soc. B.* **282**, 20151684 (2015).

124. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

125. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).

126. Halldorsson, B. V. *et al.* The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).

127. Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* **40**, 695–701 (2008).

128. Alioto, T. S. *et al.* A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat Commun* **6**, 10001 (2015).

129. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* **11**, 415–425 (2010).

130. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**, 333–351 (2016).

131. the Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279–1283 (2016).

132. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906–913 (2007).

133. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499–511 (2010).

134. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**, 955–959 (2012).

135. Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *The American Journal of Human Genetics* **103**, 338–348 (2018).

136. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284–1287 (2016).

137. Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F. & Marchini, J. Haplotype Estimation Using Sequencing Reads. *The American Journal of Human Genetics* **93**, 687–696 (2013).

138. Delaneau, O. *et al.* Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun* **5**, 3934 (2014).

139. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179–181 (2012).

140. Hehir-Kwa, J. Y. *et al.* A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat Commun* **7**, 12989 (2016).

141. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* **14**, 125–138 (2013).

142. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45**, D896–D901 (2017).

143. Rioux, J. D. *et al.* Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet* **39**, 596–604 (2007).

144. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

145. Duerr, R. H. *et al.* A Genome-Wide Association Study Identifies *IL23R* as an Inflammatory Bowel Disease Gene. *Science* **314**, 1461–1463 (2006).

146. the NIDDK IBD Genetics Consortium *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn’s disease. *Nat Genet* **40**, 955–962 (2008).

147. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic Mapping in Human Disease. *Science* **322**, 881–888 (2008).

148. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 9362–9367 (2009).

149. the DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium *et al.* Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat Genet* **47**, 1415–1425 (2015).

150. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet* **19**, 581–590 (2018).

151. Nishino, J., Ochi, H., Kochi, Y., Tsunoda, T. & Matsui, S. Sample Size for Successful Genome-Wide Association Study of Major Depressive Disorder. *Front. Genet.* **9**, 227 (2018).

152. Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. & Visscher, P. M. Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. *Cell* **173**, 1573–1580 (2018).

153. Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol* **18**, 77 (2017).

154. Guindo-Martínez, M. *et al.* The impact of non-additive genetic associations on age-related complex diseases. *Nat Commun* **12**, 2436 (2021).

155. Schurz, H. *et al.* The X chromosome and sex-specific effects in infectious disease susceptibility. *Hum Genomics* **13**, 2 (2019).

156. Balaton, B. P., Dixon-McDougall, T., Peeters, S. B. & Brown, C. J. The eXceptional nature of the X chromosome. *Human Molecular Genetics* **27**, R242–R249 (2018).

157. Lyon, M. F. Gene Action in the X-chromosome of the Mouse (Mus musculus L.). *Nature* **190**, 372–373 (1961).

158. Galupa, R. & Heard, E. X-Chromosome Inactivation: A Crossroads Between Chromosome Architecture and Gene Regulation. *Annu. Rev. Genet.* **52**, 535–566 (2018).

159. Maxfield Boumil, R. Forty years of decoding the silence in X-chromosome inactivation. *Human Molecular Genetics* **10**, 2225–2232 (2001).

160. Posynick, B. J. & Brown, C. J. Escape From X-Chromosome Inactivation: An Evolutionary Perspective. *Front. Cell Dev. Biol.* **7**, 241 (2019).

161. Borsani, G. *et al.* Characterization of a murine gene expressed from the inactive X chromosome. *Nature* **351**, 325–329 (1991).

162. Oliva, M. *et al.* The impact of sex on gene expression across human tissues. *Science* **369**, eaba3066 (2020).

163. Disteche, C. M. Dosage compensation of the sex chromosomes and autosomes. *Seminars in Cell & Developmental Biology* **56**, 9–18 (2016).

164. Khramtsova, E. A., Davis, L. K. & Stranger, B. E. The role of sex in the genomics of human complex traits. *Nat Rev Genet* **20**, 173–190 (2019).

165. Bernabeu, E. *et al.* Sex differences in genetic architecture in the UK Biobank. *Nat Genet* **53**, 1283–1289 (2021).

166. GTEx Consortium *et al.* Landscape of X chromosome inactivation across human tissues. *Nature* **550**, 244–248 (2017).

167. Robinson, E. B., Lichtenstein, P., Anckarsäter, H., Happé, F. & Ronald, A. Examining and interpreting the female protective effect against autistic behavior. *Proc. Natl. Acad. Sci.*

U.S.A. **110**, 5258–5262 (2013).

168. Ngo, S. T., Steyn, F. J. & McCombe, P. A. Gender differences in autoimmune disease. *Frontiers in Neuroendocrinology* **35**, 347–369 (2014).
169. Feuk, L. *et al.* Absence of a paternally inherited FOXP2 gene in developmental verbal dyspraxia. *Am J Hum Genet* **79**, 965–972 (2006).
170. Al-Sayed, M. D. *et al.* Mutations in NALCN cause an autosomal-recessive syndrome with severe hypotonia, speech impairment, and cognitive delay. *Am J Hum Genet* **93**, 721–726 (2013).
171. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research* **51**, D977–D985 (2023).
172. Wise, A. L., Gyi, L. & Manolio, T. A. eXclusion: toward integrating the X chromosome in genome-wide association analyses. *Am J Hum Genet* **92**, 643–647 (2013).
173. Tukiainen, T. *et al.* Chromosome X-Wide Association Study Identifies Loci for Fasting Insulin and Height and Evidence for Incomplete Dosage Compensation. *PLoS Genet* **10**, e1004127 (2014).
174. König, I. R., Loley, C., Erdmann, J. & Ziegler, A. How to Include Chromosome X in Your Genome-Wide Association Study. *Genet. Epidemiol.* **38**, 97–103 (2014).
175. Ma, L., Hoffman, G. & Keinan, A. X-inactivation informs variance-based testing for X-linked association of a quantitative trait. *BMC Genomics* **16**, 241 (2015).
176. Bonàs-Guarch, S. *et al.* Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes. *Nat Commun* **9**, 321 (2018).
177. Chen, B., Craiu, R. V., Strug, L. J. & Sun, L. The X factor: A robust and powerful approach to X-chromosome-inclusive whole-genome association studies. *Genet. Epidemiol.* **45**, 694–709 (2021).
178. Migeon, B. R. X-linked diseases: susceptible females. *Genetics in Medicine* **22**, 1156–1174 (2020).
179. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).
180. The Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113–1120 (2013).

181. Genome in a bottle—a human DNA standard. *Nat Biotechnol* **33**, 675–675 (2015).

182. Van der Auwera, G. A. *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *CP in Bioinformatics* **43**, (2013).

183. Poplin, R. *et al. Scaling accurate genetic variant discovery to tens of thousands of samples*. http://biorxiv.org/lookup/doi/10.1101/201178 (2017) doi:10.1101/201178.

184. Chiang, C. *et al.* SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods* **12**, 966–968 (2015).

185. Ye, K. *et al.* Split-Read Indel and Structural Variant Calling Using PINDEL. in *Copy Number Variants* (ed. Bickhart, D. M.) vol. 1833 95–105 (Springer New York, 2018).

186. Wala, J. A. *et al.* SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28**, 581–591 (2018).

187. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21**, 974–984 (2011).

188. Handsaker, R. E. *et al.* Large multiallelic copy number variations in humans. *Nat Genet* **47**, 296–303 (2015).

189. Kavak, P. *et al.* Discovery and genotyping of novel sequence insertions in many sequenced individuals. *Bioinformatics* **33**, i161–i169 (2017).

190. Obón-Santacana, M. *et al.* GCAT|Genomes for life: a prospective cohort study of the genomes of Catalonia. *BMJ Open* **8**, e018324 (2018).

191. Jun, G. *et al.* Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. *The American Journal of Human Genetics* **91**, 839–848 (2012).

192. The FUSION Study *et al.* Ancestry estimation and control of population stratification for sequence-based association studies. *Nat Genet* **46**, 409–415 (2014).

193. Martínez-Fundichely, A. *et al.* InvFEST, a database integrating information of polymorphic inversions in the human genome. *Nucl. Acids Res.* **42**, D1027–D1032 (2014).

194. Haraksingh, R. R., Abyzov, A. & Urban, A. E. Comprehensive performance comparison of high-resolution array platforms for genome-wide Copy Number Variation (CNV) analysis in humans. *BMC Genomics* **18**, 321 (2017).

195. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat Commun* **10**, 5436 (2019).

196. Patterson, M. *et al.* W HATS H AP : Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *Journal of Computational Biology* **22**, 498–509 (2015).

197. Galván-Femenía, I. *et al.* Multitrait genome association analysis identifies new susceptibility genes for human anthropometric variation in the GCAT cohort. *J Med Genet* **55**, 765–778 (2018).

198. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet* **5**, e1000529 (2009).

199. Roslin, N. M., Weili, L., Paterson, A. D. & Strug, L. J. *Quality control analysis of the 1000 Genomes Project Omni2.5 genotypes*. http://biorxiv.org/lookup/doi/10.1101/078600 (2016) doi:10.1101/078600.

200. Chi, M., Lee, C. & Wu, S. The prevalence of chronic conditions and medical expenditures of the elderly by chronic condition indicator (CCI). *Archives of Gerontology and Geriatrics* **52**, 284–289 (2011).

201. Friedman, B., Jiang, H. J., Elixhauser, A. & Segal, A. Hospital Inpatient Costs for Adults with Multiple Chronic Conditions. *Med Care Res Rev* **63**, 327–346 (2006).

202. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10**, 5–6 (2013).

203. Band, G. & Marchini, J. *BGEN: a binary file format for imputed genotype and haplotype data*. http://biorxiv.org/lookup/doi/10.1101/308296 (2018) doi:10.1101/308296.

204. Mägi, R. & Morris, A. P. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* **11**, 288 (2010).

205. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* **47**, D1005–D1012 (2019).

206. Pan-UKB team. https://pan.ukbb.broadinstitute.org. (2020).

207. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).

208. Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Research* **50**, D988–D995 (2022).

209. Fadista, J., Manning, A. K., Florez, J. C. & Groop, L. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *Eur J Hum Genet* **24**, 1202–1205 (2016).

210. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat Genet* **50**, 390–400 (2018).

211. Hoffmann, T. J. *et al.* Genome-wide association study of prostate-specific antigen levels identifies novel loci independent of prostate cancer. *Nat Commun* **8**, 14248 (2017).

212. Rosendahl, J. *et al.* Genome-wide association study identifies inversion in the *CTRB1-CTRB2* locus to modify risk for alcoholic and non-alcoholic chronic pancreatitis. *Gut* **67**, 1855–1863 (2018).

213. Krizman, J., Rotondo, E. K., Nicol, T., Kraus, N. & Bieszczad, K. M. Sex differences in auditory processing vary across estrous cycle. *Sci Rep* **11**, 22898 (2021).

214. Williams, O. O. F., Coppolino, M., George, S. R. & Perreault, M. L. Sex Differences in Dopamine Receptors and Relevance to Neuropsychiatric Disorders. *Brain Sciences* **11**, 1199 (2021).

215. Sacan, A., Ekins, S. & Kortagere, S. Applications and Limitations of In Silico Models in Drug Discovery. in *Bioinformatics and Drug Discovery* (ed. Larson, R. S.) vol. 910 87–124 (Humana Press, 2012).

216. Koboldt, D. C. Best practices for variant calling in clinical sequencing. *Genome Med* **12**, 91 (2020).

217. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).

218. Uffelmann, E. *et al.* Genome-wide association studies. *Nat Rev Methods Primers* **1**, 59 (2021).

219. Gangiredla, J. *et al.* GalaxyTrakr: a distributed analysis tool for public health whole genome sequence data accessible to non-bioinformaticians. *BMC Genomics* **22**, 114 (2021).

220. Roshyara, N. R., Horn, K., Kirsten, H., Ahnert, P. & Scholz, M. Comparing performance of modern genotype imputation methods in different ethnicities. *Sci Rep* **6**, 34386 (2016).

221. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* **21**, 30 (2020).

222. Schumacher Dimech, A., Ferretti, M. T., Sandset, E. C. & Santuccione Chadha, A. The role of sex and gender differences in precision medicine: the work of the Women's Brain Project. *European Heart Journal* **42**, 3215–3217 (2021).

223. Wang, J., Yu, R. & Shete, S. X-Chromosome Genetic Association Test Accounting for X-Inactivation, Skewed X-Inactivation, and Escape from X-Inactivation: X-Chromosome Genetic Association Test. *Genet. Epidemiol.* **38**, 483–493 (2014).

224. Clayton, D. G. Sex chromosomes and genetic association studies. *Genome Med* **1**, 110 (2009).

225. Zhu, Z. *et al.* Discovery of a novel genetic susceptibility locus on X chromosome for systemic lupus erythematosus. *Arthritis Res Ther* **17**, 349 (2015).

226. Wang, J., Talluri, R. & Shete, S. Selection of X-chromosome Inactivation Model. *Cancer Inform* **16**, 117693511774727 (2017).

227. Struck, T. J., Mannakee, B. K. & Gutenkunst, R. N. The impact of genome-wide association studies on biomedical research publications. *Hum Genomics* **12**, 38 (2018).

228. Keur, N., Ricaño-Ponce, I., Kumar, V. & Matzaraki, V. A systematic review of analytical methods used in genetic association analysis of the X-chromosome. *Brief Bioinform* **23**, bbac287 (2022).

229. Marigorta, U. M., Rodríguez, J. A., Gibson, G. & Navarro, A. Replicability and Prediction: Lessons and Challenges from GWAS. *Trends in Genetics* **34**, 504–517 (2018).

230. Manolio, T. A. Bringing genome-wide association findings into clinical use. *Nat Rev Genet* **14**, 549–558 (2013).

231. Sharkey, T. *et al.* Effectiveness of gender-targeted versus gender-neutral interventions aimed at improving dietary intake, physical activity and/or overweight/obesity in young adults (aged 17–35 years): a systematic review and meta-analysis. *Nutr J* **19**, 78 (2020).

232. Agbaedeng, T. A. *et al.* Polygenic risk score and coronary artery disease: A meta-analysis of 979,286 participant data. *Atherosclerosis* **333**, 48–55 (2021).

233. Zeinomar, N. & Chung, W. K. Cases in Precision Medicine: The Role of Polygenic Risk Scores in Breast Cancer Risk Assessment. *Ann Intern Med* **174**, 408–412 (2021).

234. Murray, G. K. *et al.* Could Polygenic Risk Scores Be Useful in Psychiatry?: A Review. *JAMA Psychiatry* **78**, 210 (2021).

235. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet* **9**, e1003348 (2013).

Appendix

Appendix 1. GCAT|Panel, a comprehensive structural variant haplotype map of the Iberian population from high coverage whole-genome sequencing.

Jordi Valls-Margarit*, Iván Galván-Femenía*, Daniel Matías-Sánchez*, Natalia Blay, Montserrat Puiggròs, Anna Carreras, Cecilia Salvoro, Beatriz Cortés, Ramon Amela, Xavier Farre, Jon Lerga-Jaso, Marta Puig, Jose Francisco Sánchez-Herrero, Victor Moreno, Manuel Perucho, Lauro Sumoy, Lluís Armengol, Olivier Delaneau, Mario Cáceres, Rafael de Cid, David Torrents

Contribution:

- Variant Calling Benchmarking.
- Development of the Logistic Regression Model to select a high quality set of variants.
- Validation of variants and comparison against public databases.
- Elaboration and revision of the supplementary methodology and part of the manuscript.

2464–2479 Nucleic Acids Research, 2022, Vol. 50, No. 5 *Published online 18 February 2022*
https://doi.org/10.1093/nar/gkac076

GCAT|Panel, a comprehensive structural variant haplotype map of the Iberian population from high-coverage whole-genome sequencing

Jordi Valls-Margarit[1,†], Iván Galván-Femenía[2,†], Daniel Matías-Sánchez[1,†], Natalia Blay[2], Montserrat Puiggròs[1], Anna Carreras[2], Cecilia Salvoro[1], Beatriz Cortés[2], Ramon Amela[1], Xavier Farre[2], Jon Lerga-Jaso[3], Marta Puig[3], Jose Francisco Sánchez-Herrero[4], Victor Moreno[5,6,7,8], Manuel Perucho[9,10], Lauro Sumoy[4], Lluís Armengol[11], Olivier Delaneau[12,13], Mario Cáceres[3,14], Rafael de Cid[2,*,‡] and David Torrents[1,14,*]

[1]Life Sciences Department, Barcelona Supercomputing Center (BSC), Barcelona 08034, Spain, [2]Genomes for Life-GCAT lab Group, Institute for Health Science Research Germans Trias i Pujol (IGTP), Badalona 08916, Spain, [3]Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra, Barcelona 08193, Spain, [4]High Content Genomics and Bioinformatics Unit, Institute for Health Science Research Germans Trias i Pujol (IGTP), 08916 Badalona, Spain, [5]Catalan Institute of Oncology, Hospitalet del Llobregat, 08908, Spain, [6]Bellvitge Biomedical Research Institute (IDIBELL), Hospitalet del Llobregat, 08908, Spain, [7]CIBER Epidemiología y Salud Pública (CIBERESP), Madrid 28029, Spain, [8]Universitat de Barcelona (UB), Barcelona 08007, Spain, [9]Sanford Burnham Prebys Medical Discovery Institute (SBP), La Jolla, CA 92037, USA, [10]Cancer Genetics and Epigenetics, Program of Predictive and Personalized Medicine of Cancer (PMPPC), Health Science Research Institute Germans Trias i Pujol (IGTP), Badalona 08916, Spain, [11]Quantitative Genomic Medicine Laboratories (qGenomics), Esplugues del Llobregat, 08950, Spain, [12]Department of Computational Biology, University of Lausanne, Génopode, 1015 Lausanne, Switzerland, [13]Swiss Institute of Bioinformatics (SIB), University of Lausanne, Quartier Sorge – Batiment Amphipole, 1015 Lausanne, Switzerland and [14]ICREA, Barcelona 08010, Spain

Received July 20, 2021; Revised December 24, 2021; Editorial Decision January 18, 2022; Accepted February 09, 2022

ABSTRACT

The combined analysis of haplotype panels with phenotype clinical cohorts is a common approach to explore the genetic architecture of human diseases. However, genetic studies are mainly based on single nucleotide variants (SNVs) and small insertions and deletions (indels). Here, we contribute to fill this gap by generating a dense haplotype map focused on the identification, characterization, and phasing of structural variants (SVs). By integrating multiple variant identification methods and Logistic Regression Models (LRMs), we present a catalogue of 35 431 441 variants, including 89 178 SVs (≥50 bp), 30 325 064 SNVs and 5 017 199 indels, across 785 Illumina high coverage (30x) whole-genomes from the Iberian GCAT Cohort, containing a median of 3.52M SNVs, 606 336 indels and 6393 SVs per individual. The haplotype panel is able to impute up to 14 360 728 SNVs/indels and 23 179 SVs, showing a 2.7-fold increase for SVs compared with available genetic variation panels. The value of this panel for SVs analysis is shown through an imputed rare Alu element located in a new locus associated with Mononeuritis of lower limb, a rare neuromuscular disease. This study represents the first deep characterization of genetic variation within the Iberian population and the first operational haplotype panel to systematically include the SVs into genome-wide genetic studies.

[*]To whom correspondence should be addressed. Tel: +34 934134074; Email: david.torrents@bsc.es
Correspondence may also be addressed to Rafael de Cid. Tel: +34 930330542; Email: rdecid@igtp.cat

[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

[‡]Lead contact for data access.

Present address: Iván Galván-Femenía, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08028, Barcelona, Spain.

© The Author(s) 2022. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/nar/article/50/5/2464/6527675 by eFADA user on 20 May 2023

INTRODUCTION

One of the central aims of biology and biomedicine has been the characterization of genetic variation across humans to answer evolutionary questions and to explain phenotypic variability in relation to disease. From the first genotyping and sequencing efforts, scientists have been gradually identifying specific genomic regions that vary within and across different populations, elaborating the first maps of human genetic variation (e.g. the HapMap Phase I (1)). Next-generation sequencing (NGS) technologies are now allowing to systematically evaluate the genetic variability across the entire genome of hundreds and thousands of individuals. This has increased >200-fold the number of known genomic variants over the past 10 years, resulting in much richer reference catalogues of genetic variability. One example is HRC (2) or Trans-Omics for Precision Medicine (TOPMed) (3), listing more than 39.2M and 410M polymorphic positions, respectively, from several human populations. The extensive genetic and phenotypic characterization of cohorts using rich variability reference panels is now fuelling up Genome-Wide Association Studies (GWAS). A total of 151 703 unique genetic variants are already reported to be associated across 5193 unique traits (GWAS catalog, version1.0.2 release 05/05/2021, https://www.ebi.ac.uk/gwas/). Despite these advances, a large fraction of the genetic variability underlying complex diseases still remains unexplored, as studies have been mostly restricted to single nucleotide variants (SNVs) and small insertions and deletions (indels) (<50 bp). Large structural variants (SVs) are known to play an important role in disease (4–7) and could actually explain part of the well-known missing heritability paradox (8,9). However, the technical and methodological challenges associated with the identification and classification of this type of variation from whole-genome sequences (WGS) have left this type of variation out of GWASs.

Large-scale efforts combining improved sequencing methodologies are now identifying a much larger and richer spectrum of structural variation in humans. For example, by increasing the sequencing coverage and sample size across different populations, the gnomAD-SV project (10) detected a median of 7439 SVs per individual, generating one of the most extensive catalogues of structural variation so far. Other whole-genome studies have gone a step further by phasing the variants and constructing haplotype panels, such as the 1000 Genomes project (1000G) (11), becoming a reference within the GWAS community. However, the SVs are less represented in the current 1000G phase3, including a median of 3441 SVs per individual (11). The use of costly family trios and an increase in the sequencing coverage, allowed the Genome of the Netherlands consortium (GoNL) to increase a median of 7006 SVs per individual (12). In parallel, the recent inclusion of long-read sequencing technologies has made it possible to uncover many new SVs, reaching >20 000 per individual (13–16), including repeat-rich regions, where short-read sequencing has traditionally shown low call rates.

Genome-wide imputation from SNP-genotyping array data is still the most practical and powerful strategy to predict SVs, and test them for association with particular phenotypes. Current haplotype reference panels allow a high-quality imputation (info score ≥ 0.7) of ∼9000–14 000 SVs (≥50 bp), but considering the ranges of SVs that the community is now reporting across individuals, this is still incomplete. Therefore, it is necessary to generate improved variability reference panels of controlled populations by including SVs in the discovery and functional interpretation of associated variants to power-up current genetic studies.

In this study, we contribute to fill this gap by generating a new SV-enriched haplotype reference panel of human variation, through the analysis of whole-genome sequences (30×) of Iberian individuals from the GCAT|Genomes for Life Cohort (www.genomesforlife.com) (17,18). For this, we developed and applied a comprehensive genomic analysis pipeline based on the weighted integration and orthogonal validation of the results of multiple variant callers to generate a robust catalogue of genetic variability that covers from SNVs to large SVs. These variants were further phased and converted into haplotypes that can be incorporated into GWAS. This study represents an important step towards the completion of the annotation and characterization of the human genome and provides a unique resource for the incorporation of SVs into genetic studies.

MATERIALS AND METHODS

Benchmarking samples

To benchmark our variant calling strategy, an *in-silico* sample genome was generated, by inserting a controlled set of 5 334 669 variants into the hs37d5 reference genome (excluding telomeres and centromeres). These variants cover from single nucleotide variants (SNVs) to large structural variations (SVs). The majority of them correspond to variants identified in real samples of the 1000G (11) and the ICGC-PanCancer (19) projects. In addition, to have a wider and more complex range of benchmarking variants, we designed and inserted randomly an additional set of 3925 Structural Variants (SVs) (Supplementary Table S2), reinforcing the support for insertions and translocations, among others (Supplementary Figure S1). We then used the *in-silico* sequencing ART software (ART-Illumina version 2.5.8) (20) to obtain simulated FASTQ files (Supplementary Table S1) that were further aligned to the hs37d5 reference genome using BWA (21) (version 0.7.15-r1140) and Samtools (22) (version 1.5). Best Practices of GATK (23) were followed for marking duplicates (PICARD version 1.108) and recalibrating Base Quality Scores of the BAM file with the VariantRecalibrator and ApplyVQSR modules of GATK4 (version 4.0.11). A detailed description is available at Supplementary Information Material.

The sample NA12878 from the genome in a Bottle (GIAB) Consortium (24) and the *in-silico* were used to validate SNVs and indels detection. BAM files were reconstructed using the hs37d5 reference genome and following the GATK Best Practices guidelines.

Variant calling

We originally selected 17 candidate programs for variant identification and classification, representing different calling algorithms and strategies: Split Read, Discordant Read,

Downloaded from https://academic.oup.com/nar/article/50/5/2464/6527675 by eFADA user on 20 May 2023

de novo Assembly and Read-depth. Variant callers were Haplotype Caller (25) (version 4.0.2.0), Deepvariant (26) (version 0.6.1), Strelka2 (27) (version 2.9.2), Platypus (28) (version 0.8.1), and VarScan2 (29) (version 2.4.3) for SNVs and indels and Delly2 (30) (version 0.7.7), Manta (31) (version 1.2), Pindel (32) (version 0.2.5b9), Lumpy (33) (version 0.2.13), Whamg (34) (version v1.7.0-311-g4e8c), SvABA (35) (version 7.0.2), CNVnator (36) (version v0.3.3), PopIns (37) (version damp v1-151-g4010f61), Genome Strip (38) (Version 2.0), Pamir (39) (version 1.2.2), AsmVar (40) (version 2.0) and MELT (41) (version 2.1.4) (Supplementary information section 3) for SVs. To keep consistency on the type of variables provided by these callers that will later be used by the Logistic Regression Model (LRM), we have only considered mapping-based methods, despite mapping-free methods can also identify SV efficiently.

Recall, precision, and *F*-score metrics were calculated to evaluate the performance of each variant caller for each variant type. The NA12878 sample was used as a gold standard to calculate performance metrics for SNVs and indels, and the *in-silico* was used to benchmark SVs. For SNVs and indels, a variant was considered a true positive when the calling matched with the exact position and alternative allele shown on the benchmarking set. The criteria to classify SVs as true positives were: (i) the chromosome and the breakpoint position ± the breakpoint-error of the variant caller overlaps with the gold standard (Supplementary Table S4), (ii) the SV type label matched with the gold standard, and (iii) the variant length reported by the caller has a $\geq$80% reciprocal overlap with the variant length in the gold standard sample. In addition, for SVs, we also captured information from the callers regarding breakpoint resolution, the size effect on variant calling, and the genotyping accuracy. Platypus, Varscan2, Genome Strip, Pamir and AsmVar (Supplementary Information section 4.2) were finally discarded due to either technical incompatibilities with our computing environment or the low performance in benchmarking, leaving 12 final variant callers to be applied to the GCAT–WGS samples.

The effect of the coverage on the variant calling was done by read downsampling of a group of 10 randomly selected individuals from our cohort, reproducing 5×, 10×, 15×, 20× and 25× coverage. We applied the complete variant calling strategy to the resulting samples.

Logistic regression model

Logistic Regression Model (LRM) was used on indels and SVs to merge and filter the results from all callers, generating a final set of high-quality variants with the highest recall and precision values. This method is proposed as an improved alternative to other strategies based on the number of coincident callers, which were also included for comparison and evaluation purposes. As discriminative variables, LRM used variant and calling-related parameters, like size, reciprocal overlap and breakpoint resolution (Supplementary Table S5).

Logistic regression model for indels. LRM was trained using indels of the NA12878 sample and tested using the *in-silico* sample. The LRM input was a merged dataset of the VCF outputs from all included callers, a matrix of unique variants and variant callers together. The criteria to obtain this dataset is described in the 'Variant calling, filtering and merging' section. True positive detection of the variants was assessed via logistic regression as follows: $Y \sim X_{c1} + X_{c2} + \ldots + X_{cn}$, where Y is the presence (true positive) or absence (false positive) of the variant in the training set, and X_{c1}, X_{c2}, .., X_{cn} are the genotypes reported by each variant caller respectively. Predictions derived from the LRM were converted into a binary variable, indicating if the variant was considered a true (PASS, if predicted probability $\geq$ 0.5) or a false positive (NO PASS). The genotype considered in the LRM is a consensus genotype reported by Haplotype caller, Deepvariant, and Strelka2 (Supplementary information section 5.1). The LRM was developed using R software (version 3.3.1) and the ISLR package.

Logistic regression model for SVs. For SVs, we randomly splitted the *in-silico* sample into training, with 70% of the variants, and the test set, with the rest. True positive detection of the variants was assessed via logistic regression using 10-fold cross-validations as follows: $Y \sim X_{c1} + X_{c2} + \ldots + X_{cn} + G_1 + G_2 + G_3 + G_4$, where Y is the presence (true positive) or absence (false positive) of the variant in the training set X_{c1}, X_{c2}, .., X_{cn} are the genotypes reported by each variant caller; and *G1, G2, G3* and *G4* are the genomic covariates such as size, number of callers, number of strategies and reciprocal overlap (Supplementary Table S5). Similar to indels, the input of the LRM for SVs is a merged dataset of the VCF outputs from the callers ('Variant calling, filtering and merging' section). Prediction is a binary variable depending on the predicted probability (PASS, if predicted probability $\geq$ 0.5; NO PASS otherwise). Using stepwise backward criteria for determining which genomic covariates contribute to the true positive detection of the variants, we fitted an LRM for each SV type using the caret (version 6.0–85) and e1071 (version 1.7–3) R packages. Finally, to determine the performance of the model, the receiver operating characteristic (ROC) curves and area under the curve (AUC) of the LRM were computed for the test sets of each SV type using the 'ROCR' R package. The largest AUC values correlate with the highest *F*-scores suggesting that the LRM predictions are close to the 0 (false positive) and 1 (true positive) values.

The strategy to determine the position of a variant in the LRM was different for each SV type. First, variant callers were ranked according to the accuracy in resolving the breakpoint (with an interval of error of ±10 bp; Supplementary Table S6) and the number of variants detected. This was used to select unique variants according to the position of the caller for that particular variant. In the case that a variant was not detected by the best-ranked algorithms (Supplementary Table S6), the final position of the variant was considered as the median position and the length reported by the rest of the callers.

The strategy to determine the genotype of a variant in the LRM was adapted to each SV type (Supplementary Figure S3). For Deletions and Insertions, we selected the final genotype based on the highest recurrence across callers that identified a particular variant. For Inversions, we directly reported the genotype obtained from the caller with

Downloaded from https://academic.oup.com/nar/article/50/5/2464/6527675 by eFADA user on 20 May 2023

the smallest genotyping error in the benchmarking analysis. For Duplications and Translocations, which show the lowest genotyping accuracy in the benchmarking, we applied a customised genotyping method strategy. This is based on the proportion of altered reads from the *in-silico* sample around the breakpoint: if the proportion of altered reads was <0.20, the genotype was 0/0; if the proportion was between 0.20 and 0.80, the genotype was 0/1; and if the proportion was > 0.80, the genotype was 1/1 (Supplementary information section 5.2.3).

Quality control

The GCAT Cohort is a prospective cohort study that includes 19 267 volunteers from Catalonia, in the Northeast of Spain (http://www.genomesforlife.org/). The participants were recruited from the general population (2014–2017) with the only restriction to live at least five years in Catalonia and aged between 40 and 65 years. All participants who agreed to be part of the study provided informed consent and were asked to sign a consent agreement. Whole-genome sequencing data from 808 individuals using HiSeq 4000 sequencer (Illumina, 30× coverage, read length 150 bp, insert size 600 bp) was obtained in FASTQ format (Supplementary Tables S7 and S8). BAM files were built using the hs37d5 reference genome and following the GATK Best Practices (Supplementary Figure S4). FASTQ and BAM files corresponding to these samples were deposited to the European Genome-Phenome Archive (EGA, EGAS00001003018). The GCAT cohort protocol, including sampling and processing, data generation and health status is described elsewhere (www.genomesforlife.com) (17,18).

Quality control was applied by assessing the quality alignment of the BAM files, the presence of contamination traces, possibly swapped samples, population structure and relatedness (Supplementary information section 6.3). Alignment quality was analysed using PICARD (version 2.18.11), Biobambam (42) (version 2–2.0.65), and Alfred (43) (version 0.1.16). Contamination or swapped ID samples was determined by VerifyBamID (44) (Supplementary Table S9 and Figure S6). Population structure was assessed using reference ancestry populations. Identity by descent (IBD) estimates was used to remove up to third-degree relatives.

The GCAT sample was characterized by Principal Component Analysis (PCA). Firstly, we ran the Haplotype Caller tool and only PASS variants from the VCF file were retained. Then, SNVs with minor allele frequency (MAF) >0.01 and independent variants (LD, $r^2 < 0.2$) were selected with PLINK (version 1.90b6.7 64-bit). Finally, on retained variants (∼1M) we ran PCs together with reference samples of known ancestry (i.e. 1000G project sample and the Population Reference Sample (45) (POPRES)). The genetic homogeneity of the GCAT sample was confirmed by PCA in the retained cohort samples (Figure 1 and Supplementary Figure S7).

Variant calling, filtering and merging

Each of the 12 selected variant callers was first executed independently on all samples (Supplementary information section 7, Supplementary Figure S8), then merged by call and individual according to our benchmarking strategy to produce the VCF.

SNVs and indels calls were merged by (i) the chromosome, (ii) position and (iii) REF/ALT allele. SVs were merged by (i) variant type, (ii) chromosome, (iii) position, considering the breakpoint error estimated for each variant caller (Supplementary Table S4) and finally (iv) reciprocal overlap ≥80% between callers (Supplementary information section 8.2) and individuals (Supplementary information section 8.3). Given the consistently high accuracy in detecting SNVs for most callers, we considered one of these variants as a true positive if it was detected by at least two callers. For indels and SV, we applied LRM considering a variant as true positive if the prediction probability was ≥0.5.

We calculated the true positive proportion for each variant determined by the LRM prediction in all GCAT samples. We referred to this proportion as the quality score of the merged variant. Then, we considered a variant as PASS if the quality score was ≥0.5. We reported the length and position of each SV as the median length and median position of all the samples that have that SV (Supplementary methods). Finally, monomorphic variants, variants out of Hardy-Weinberg equilibrium (Bonferroni correction *P*-value < 5×10^{-8}), and variants with ≥10% of missingness were excluded from subsequent analysis. Data and Code availability is described below. Summarized later at the resource availability section.

Variant validation

Comparison with public datasets. SNVs and indels from the GCAT dataset were compared with the NCBI dbSNP database (46) (Build version 153) (https://ftp.ncbi.nlm.nih.gov) to determine the number of unique/shared variants between them. GCAT SVs were compared with the following public databases: (i) The Genome Aggregation Database (gnomAD.v.2) (10) (https://gnomad.broadinstitute.org/downloads), (ii) the Database of Genomic Variants (DGV) (http://dgv.tcag.ca/dgv/app/downloads?ref=GRCh37/hg19) (47), (iii) the Human Genome Structural Variation Consortium set (HGSVC) (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/) (13), (iv) the Ira M Hall dataset (https://github.com/hall-lab/sv_paper_042020) (48), (v) the 1000G project (Phase3) (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/) (11) and (vi) GoNL (release 6.2) (on request) (12). Finally, we determine the number of shared variants between the GCAT and at least one other public dataset and the number of unique variants in the GCAT derived (Supplementary Information section 9.1.2).

We also carried out a comparison with the emerging long-read sequencing technologies. We analysed with our pipeline 30× short-read sequencing information from a 1000G sample (id: NA12878) that had been also independently sequenced and analysed using long-read technologies. We ran our variant calling and filtering strategies in this sample and matched the results obtained with those reported in Audano's study (15) (long-read sequencing) and

Downloaded from https://academic.oup.com/nar/article/50/5/2464/6527675 by eFADA user on 20 May 2023

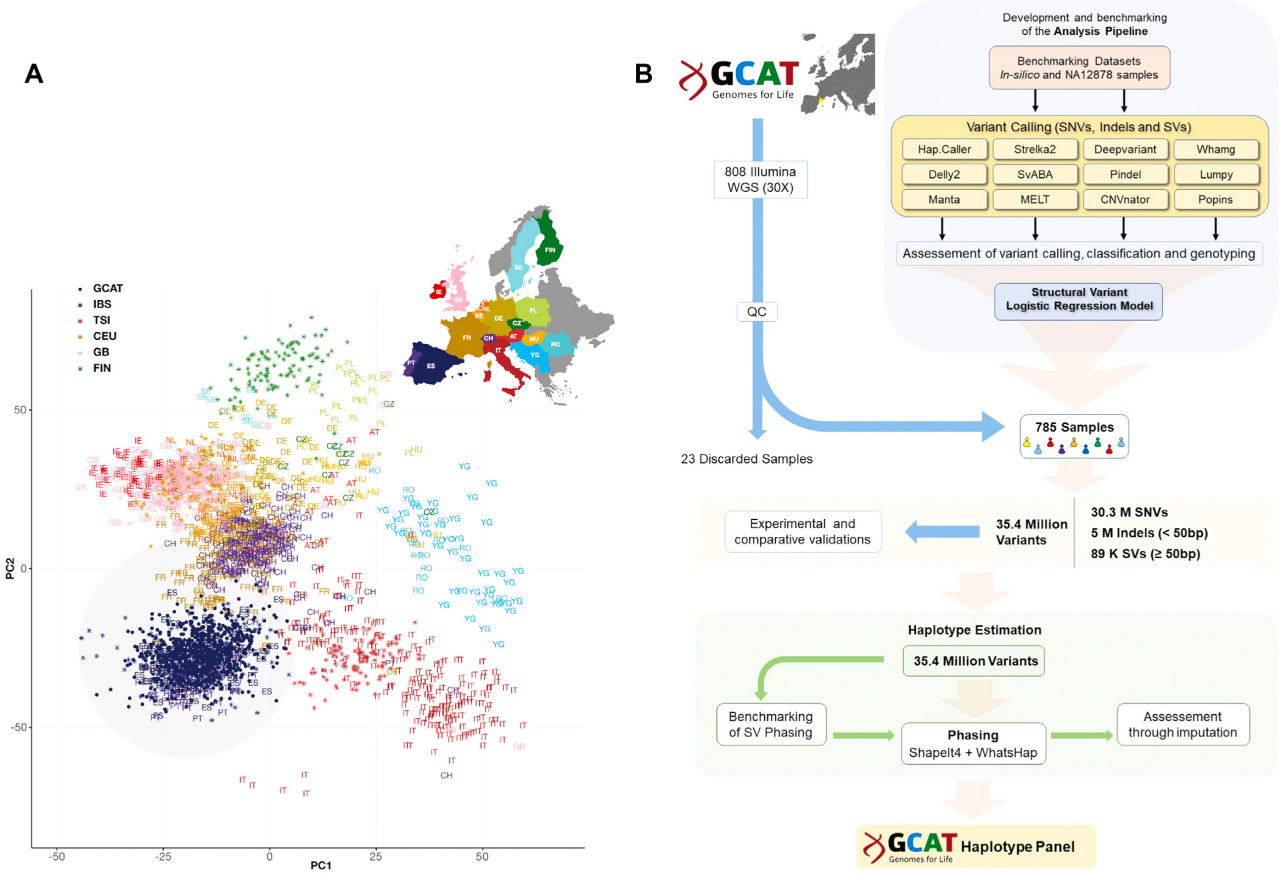


Figure 1. Overview of data and overall strategy. (**A**) Distribution of genetic data (SNVs) based on principal component analysis (PCA) (adapted from Novembre *et al.* (45)). The PC grouped by geographic localization (coloured in grey) the individuals of the GCAT cohort (blue dots) with Iberian samples from 1000G (asterisk) and POPRES (letters) projects in the context of other European samples. (**B**) Flowchart of the overall strategy followed in this study, covering from the quality control of the initial data, to the final generation of the GCAT haplotype panel, with particular focus on SVs. Overall, the complete strategy consumed ~3.5 million CPU/hour, which highlights part of the computational challenges associated with this type of analysis (Supplementary Table S11) (See also Supplementary Figure S7).

1000G Phase 3 (3–7× coverage), obtaining the number of variants shared between projects.

Experimental validation. The validation of SNV and indel calling was performed using the SNP-array data available from 570 of the 785 individuals analysed in this study. We include QCed genotypes generated in the GCAT cohort with the Infinium Expanded Multi-Ethnic Genotyping Array (MEGAEx) (ILLUMINA, San Diego, CA, USA) as described elsewhere (18) (i.e. 732 978 SNPs and 1168 indels). Genotypes from both strategies were compared by (i) chromosome and position at base-pair resolution and (ii) REF/ALT alleles; the recall and genotype concordance for each individual sample was calculated.

Inversions were validated using a recent benchmark dataset, consisting of 59 validated human polymorphic inversions from the InvFEST project (49). Allele frequency (using CEU and TSI European populations) and length concordance was determined using an overlapping window of ±1 kb around the inversion breakpoints. Accuracy of inversion genotyping was assessed for the 785 WGS samples, using the available reference panel of experimentally-resolved genotypes (49). GCAT genotypes were imputed with IMPUTE2 (50) with a genotype posterior probability ≥0.8 and classified as missing otherwise. Missing genotypes were recovered if they had a perfect tag SNP in the reference panel ($r^2 = 1$).

Comparative genomic hybridization (CGH) method was used to validate deletions and duplications using the NA12878 sample from 1000 Genomes project as reference, for which the lists of variants had been previously described (51). For each sample, we determined gains and losses and compared them with those reported from our variant calling analysis.

Phasing and imputation performance

In order to analyse the performance of the phasing and imputation processes, all 785 GCAT samples were divided into two subsets, (i) a subset including 690 samples were first used to construct a pilot reference panel and (ii) the remaining 95 samples, with WGS and SNP-genotyping array data available, were then used as a test sample in the different analyses.

Downloaded from https://academic.oup.com/nar/article/50/5/2464/6527675 by eFADA user on 20 May 2023

The evaluation of phasing strategies was carried out by determining the imputation accuracy of SVs, using the genotypes independently generated by WGS and imputation techniques across the 95 test GCAT samples, and with the pilot reference panel of 690 individuals (Supplementary Information section 10.1). Accuracy was determined for chromosome 22, and the quality score of imputed variants was considered as a validation proxy of the best phasing strategy. Each phasing strategy was evaluated by counting the number of variants with an info score $\geq$0.7, and by calculating the genotype concordance between imputed data and the calling. The phasing algorithms evaluated were ShapeIt2 (52) (version v2.r904), MVNcall (53) (version 1.0), ShapeIt4 (54) (version 4.1.3) and WhatsHap (55) (version 0.18). We used IMPUTE2 (50) (version 2.3.2) for imputation analysis (Supplementary methods).

In order to evaluate the imputation performance of the GCAT|Panel for distant ethnicities, we used the 1000G SNP-genotyping array data covering 2318 samples from 19 populations (56) (Supplementary Table S13). First, quality control was applied to the 1000G SNP-genotyping array per population by removing variants the met the following criteria: (i) $\geq$10% of missingness; (ii) matching A–T, C–G sites; (iii) in Hardy–Weinberg disequilibrium (P-value < 0.05); and by discarding samples with (i) $\geq$10% of missing, (ii) Kinship coefficient $\geq$0.05 and (iii) an excess of heterozygosity ±2SD, obtaining finally 1880 individuals covering 19 populations. Each population group was pre-phased with ShapeIt4 and imputed separately using IMPUTE2. Then, we compared the allele frequency, type of variant distribution, and the quality of the imputed SVs across populations.

To evaluate the imputation of SVs, we used as reference the Audano *et al.* (15) study that includes SVs identified using long-read sequencing. Imputed SVs with an info score $\geq$0.7 were compared considering a window of ±50 bp around the breakpoint. Furthermore, we evaluated the concordance of SV type and SV length error reported by WGS calling. On the other hand, we also evaluated the concordance of the genotype of our imputed SVs, using the SV list generated on the same samples, by Hickey *et al.* (57).

Benchmarking different panels of genetic variability

QCed genotypes generated in the GCAT cohort with the Infinium Expanded Multi-Ethnic Genotyping Array (MEGAEx) (i.e. 756 773 SNVs) were used to impute 4448 individuals (e.g. excluding those 785 with WGS) using the GCAT|Panel and the publicly available 1000G phase3 (11), GoNL-SV (12), UK10K (58) and HRC (2) reference panels. Multiple reference panel imputation was conducted using GUIDANCE (59). For comparative purposes, we considered imputed variants with info imputation score $\geq$0.7 and MAF >0.001. For SNVs and indels, variants were considered coincident when the position and change matched. For SVs, matching variants were considered if their positions were within a ±1 kb window, and the variant type was the same. Since allele frequency impacts imputation, we calculated the average of the info imputation score (r^2) by frequency categories: rare (MAF < 0.01), low frequency ($0.01 \leq$ MAF < 0.05), and common (MAF ≥ 0.05).

Functional impact of structural variants

Variant annotation. Functional, regulatory, and clinical annotations of SVs were predicted using AnnotSV (60). The functional impact of SVs was evaluated by considering (i) the level of overlap with known genes, (ii) the level of overlap with regulatory regions (61), (iii) the predicted loss of function intolerance (pLI) effect and (iv) the reported disease association studies. In addition, we used SVFX (62), a mechanism-agnostic machine learning-based workflow, to evaluate the potential pathogenicity of large deletions and duplications (>50 bp), in four major cardiometabolic conditions from the GCAT cohort; diabetes, obesity, cardiovascular diseases, and hypertension. SVs were classified using the annotations of the SVFX tool into pathogenic (SV pathogenic score $\geq$ 0.9) or benign (SV pathogenic score $\leq$ 0.2). Finally, SNVs and indels (up to 50 bp) were annotated using SnpEff (63) and SnpSif (64) (v5.0e) tools, covering LoF and pathogenicity descriptors from ClinVar (65) and CADD (66) resources.

Comparison with the GWAS catalog. GWAS catalog version 1.0.2 (r2021-05-05) was downloaded from https://www.ebi.ac.uk/gwas/docs/file-downloads. First, we selected 106 906 variant-phenotype associations of 72 849 unique autosomal entries identified in European ancestry. Second, we intersected with PLINK2.0 (67) 68 323 unique variant-phenotype associations (MAF $>$ 0.01) with the GCAT dataset (~30M variants) by breakpoint coordinates. Finally, we identified 1374 unique SVs (MAF > 0.01) in strong linkage disequilibrium ($r^2 > 0.80$) with variant-phenotype associations in 1Mb window (Supplementary Figure S27). From these 1374 SVs, we evaluated the SV type, as well as the overlap with genes and regulatory regions.

Genome-wide association analysis. Association analysis was performed by 70 independent GWAS of chronic conditions. Phenotype selection was derived from the Electronic Health Records registry from the cohort (2012–2017) and chronicity was defined using public guidelines for chronic condition definitions (68), and the Chronic Condition Indicator (CCI) (http://www.hcup-us.ahrq.gov/toolssoftware/chronic/chronic.jsp) (69,70), then grouped considering ICD-9 codes and chapter descriptions. Conditions with more than 50 cases were retained for the GWAS analysis (i.e. 70). Each association test was performed as independent logistic regression for each cohort, under the assumption of an additive model for allelic effects, with adjustments made for age, sex and the first five principal components. Gender-specific conditions were analysed only for a specific gender. The analysis was performed using PLINK2.0 (67) for autosomal chromosomes. A Bonferroni correction accounting for the 10 ICD-9 categories used (i.e. body systems) was applied. Locus Zoom was derived for specific regions, and suggestive tower profiles were analysed, based on LD patterns and gene-centered impact.

Experimental validation of the Alu element. PCR amplicon analysis was designed using Primer 3.0 software using the hg19_dna range = chr3:49 492 813–49 496 062 sequence, including the Alu element. Sequence primers are for F-primer (5′CATTGACTCATTCAGCAAGCA 3′) and

Downloaded from https://academic.oup.com/nar/article/50/5/2464/6527675 by eFADA user on 20 May 2023

for R-primer (5′AAATTAAGCCCCACCCTAG3′). Using standard conditions (35×, $T_m = 60^oC$) in a Veriti™ 96-Well Thermal Cycler (Thermo Fisher Scientific), we obtain a 515 bp fragment corresponding to the control-allele and an 848 bp one for the Alu-allele. Fragments were resolved by e-agarose gel, in a TapeStation (Agilent). Further, the amplicon of a non-ALU allele carrier was analysed by Sanger Sequence Method to verify the insertion point (i.e. at hg19 Chr3:49 494 280) and the ALU sequence insertion (324 bp).

Statistical analyses

R software was used for data visualization and statistical analyses. 95% confidence intervals (CI) for recall, precision, and genotype error metrics were assessed as point estimation ±1.96SD. Risk ratios with 95% CI and two-tailed *P*-values from the functional enrichment of common and rare SVs were calculated using the risk ratio function from the epitools R package. Pearson correlation coefficient with 95% CI and two-tailed *P*-value were estimated using the cor.test() function implemented in R.

RESULTS

Evaluation of cohort data quality and consistency

From the GCAT cohort (17) we randomly selected 808 individuals (gender-balanced) for new Illumina whole-genome sequencing at 30× coverage. Twenty three samples were excluded based on sequence quality, ethnicity, and relatedness parameters (see Methods, Supplementary Table S10). Principal component analysis (PCA) on the remaining 785 individuals identified a unique and separated cluster compared with neighbouring populations (Figure 1A, Supplementary Figure S7), in agreement with their geographic origin (45).

Generation of a comprehensive variant identification strategy

We designed, benchmarked, and implemented a comprehensive strategy for capturing, classifying, and phasing a wide range of germline variants from short-read Illumina sequences, with particular efforts devoted to the identification and subclassification of larger structural variants (Figure 1B). Using sequencing data from an *in-silico* genome (Supplementary information 1, Supplementary Table S2), and a real sample (NA12878, from the Genome In A Bottle (GIAB) project (24)), we assessed the performance (i.e. recall, precision and *F*-score metrics) of 17 variant callers covering SNVs, small indels (<50 bp), and large SVs (≥50 bp) (see Materials and Methods), and retained the best twelve (Supplementary Table S3). SNVs were first filtered based on a minimum constraint of having the support from at least two callers, which provided high recall (>95%) and precision (>96%) values. On the other hand, for the filtering of small indels and SVs, which show high levels of discrepancy across individual callers and their combinations (Figure 2A), we built a Logistic Regression Model (LRM), to prioritize caller results through a reliability score from the weighted combination of different calling parameters (Figure 2B, Supplementary Figure S2) (see Materials and Methods), accordingly higher *F*-scores correlated with larger AUC values (Supplementary Figure S30). This approach outperformed other typical curation strategies over the entire spectrum of SV sizes (Figure 2C, Supplementary Figure S5). Furthermore, because accurate genotype calling is also key for downstream analyses, on top of this LRM, we prioritized those callers that best resolved the heterozygosity (i.e. genotypes) (see Materials and Methods), resulting in a lower rate of genotype error (<6%) across all variant types when compared to the *in-silico* sample (Figure 2D, Supplementary Figure S3).

Genome-wide variation analysis of the GCAT cohort

The application of this strategy to the selected 785 whole-genome Illumina sequences (30×), let us identify 35 431 441 unique variants across the cohort. Of these, 85.6% correspond to SNV, 14.1% to indels (<50 bp) and 0.3% (n = 89 178) to SVs (≥50 bp) (Figure 3A). Median values of variants per individual were 3.52M SNVs (SD = 24 983), 606 336 indels (SD = 8060) and 6393 SVs (SD = 222), showing good consistency across the cohort (Figure 3B). SV sizes ranged from 50 bp to 197MB (duplication), with median values of 291 bp and a different distribution for each type of variation (Figure 3C), affecting globally a median of 7% of the entire genome per individual. Frequency ranges across all SVs were in agreement with other public WGS-based studies (Figure 3D), with 31% of them being common or low-frequency (MAF ≥ 0.01), and 69% being rare (MAF < 0.01), including a large fraction (50%) present only in one or two individuals (i.e. MAF ≤ 0.0025).

The robustness of these results was evaluated using comparative and experimental approaches. A large fraction of SNVs and indels (i.e. >79% and >93% respectively) matched with dbSNP (Build 153.v) (46) entries (Supplementary Figure S9a, b). Regarding SVs, the comparison against different public databases (i.e. gnomAD-SV (10), 1000G (11), GoNL (12), HGSVC (13), DGV (47), dbVar (47), Ira M. Hall Lab dataset (48); see Materials and Methods) highlighted 49,333 novel SVs (i.e. 61% of all SVs), of which 27% were present in more than two individuals (Supplementary Figure S9). As to the type, 26% of these novel variants correspond to deletions, 8% to duplications, 20% to insertions, 20% to inversions, 4% to LINEs, 1% to SVAs, and 21% to Alu elements. The comparison of our results with array-based genotypes in a fraction of our cohort (n = 570 individuals) validated 96% and 87% of SNVs and indels, respectively, with a genotype concordance of 97% and 96% (Supplementary Figure S10). Furthermore, we also used a benchmarking set of 59 manually-curated and experimentally-genotyped inversions with MAF >0.01 from the InvFEST project (49) to evaluate this type of variants within our catalogue. Of these 59 inversions, we detected 51 (86%), with concordant size and allele frequency values (Supplementary Figure S11a, b; see Materials and Methods). This validates ∼38 000 of ∼40 000 independent inversion calls across the entire cohort, with an average genotype concordance of 95% (Supplementary Figure S11c). In addition, we have applied CGH, which best targets duplications, as well as large deletions (>20 kb). Using this

Downloaded from https://academic.oup.com/nar/article/50/5/2464/6527675 by eFADA user on 20 May 2023

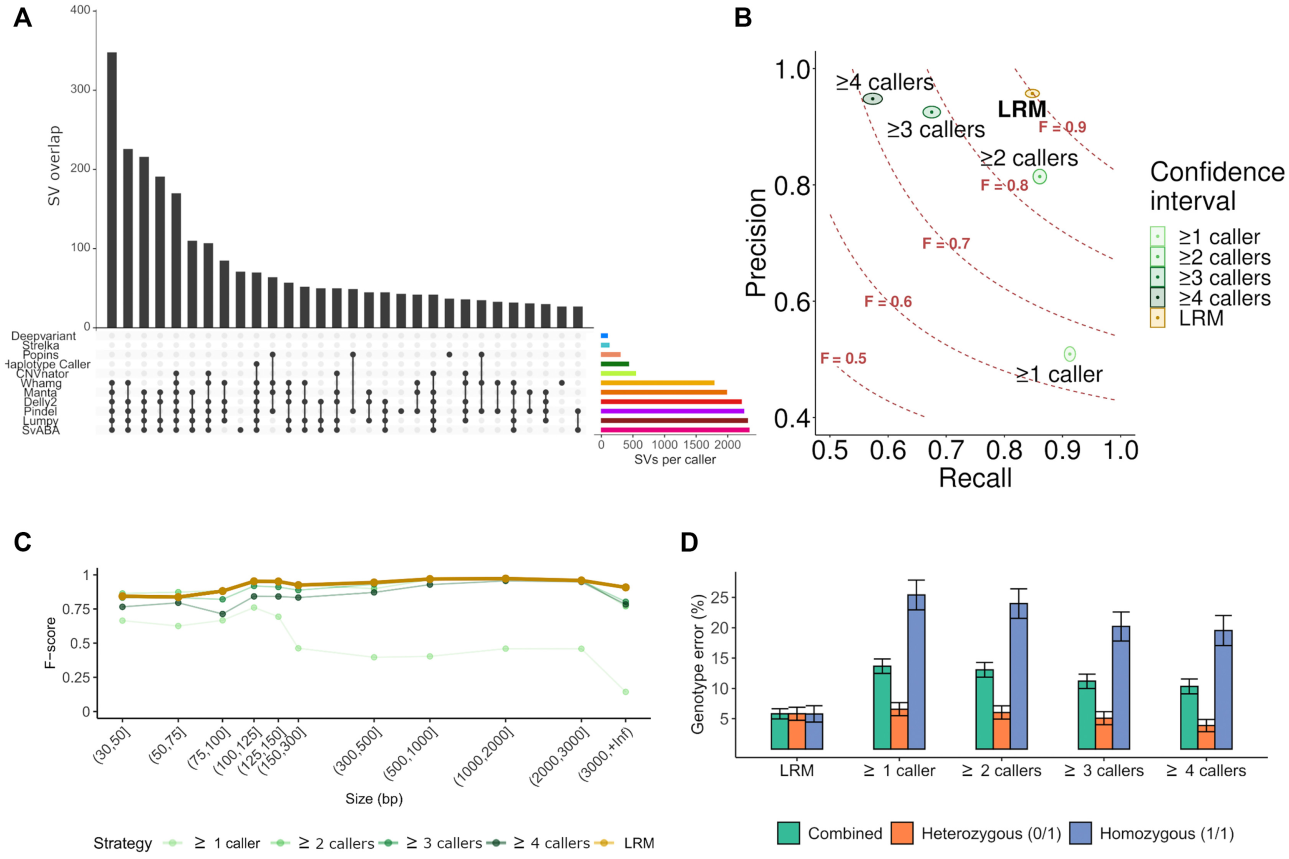


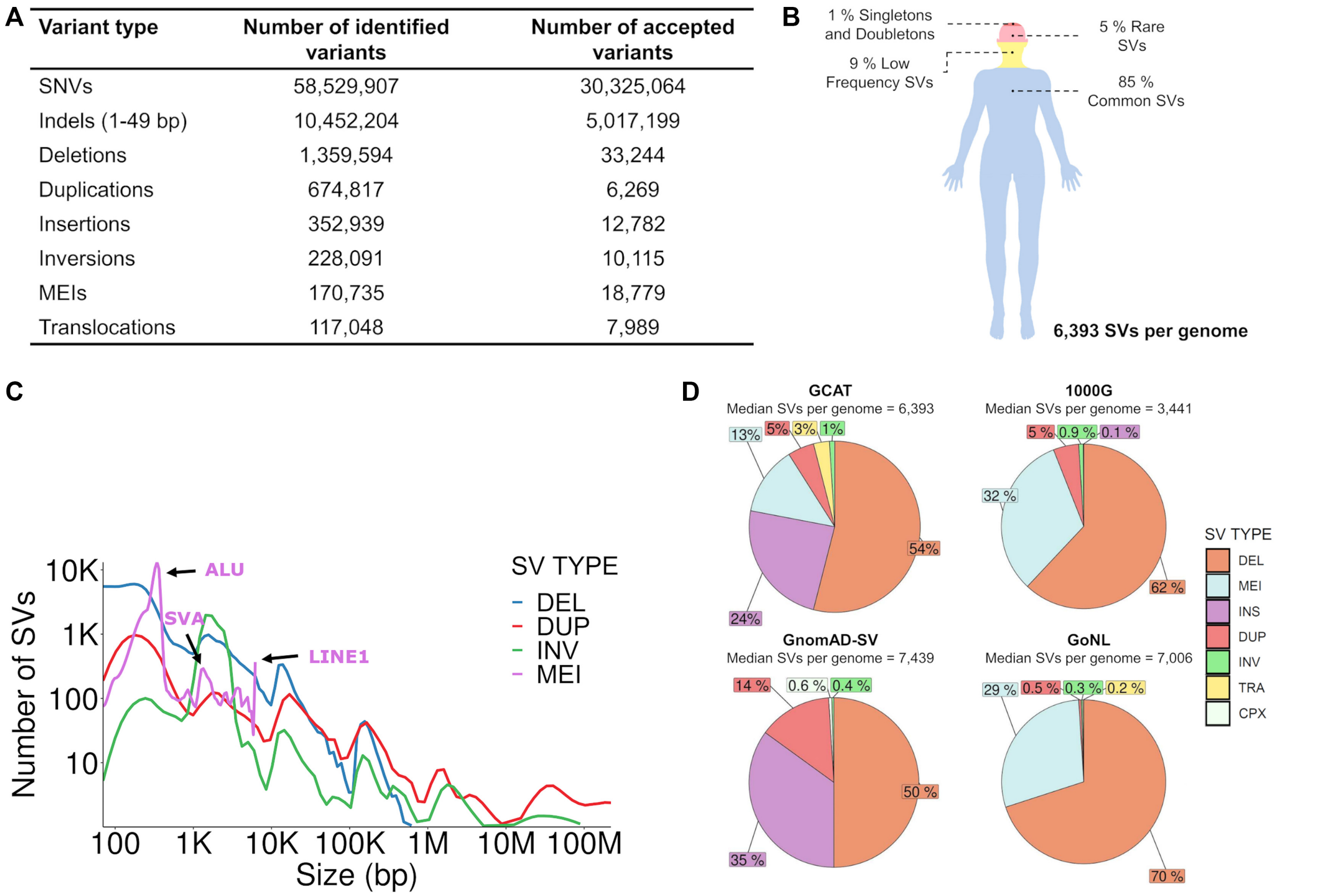
Figure 2. Benchmarking of the structural variant identification and classification pipeline. (**A**) Structural variant (SV) detection patterns according to the programs used. Lines and dots indicate the programs used and bars the number of overlapping calls resulting from that combination. The first 30 patterns with more coincident SV calling are shown. Right coloured horizontal bars indicate the total number of SVs detected by each caller. Variant callers that detect all SV types and sizes tend to recover more SVs than those that detect specific SV types (i.e. CNVnator) and smaller SVs (i.e. Strelka2). (**B**) Overview of the detection performance of different strategies and filtering results from multiple variant callers. Each strategy is plotted according to the recall and precision ratios ($F = F$-score) using the benchmarking dataset. The logistic regression model (LRM), with a F-score of 0.9, outperformed other commonly used strategies that are based on the number of coincident callers (logical rules). The confidence interval for each case is represented by coloured area of each strategy. (**C**) Comparison of performances (F-score) of different merging and filtering strategies according to the size of the structural variant. (**D**) Comparative overview of the genotype error, associated to each strategy for each allelic state. Error values and their intervals were inferred from the benchmarking dataset (see supplementary Figures S2, S3 and S5 for the information across the different SV types).

technique, we could validate 76% of our deletions, as well as 20% of the duplications (Supplementary Table S16). Finally, we contextualized our results in the frame of other SV identification efforts, through the analysis of the NA12878 sample from the 1000G project that has been sequenced and analysed using long and short read technologies at different coverages. From all SVs identified with long-read technology (15), our strategy was able to identify 24% of them when applied to NA12878 at 30× short-read sequence. This overlap is different across different SV types, as we detected 14% of the insertions and duplications, but up to 48 and 57% of the inversions and deletions, respectively. The same comparison using the 1000G annotation of NA12878 at 3–7× coverage showed a coincidence with long-read results of 4, 2 and 0.1% for deletions, inversions, and duplications respectively (Supplementary Table S17), showing a significant detection improvement when using higher coverages, identifying between a 2- and 7-fold the number of variants with 30× coverage, compared with 15× and 5× coverages, respectively (Supplementary Figure S31).

Predicted functional impact of SVs

A first assessment of the potential functional impact and pathogenicity of our SVs was obtained using AnnotSV (60). 46% of all SVs overlapped with genes, affecting a median of 2868 per individual, whereas 18% overlapped with gene regulatory regions (see Data and Code Availability at the resource availability section for the corresponding gene lists). While the majority (88%) of gene-overlapping SVs mapped within intronic regions (Supplementary Figure S24a), 9% of them affected coding sequencing regions (CDS). In agreement with known variant fixation patterns within populations, we observed that rare SVs (MAF < 0.01) tend to be more disruptive, compared to common variants (MAF ≥ 0.05), as 13% of rare SVs are overlapping coding regions, compared to 5% of the common ones (RR = 0.13/0.054 = 2.4, 95% CI = [2.14,2.69], P-value $= 2.6 \times 10^{-67}$, Supplementary Table S15a, b). Of the affected genes, 28% (10 600 SVs) are related to disease, as indicated by the predicted loss-of-function intolerance parameter (pLI) (71) (Supplementary Figures S25a, S26 and

Downloaded from https://academic.oup.com/nar/article/50/5/2464/6527675 by eFADA user on 20 May 2023



Variant type	Number of identified variants	Number of accepted variants
SNVs	58,529,907	30,325,064
Indels (1-49 bp)	10,452,204	5,017,199
Deletions	1,359,594	33,244
Duplications	674,817	6,269
Insertions	352,939	12,782
Inversions	228,091	10,115
MEIs	170,735	18,779
Translocations	117,048	7,989

Figure 3. Overview of the GCAT variant catalogue. (**A**) Table with the numbers of identified and accepted variants after applying the filters 'at least two callers detecting the same variant' for SNVs, the LRM for indels and SVs, Hardy–Weinberg equilibrium, and discard monomorphic variants and those with >10% missingness within the GCAT cohort, according to their class. (**B**) Overview of the variant distribution within an average individual in the GCAT cohort, according to their observed minor allele frequency (MAF). (**C**) Distribution of SV type according to their genomic sizes. (**D**) Comparative overview of the SV type number and distribution across the GCAT, 1000G, GnomAD and GoNL catalogues.

Table S14). Additionally, when we analysed the putative causal role of our SVs variants across multiple phenotypes, we observed that 1374 SVs (MAF ≥ 0.01) are in strong linkage disequilibrium (LD) ($r^2 \geq 0.8$) with loci associated with human traits from the GWAS Catalog (version1.0.2 release 05/05/2021), tagging mainly deletions (Supplementary Figure S27), with more than half of them (799) directly overlapping genes or regulatory regions. Finally, we further refined these results with annotations from the SVFX tool (62) for four major cardiometabolic conditions; obesity, cardiovascular traits, hypertension, and diabetes. Our analysis identified 106 GWAS catalog (P-value $< 10^{-8}$) hits (i.e. 8% of total hits) that overlap with pathogenic annotated variants in the four analysed traits; 55% variants overlap with obesity and related obesity traits, 20% with diabetes, 16% with cardiovascular-related diseases and 9% with Hypertension and related traits. Of these, 95% were common and 5% were low-frequency variants. We observed a ratio of pathogenic to benign deletions of 0.95, 1.93, 1.85 and 0.40 for diabetes, hypertension, obesity, and cardiovascular traits, respectively. In the case of duplications, these ratios were 2.06, 4.42, 4.06 and 0.82 for diabetes, hypertension, obesity, and cardiovascular traits, respectively, suggesting that duplications are twice more likely to be involved in these traits.

From the annotation obtained using SnpEff (63) we extracted 2855 variants that were classified as LoF and obtained their pathogenicity using ClinVar (65) and CADD (66) data. ClinVar data was available for 243 variants 70 of which were reported as pathogenic or likely pathogenic, and CADD data was available for 2850 variants, 2330 of which were classified as deleterious (CADD PHRED score > 20).

Iberian Haplotypes estimation

As a resource for the enrichment of SVs within genome-wide association studies, we built a haplotype reference panel by phasing together all the variants identified within all GCAT samples. We first generated a cross-validation framework to identify the best available phasing strategy for SV (see Materials and Methods), using downstream imputation results as the evaluation and ranking criteria (Supplementary Figures S12 and S13 and Table S12). In our hands, the combination of ShapeIt4 (54) and WhatsHap (55), which include phase informative reads (PIRs),

Downloaded from https://academic.oup.com/nar/article/50/5/2464/6527675 by eFADA user on 20 May 2023

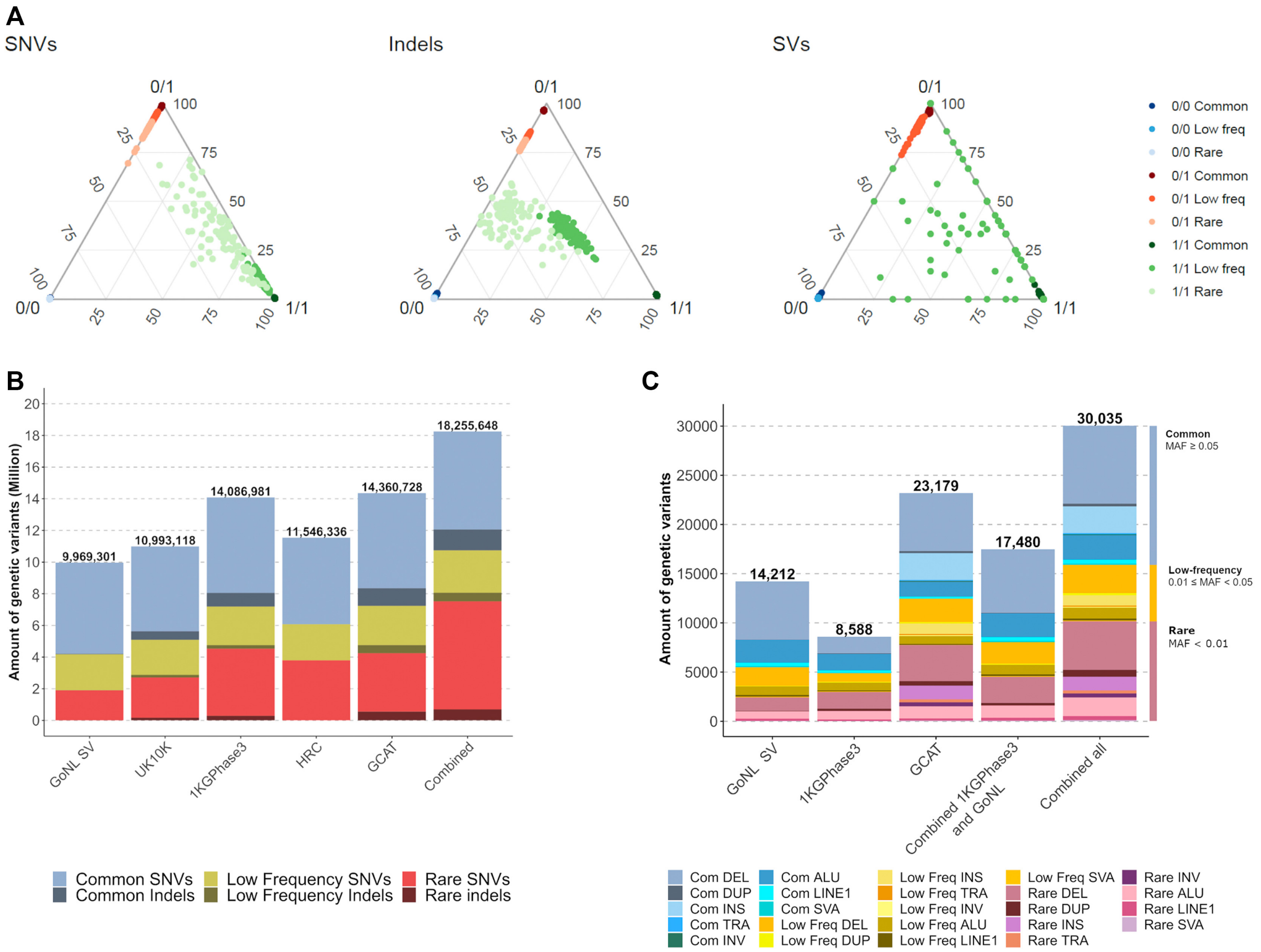


Figure 4. Phasing and Imputation performance of the GCAT|Panel. (**A**) Ternary diagram of the genotype imputation accuracy by variant type and frequency, considering the genotype calling as reference. Three dots evaluate each genotype state per sample. The samples with high concordances between genotype imputation and genotype calling were located at ternary diagram vertices. (**B**) Number of SNVs and indels imputed (info score ≥ 0.7) using different reference panels and combining their imputation results. More indels were recovered by GCAT|Panel. (**C**) Number of SVs imputed (info score ≥ 0.7) using different panels, and combining the imputation results with and without GCAT|Panel. (See also Supplementary Figure S21).

provided the best results. Using this protocol (Supplementary Figure S14), the resulting haplotype panel allowed the imputation (info scores > 0.7) of 98%, 92%, and 90% of our common SNVs, indels and SVs, respectively, recovering a median of 5120 SVs (SD = 50), from a maximum of 6393 SVs estimated per individual. While the best imputation results came from *de novo* insertions and deletions, with 96% and 95% recovery rates, respectively, duplications and translocations were imputed at lower rates, i.e. 48% and 19%, respectively (Supplementary Figure S15). Overall we imputed common SNVs, indels and SVs with a genotyping concordance of 99% (SD = 0.4), 97% (SD = 0.6) and 98% (SD = 1.2) (Figure 4A), respectively. The lowest values were observed for duplications and translocations, with genotype concordances of 84% (SD = 9.2) and 73% (SD = 27.6), respectively (Supplementary Figure S16).

As the possibilities of accurately imputing SVs are expected to correlate with the number of neighbouring SNVs and indels in LD, we next analysed the variation context of our SVs. Using one megabase window, we observed that the number of SNVs and indels in strong LD ($r^2 \geq 0.8$) with common deletions, insertions, inversions, and mobile element insertions (MEIs) was in the range of 39–42, in contrast to duplications and translocations, which showed mean values of 12 and 8 variants respectively (Supplementary Figure S17a). In fact, as expected, a positive significant correlation was observed between the number of variants in LD and the score of imputation for common SVs (Pearson's $r = 0.38$, 95% CI = [0.37, 0.40], P-value $< 2 \times 10^{-16}$) (Supplementary Figure S17b), and for all SV types (except translocations) (Supplementary Figure S18).

Imputation performance of the haplotype panel

Following this strategy, we generated a complete and operational panel of Iberian haplotypes, with all the variants of our 785 individuals. To assess the value and benefits of the resulting GCAT|Panel, as an imputation resource for enriching genetic association studies with SVs, we first imputed the genotyping array data of 4448 GCAT individuals and compared the results with those of other reference panels, such as 1000G (11), GoNL (12), HRC (2), and UK10K

Downloaded from https://academic.oup.com/nar/article/50/5/2464/6527675 by eFADA user on 20 May 2023

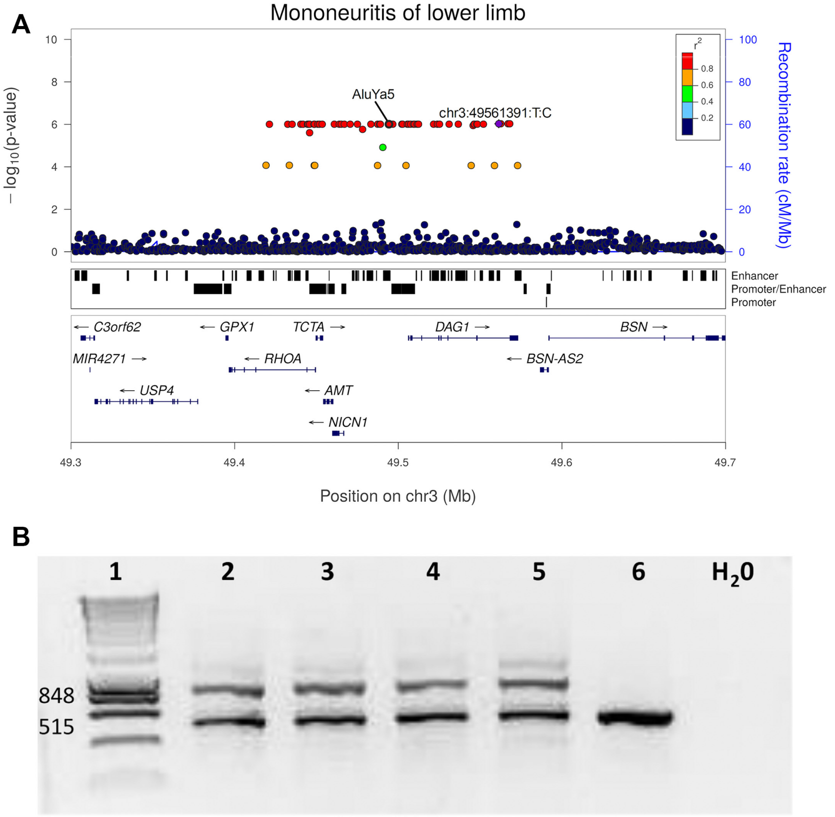


Figure 5. Genome-wide association analysis using GCAT|Panel and experimental validation of an AluYa5-element. (**A**) Locus zoom plot of the locus associated with mononeuritis of lower limb (ICD-9 355) (P-value = 9.84 × 10^{-7}), showing the lead variant in purple. The AluYa5-element (g.49494276_49494600ins (hs37d5) maps in an enhancer element upstream of the *DAG1*. (**B**) Experimental validation of an AluYa5-element, agarose e-gel electrophoresis of PCR products after amplification of Alu-insertion-specific DNA fragments from blood DNA Lanes: 1, 100 bp DNA ladder marker (Life Technologies), expected sizes of both states are shown to the left; 2–5 Alu carriers (EGA_04200, EGA_01901, EGA_13378, EGA_03940); six control individual (EGA_01399). The numbers to the left refer to the size (bp) of marker DNA fragments. Electrophoresis analysis of Alu carriers show two-band amplicons (515 bp and 848 bp) detected in Alu carriers (lanes 2–5) and one-band amplicon (515 bp) in control non-Alu-allele individuals (lane 6) (See also Supplementary Figure S29).

(58). With IMPUTE2 (50), the GCAT|Panel was able to impute a total of 14 383 907 variants with MAF > 0.001 and high quality (info score ≥ 0.7). Across different reference panels, the overall imputation performance for SNVs and indels (<50 bp) was generally high (Figure 4B), with slight overperformances of the GCAT|Panel on indels, and of 1000G and HRC panels on SNVs. While HRC and 1000G recovered rarer SNVs, likely because of their larger sample sizes, the GCAT|Panel was able to recover rarer indels (Figure 4B). At the structural variation level, the GCAT|Panel was able to impute a total of 23, 179 SVs with info scores ≥0.7, resulting in a 1.6-, 2.7- and 1.3-fold increase, compared with the 1000G, the GoNL, and both panels combined, respectively (Figure 4C). For common SNVs/Indels (MAF > 0.05) the GCAT|Panel showed similar performance as HRC, 1000G, GoNL and UK10K reference panels (mean r^2 > 0.96, Supplementary Figure S21a). For common SVs, the GCAT|Panel outperformed (mean r^2 = 0.91, SD = 0.15) 1000G (mean r^2 = 0.80, SD = 0.21) and GoNL-SV reference panels (mean r^2 = 0.82, SD = 0.21, Kruskal–Wallis P-value < 2.2 × 10^{-16}, Supplementary Figure S21b).

In an exploratory analysis, structural variants imputed by the GCAT|Panel were also tested (together with SNV and indels) for association across 70 identified chronic conditions within the cohort. Conservatively, only structural variants with an info score >0.7 and conditions with >50 cases were included in this analysis. Forty six SV loci showed suggestive association with 26 conditions after Bonferroni correction (P-value ≤ 1 × 10^{-6}) (Supplementary Figure S28). Of all these associations, 63% could potentially be functionally explained through SVs, as they either lead the association (37%) or are in strong LD (r^2 ≥ 0.8) with the lead variant (26%). A notable example is a rare AluYa5-element in chr3 (g.49494276_49494600ins (hs37d5), MAF = 0.0013), located near the dystroglycan gene (*DAG1*) and associated (P-value = 9.84 × 10^{-7}) with Mononeuritis of lower limb (ICD-9 355) (Figure 5A). The presence of this Alu element, imputed only with the GCAT|Panel (info score = 0.98), was experimentally confirmed in all carrier individuals (Figure 5B, Supplementary Figure S29).

Finally, we evaluated the portability of the GCAT|Panel to infer SVs across 19 different ethnic groups using 1880 individuals from the 1000G project. While the imputation

Downloaded from https://academic.oup.com/nar/article/50/5/2464/6527675 by eFADA user on 20 May 2023

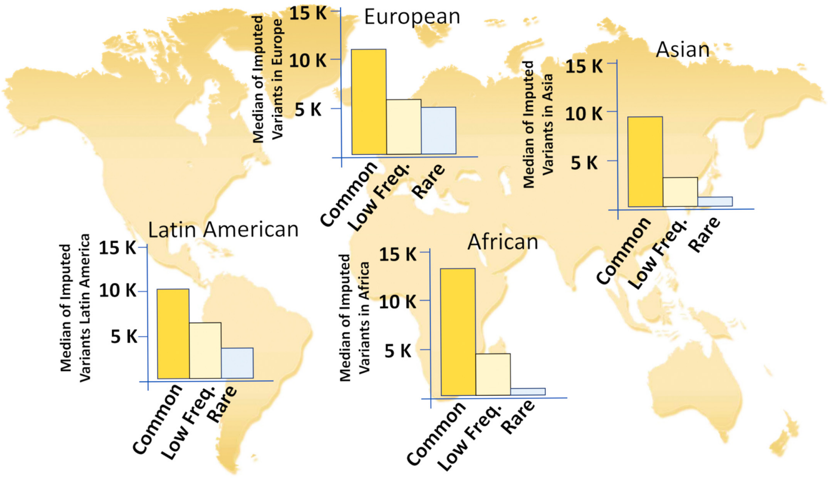


Figure 6. Structural Variant imputation performance using GCAT|Panel across all continents. European and Latin American populations recover more low frequency and rare SVs at high info scores (≥0.7) than African and Asian populations (see also Supplementary Figures S22 and S23).

quality of SVs was higher within the European populations (Supplementary Figure S22), the GCAT|Panel was also able to impute a large fraction of SVs across all other ethnicities (Figure 6, Supplementary Figure S23a). Of nearly 50K unique SVs imputed across all groups, 25%, 35% and 40% of them were detected within the Asian, African and Latin American populations, respectively (Figure 6, Supplementary Figure S23). In agreement with the mixed origin of Latin Americans, nearly half of all imputed variants within this group showed low-frequency values (MAF < 0.05), compared with other non-European groups, where the imputation covered predominantly common variants (Figure 6). In addition, 73% of all the structural variants identified and genotyped in previous studies, using long and short WGS (15,57) were also imputed by our panel on the same individuals, with 88% of matching genotypes (Supplementary Figures S19b and S20a).

DISCUSSION

Here, we present the GCAT|Panel, the first Iberian Haplotype reference panel derived from high-coverage whole-genome sequencing. The strategy developed for variant identification, classification, and phasing, has provided a comprehensive and high-quality catalogue of genetic variants, with low rates of false-positive calls and genotyping errors for all variant types, including SVs. This is due to the combination of high sequencing coverage (30×) with a comprehensive analysis strategy that integrates multiple variant callers and a Logistic Regression Model for maximzing recall and precision for each SV type and size.

Increasing the sequencing coverage to 30× allowed us to resolve a large fraction of SVs and accurately define the genotypes that cannot be properly defined with lower sequencing depths. In addition, while previous projects inferred SVs into phased haplotype scaffolds (11,12), our sequencing coverage allows us, for the first time, to phase SVs together with biallelic SNVs and indels, and to use phase informative reads (PIRs), which are expected to improve the imputation of rare variants (72). With this sequencing technology, we also expect a slight detection bias against low complexity (repeated) regions of the genome, where short-read sequencing tends to be less informative, in contrast to long-read sequencing technology (13–16). This is further highlighted by the high portion (54%) of our SVs affecting genes or regulatory regions, which also tend to be within the non-repetitive portion of the genome.

Given the increasing incorporation of whole-genome sequencing into genetic studies, it is crucial to highlight the importance of accurately identifying and resolving SVs with the correct genotype, to then obtain robust and meaningful results during the imputation in a different cohort. Here, we found a positive correlation between the number of neighbouring variants in LD with SVs and their quality of imputation, suggesting that variants with a high genotyping error show a lower number of variants in LD, which translates into a lower imputation accuracy for those variants (Supplementary Figure S17). On the other hand, software limitations (PLINK or ShapeIt4), can translate into poor estimations of haplotypes and LD, directly hampering the association test, which relies on accurate counts of variant allele frequencies and states. Improved variant calling strategies that can accurately identify and define complex structural variation events are still needed, together with new and dedicated analysis frames (e.g. phasing and LD) for SVs, where the actual size and type of the variant is considered, in contrast to the current scenario where SVs are taken as SNVs.

In our cohort, the GCAT|Panel led to the identification of potential risk SV, including those within the rare spectrum. Here, we highlight the identification of a rare polymorphic 324 bp-long AluYa5 element in chromosome 3 (g.49494276_49494600, MAF = 0.0013) associated with Mononeuritis of the lower limb (ICD-9 355). This SV is located within a multi enhancer-elite element (GeneCards) (73), proximal to *DAG1*, a gene involved in pathways responsible for neuromuscular diseases, and already causing severe limb-girdle muscular dystrophy type 2P (LGMD2P) through missense point mutations (74). Further studies are now needed to validate the resulting hypothesis, in which

Downloaded from https://academic.oup.com/nar/article/50/5/2464/6527675 by eFADA user on 20 May 2023

this Alu element could be affecting the expression of the *DAG1* gene in this disease.

This study also provides detailed guidance for the comprehensive analysis of whole-genome sequences, including the identification, classification, and phasing of SVs. We expect that this type of analysis will soon become the standard within large genetic studies that are already incorporating whole-genome Illumina sequences and combining them with existing genotyping array information.

Taken together, the availability of a high-quality haplotype panel, including a comprehensive fraction of structural variability, will significantly impact evolutionary and biomedical studies at different levels. The possibility of enriching current genome-wide association studies (e.g. GWAS and eQTL) with SVs through imputation, directly increases the chances of variant discovery, as well as of their functional interpretations. Our analysis evidence the potential of using population-matched reference panels, for the identification of rare structural variants involved in disease, and the important contribution to the understanding of the underlying genomic architecture of genetic diseases.

RESOURCE AVAILABILITY

Below we attach the information of the data and code availability used in this study.

DATA AVAILABILITY

The data generated in this study, including the FASTQ, BAM and VCF files of the 808 individuals with their genotyping information, as well as the entire GCAT|Panel, are accessible upon request (rdecid@igtp.cat) from the European Genome-phenome Archive (EGA), under the accession number EGAS00001003018. All the GCAT catalogue variants, the SV (Figure 3A), SNVs and indels annotations files, and the *in-silico* information (i.e. FASTQ, BAM files, catalogue of variants inserted) are available at http://cg.bsc.es/GCAT_BSC_iberianpanel.

All original code has been deposited at (https://github.com/gcatbiobank/GCAT_panel) and is publicly available as of the date of publication. DOIs are listed in the key resources table.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Lluís Puig and Vanessa Plegezuelos on behalf of Blood and Tissue Bank from Catalonia (BST), who collaborated in GCAT recruitment, and all the GCAT volunteers that participated in the study. We also thank the Centro Nacional de Análisis Genómico (CNAG-CRG), who collaborated in the sequence analysis of the study; members of the Comparative and Functional Genomics Group at the UAB for helping with the inversion validation; Dr. Francesc Calafell for his comments on the manuscript; Marta Morell from qGenomics for the technical assistance in the Alu validation; the Computational Genomics Group at the BSC, specially Ignasi Morán and Lorena Alonso, for their helpful discussions and valuable comments on the manuscript; and the technical support group from the Barcelona Supercomputing Center Center. We acknowledge Red Española de Supercomputación (RES) for awarding us access to MareNostrum4 supercomputer from Barcelona Supercomputing Center (proposal numbers BCV-2018-3-0010 and BCV-2019-1-0015). Figure 1a has been Adapted/Translated by permission from Springer Nature: Nature Genes mirror geography within Europe, John Novembre *et al.*, August 31, 2008.

Author contributions: J.V.M., I.G.F., D.M.S., D.T. and R.dC. designed and planned the whole study. J.V.M., I.G.F., D.M.S., D.T. and R.dC. contributed to the writing and editing of the manuscript. A.C. prepared and QCed samples for NGS. J.V.M. created the *in-silico* and D.M.S. prepared the NA12878 sample. J.V.M., I.G.F. and D.M.S. performed the variant calling and designed the Logistic Regression Models. J.V.M., I.G.F., D.M.S. contributed to the creation of BAM files, quality control, variant merging, filtering and genotyping, with the collaboration of M.P. J.V.M., I.G.F. and D.M.S. performed a comprehensive comparative analysis of the Iberian catalogue with other repositories. J.V.M., I.G.F. and D.M.S. conducted the SNV, indel and large deletion and duplication validations, collaborating with L.A. on behalf of qGenomics conducting CGH validation analysis. The inversion validation and genotype data analysis was provided by J.L.J., M.Pu and M.C. J.V.M., I.G.F. and D.M.S. performed the SV annotation, the creation of the GCAT|Panel, and their benchmarking, in collaboration with C.S. R.A. executed and adapted GUIDANCE with ShapeIt4 and GCAT|Panel. I.G.F., N.B., X.F., B.C. conducted and analysed the Phenome analysis, from phenotype extraction to GWAS analysis. N.B., R.dC., X.F. and L.A. conducted the AluY5a validation. L.S., J.F.S.H. conducted SNP-Array validation analysis of SV. V.M. and M.Pe contributed to the editing of the manuscript. D.T. and R.dC. supervised the study. All authors reviewed and approved the manuscript. J.V.M., I.G.F. and D.M.S. contributed equally to this study.

FUNDING

GCAT|Genomes for Life, a cohort study of the Genomes of Catalonia, Fundació Institut Germans Trias i Pujol (IGTP); IGTP is part of the CERCA Program/Generalitat de Catalunya; GCAT is supported by Acción de Dinamización del ISCIII-MINECO; Ministry of Health of the Generalitat of Catalunya [ADE 10/00026]; Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) [2017-SGR 529]; B.C. is supported by national grants [PI18/01512]; X.F. is supported by VEIS project [001-P-001647] (co-funded by European Regional Development Fund (ERDF), 'A way to build Europe'); a full list of the investigators who contributed to the generation of the GCAT data is available from www.genomesforlife.com/; Severo Ochoa Program, awarded by the Spanish Government [SEV-2011-00067 and SEV2015-0493]; Spanish Ministry of Science [TIN2015-65316-P]; Innovation and by the Generalitat de Catalunya [2014-SGR-1051 to D.T.]; Agencia Estatal de Investigación (AEI, Spain) [BFU2016-77244-

Downloaded from https://academic.oup.com/nar/article/50/5/2464/6527675 by eFADA user on 20 May 2023

R and PID2019-107836RB-I00]; European Regional Development Fund (FEDER, EU) (to M.C.); Spanish Ministry of Science and Innovation [FPI BES-2016-0077344 to J.V.M.]; C.S. received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement [H2020-MSCA-COFUND-2016-754433]; this study made use of data generated by the UK10K Consortium from UK10K COHORT IMPUTATION [EGAS00001000713]; formal agreement with the Barcelona Supercomputing Center (BSC); this study made use of data generated by the Genome of the Netherlands' project, which is funded by the Netherlands Organization for Scientific Research [184021007], allowing us to use the GoNL reference panel containing SVs, upon request (GoNL Data Access request 2019203); this study also used data generated by the Haplotype Reference Consortium (HRC) accessed through the European Genome-phenome Archive with the accession numbers EGAD00001002729; formal agreement of the Barcelona Supercomputing Center (BSC) with WTSI; this study made use of data generated by the 1000 Genomes (1000G), accessed through the FTP portal (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/); this study used the GeneHancer-for-AnnotSV dump for GeneCards Suite Version 4.14, through a formal agreement between the BSC and the Weizmann Institute of Science. Funding for open access charge: GCAT|Genomes for Life, a cohort study of the Genomes of Catalonia, Fundació Institut Germans Trias i Pujol (IGTP); IGTP is part of the CERCA Program/Generalitat de Catalunya; GCAT is supported by Acción de Dinamización del ISCIII-MINECO; Ministry of Health of the Generalitat of Catalunya [ADE 10/00026]; Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) [2017-SGR 529]; B.C. is supported by national grants [PI18/01512]; X.F. is supported by VEIS project [001-P-001647] (co-funded by European Regional Development Fund (ERDF), 'A way to build Europe'); a full list of the investigators who contributed to the generation of the GCAT data is available from www.genomesforlife.com/; Severo Ochoa Program, awarded by the Spanish Government [SEV-2011-00067 and SEV2015-0493]; Spanish Ministry of Science [TIN2015-65316-P]; Innovation and by the Generalitat de Catalunya [2014-SGR-1051 to D.T.]; [Agencia Estatal de Investigación (AEI, Spain) [BFU2016-77244-R and PID2019-107836RB-I00]; European Regional Development Fund (FEDER, EU) (to M.C.); Spanish Ministry of Science and Innovation [FPI BES-2016-0077344 to J.V.M.]; C.S. received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement [H2020-MSCA-COFUND-2016-754433]; this study made use of data generated by the UK10K Consortium from UK10K COHORT IMPUTATION [EGAS00001000713]; formal agreement with the Barcelona Supercomputing Center (BSC); this study made use of data generated by the Genome of the Netherlands' project, which is funded by the Netherlands Organization for Scientific Research [184021007], allowing us to use the GoNL reference panel containing SVs, upon request (GoNL Data Access request 2019203); this study also used data generated by the Haplotype Reference Consortium (HRC) accessed through the European Genome-phenome Archive with the accession numbers EGAD00001002729; formal agreement of the Barcelona Supercomputing Center (BSC) with WTSI; this study made use of data generated by the 1000 Genomes (1000G), accessed through the FTP portal (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/); this study used the GeneHancer-for-AnnotSV dump for GeneCards Suite Version 4.14, through a formal agreement between the BSC and The Weizmann Institute of Science.

Conflict of interest statement. None declared.

REFERENCES

1. Belmont,J.W., Boudreau,A., Leal,S.M., Hardenbol,P., Pasternak,S., Wheeler,D.A., Willis,T.D., Yu,F., Yang,H., Gao,Y. *et al.* (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
2. Loh,P., Danecek,P., Palamara,P.F., Fuchsberger,C., Reshef,A., Finucane,H.K., Schoenherr,S., Forer,L., Mccarthy,S., Abecasis,G.R. *et al.* (2016) Reference-based phasing using the haplotype reference consortium panel. *Nat. Genet.*, **48**, 1443–1448.
3. Taliun,D., Harris,D.N., Kessler,M.D., Carlson,J., Szpiech,Z.A., Torres,R., Taliun,S.A.G., Corvelo,A., Gogarten,S.M., Kang,H.M. *et al.* (2021) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature*, **590**, 290–299.
4. Weischenfeldt,J., Symmons,O., Spitz,F. and Korbel,J.O. (2013) Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.*, **14**, 125–138.
5. Brandler,W.M., Antaki,D., Gujral,M., Kleiber,M.L., Whitney,J., Maile,M.S., Hong,O., Chapman,T.R., Tan,S., Tandon,P. *et al.* (2018) Paternally inherited cis-regulatory structural variants are associated with autism. *Science*, **20**, 327–331.
6. González,J.R., Ruiz-Arenas,C., Cáceres,A., Morán,I., López-Sánchez,M., Alonso,L., Tolosana,I., Guindo-Martínez,M., Mercader,J.M., Esko,T. *et al.* (2020) Polymorphic inversions underlie the shared genetic susceptibility of obesity-related diseases. *Am. J. Hum. Genet.*, **106**, 846–858.
7. Thibodeau,M.L., O'Neill,K., Dixon,K., Reisle,C., Mungall,K.L., Krzywinski,M., Shen,Y., Lim,H.J., Cheng,D., Tse,K. *et al.* (2020) Improved structural variant interpretation for hereditary cancer susceptibility using long-read sequencing. *Genet. Med.*, **22**, 1892–1897.
8. Manolio,T.A., Collins,F.S., Cox,N.J., Goldstein,D.B., Hindorff,L.A., Hunter,D.J., McCarthy,M.I., Ramos,E.M., Cardon,L.R., Chakravarti,A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
9. Becker,T., Lee,W.P., Leone,J., Zhu,Q., Zhang,C., Liu,S., Sargent,J., Shanker,K., Mil-homens,A., Cerveira,E. *et al.* (2018) FusorSV: an algorithm for optimally combining data from multiple structural variation detection methods. *Genome Biol.*, **19**, 38.
10. Collins,R.L., Brand,H., Karczewski,K.J., Zhao,X., Alföldi,J., Francioli,L.C., Khera,A. V., Lowther,C., Gauthier,L.D., Wang,H. *et al.* (2020) A structural variation reference for medical and population genetics. *Nature*, **581**, 444–451.
11. Sudmant,P.H., Rausch,T., Gardner,E.J., Handsaker,R.E., Abyzov,A., Huddleston,J., Zhang,Y., Ye,K., Jun,G., Hsi-Yang Fritz,M. *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**, 75–81.
12. Hehir-Kwa,J.Y., Marschall,T., Kloosterman,W.P., Francioli,L.C., Baaijens,J.A., Dijkstra,L.J., Abdellaoui,A., Koval,V., Thung,D.T., Wardenaar,R. *et al.* (2016) A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat. Commun.*, **7**, 12989.
13. Chaisson,M.J.P., Sanders,A.D., Zhao,X., Malhotra,A., Porubsky,D., Rausch,T., Gardner,E.J., Rodriguez,O.L., Guo,L., Collins,R.L. *et al.* (2019) Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.*, **10**, 1784.
14. Ho,S.S., Urban,A.E. and Mills,R.E. (2020) Structural variation in the sequencing era. *Nat. Rev. Genet.*, **21**, 171–189.
15. Audano,P.A., Sulovari,A., Graves-Lindsay,T.A., Cantsilieris,S., Sorensen,M., Welch,A.E., Dougherty,M.L., Nelson,B.J., Shah,A., Dutcher,S.K. *et al.* (2019) Characterizing the major structural variant alleles of the human genome. *Cell*, **176**, 663–675.

Downloaded from https://academic.oup.com/nar/article/50/5/2464/6527675 by eFADA user on 20 May 2023

16. Ebert,P., Audano,P.A., Zhu,Q., Rodriguez-Martin,B., Porubsky,D., Bonder,M.J., Sulovari,A., Ebler,J., Zhou,W., Mari,R.S. *et al.* (2021) Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, **372**, eabf7117.
17. Obón-Santacana,M., Vilardell,M., Carreras,A., Duran,X., Velasco,J., Galván-Femenía,I., Alonso,T., Puig,L., Sumoy,L., Duell,E.J. *et al.* (2018) GCAT|Genomes for life: a prospective cohort study of the genomes of catalonia. *BMJ Open*, **8**, e018324.
18. Galván-Femenía,I., Obón-Santacana,M., Piñeyro,D., Guindo-Martinez,M., Duran,X., Carreras,A., Pluvinet,R., Velasco,J., Ramos,L., Aussó,S. *et al.* (2018) Multitrait genome association analysis identifies new susceptibility genes for human anthropometric variation in the GCAT cohort. *J. Med. Genet.*, **55**, 765–778.
19. Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.M., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C., Stuart,J.M. and Cancer Genome Atlas Research Network. (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
20. Huang,W., Li,L., Myers,J.R. and Marth,G.T. (2012) ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–594.
21. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**, 1754–1760.
22. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
23. Van der Auwera,G.A., Carneiro,M.O., Hartl,C., Poplin,R., Del Angel,G., Levy-Moonshine,A., Jordan,T., Shakir,K., Roazen,D., Thibault,J. *et al.* (2013) From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics*,**43**:11.10.1–11.10.33.
24. Zook,J.M., McDaniel,J., Olson,N.D., Wagner,J., Parikh,H., Heaton,H., Irvine,S.A., Trigg,L., Truty,R., McLean,C.Y. *et al.* (2019) An open resource for accurately benchmarking small variant and reference calls. *Nat. Biotechnol.*, **37**, 561–566.
25. Poplin,R., Ruano-Rubio,V., DePristo,M.A., Fennell,T.J., Carneiro,M.O., Auwera,G.A. Van der, Kling,D.E., Gauthier,L.D., Levy-Moonshine,A., Roazen,D. *et al.* (2017) Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv doi: https://doi.org/10.1101/201178, 24 July 2018, preprint: not peer reviewed.
26. Poplin,R., Chang,P.C., Alexander,D., Schwartz,S., Colthurst,T., Ku,A., Newburger,D., Dijamco,J., Nguyen,N., Afshar,P.T. *et al.* (2018) A universal snp and small-indel variant caller using deep neural networks. *Nat. Biotechnol.*, **36**, 983.
27. Kim,S., Scheffer,K., Halpern,A., Bekritsky,M., Enhuo,N., Källberg,M., Chen,X., Yeobin,K., Beyter,D., Krusche,P. *et al.* (2018) Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods*, **15**, 591–594.
28. Rimmer,A., Phan,H., Mathieson,I., Iqbal,Z. and Twigg,S.R.F. (2014) Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet*, **46**, 912–918.
29. Koboldt,D.C., Larson,D.E. and Wilson,R.K. (2013) Using varscan 2 for germline variant calling and somatic mutation detection. *Curr Protoc Bioinforma.*, **44**, 15.4.1–15.4.17.
30. Rausch,T., Zichner,T., Schlattl,A., Stütz,A.M., Benes,V. and Korbel,J.O. (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**, 333–339.
31. Chen,X., Schulz-Trieglaff,O., Shaw,R., Barnes,B., Schlesinger,F., Källberg,M., Cox,A.J., Kruglyak,S. and Saunders,C.T. (2016) Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, **32**, 1220–1222.
32. Ye,K., Guo,L., Yang,X., Lamijer,E.W., Raine,K. and Ning,Z. (2018) Split-read indel and structural variant calling using PINDEL. *Methods Mol. Biol.*, **1833**, 95–105.
33. Layer,R.M., Chiang,C., Quinlan,A.R. and Hall,I.M. (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.*, **15**, R84.
34. Kronenberg,Z.N., Osborne,E.J., Cone,K.R., Kennedy,B.J., Domyan,E.T., Shapiro,M.D., Elde,N.C. and Yandell,M. (2015) Wham: identifying structural variants of biological consequence. *PLoS Comput. Biol.*, **11**, e1004572.
35. Wala,J.A., Bandopadhayay,P., Greenwald,N.F., O'Rourke,R., Sharpe,T., Stewart,C., Schumacher,S., Li,Y., Weischenfeldt,J., Yao,X. *et al.* (2018) SvABA: Genome-wide detection of structural variants and indels by local assembly. *Genome Res.*, **28**, 581–591.
36. Abyzov,A., Urban,A.E., Snyder,M. and Gerstein,M. (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.*, **21**, 974–984.
37. Kehr,B., Melsted,P. and Halldórsson,B.V. (2016) PopIns: Population-scale detection of novel sequence insertions. *Bioinformatics*, **32**, 961–967.
38. Handsaker,R.E., Van Doren,V., Berman,J.R., Genovese,G., Kashin,S., Boettger,L.M. and Mccarroll,S.A. (2015) Large multiallelic copy number variations in humans. *Nat. Genet.*, **47**, 296–303.
39. Kavak,P., Lin,Y.Y., Numanagić,I., Asghari,H., Güngör,T., Alkan,C. and Hach,F. (2017) Discovery and genotyping of novel sequence insertions in many sequenced individuals. *Bioinformatics*, **33**, i161–i169.
40. Liu,S., Huang,S., Rao,J., Ye,W., Krogh,A. and Wang,J. (2015) Discovery, genotyping and characterization of structural variation and novel sequence at single nucleotide resolution from de novo genome assemblies on a population scale. *Gigascience*, **4**, 64.
41. Gardner,E.J., Lam,V.K., Harris,D.N., Chuang,N.T., Scott,E.C., Stephen Pittard,W., Mills,R.E. and Devine,S.E. (2017) The mobile element locator tool (MELT): Population-scale mobile element discovery and biology. *Genome Res.*, **27**, 1916–1929.
42. Tischler,G. and Leonard,S. (2014) Biobambam: tools for read pair collation based algorithms on BAM files. *Source Code Biol. Med.*, **9**, 13.
43. Rausch,T., Fritz,Hsi-Yang, Korbel,M. and Benes,V. (2019) Alfred: interactive multi-sample BAM alignment statistics, feature counting and feature annotation for long- and short-read sequencing. *Bioinformatics*, **35**, 2489–2491.
44. Jun,G., Flickinger,M., Hetrick,K.N., Romm,J.M., Doheny,K.F., Abecasis,G.R., Boehnke,M. and Kang,H.M. (2012) Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.*, **91**, 839–848.
45. Novembre,J., Johnson,T., Bryc,K., Kutalik,Z., Boyko,A.R., Auton,A., Indap,A., King,K.S., Bergmann,S., Nelson,M.R. *et al.* (2008) Genes mirror geography within europe. *Nature*, **456**, 98–101.
46. Sherry,S.T., Ward,M. and Sirotkin,K. (2001) dbSNP-Database for Single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.*, **1**, 1–8.
47. Lappalainen,I., Lopez,J., Skipper,L., Hefferon,T., Spalding,J.D., Garner,J., Chen,C., Maguire,M., Corbett,M., Zhou,G. *et al.* (2013) DbVar and DGVa: public archives for genomic structural variation. *Nucleic. Acids. Res.*, **41**, 936–941.
48. Abel,H.J., Larson,D.E., Regier,A.A., Chiang,C., Das,I., Kanchi,K.L., Layer,R.M., Neale,B.M., Salerno,W.J., Reeves,C. *et al.* (2020) Mapping and characterization of structural variation in 17,795 human genomes. *Nature*, **583**, 83–89.
49. Lerga-Jaso,J. (2019) In: *Integrative Analysis of the Functional Consequences of Inversions in the Human Genome*. Univ. Autònoma Barcelona.
50. Howie,B.N., Donnelly,P. and Marchini,J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLos Genet.*, **5**, e1000529.
51. Haraksingh,R.R., Abyzov,A. and Urban,A.E. (2017) Comprehensive performance comparison of high-resolution array platforms for genome-wide copy number variation (CNV) analysis in humans. *BMC Genomics*, **18**, 321.
52. Delaneau,O., Howie,B., Cox,A.J., Zagury,J.F. and Marchini,J. (2013) Haplotype estimation using sequencing reads. *Am. J. Hum. Genet.*, **93**, 687–696.
53. Menelaou,A. and Marchini,J. (2013) Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics*, **29**, 84–91.
54. Delaneau,O., Zagury,J.F., Robinson,M.R., Marchini,J.L. and Dermitzakis,E.T. (2019) Accurate, scalable and integrative haplotype estimation. *Nat. Commun.*, **10**, 24–29.
55. Patterson,M.D., Marschall,T., Pisanti,N., Van Iersel,L., Stougie,L., Klau,G.W. and Schönhuth,A. (2015) WhatsHap: weighted haplotype

Downloaded from https://academic.oup.com/nar/article/50/5/2464/6527675 by eFADA user on 20 May 2023

assembly for future-generation sequencing reads. *J. Comput. Biol.*, **22**, 498–509.

56. Via,M., Gignoux,C. and Burchard,E.G. (2010) The 1000 genomes project: new opportunities for research and social challenges. *Genome Med.*, **2**, 8–10.
57. Hickey,G., Heller,D., Monlong,J., Sibbesen,J.A., Sirén,J., Eizenga,J., Dawson,E.T., Garrison,E., Novak,A.M. and Paten,B. (2020) Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.*, **21**, 35.
58. Walter,K., Min,J.L., Huang,J., Crooks,L., Memari,Y., McCarthy,S., Perry,J.R.B., Xu,C., Futema,M., Lawson,D. *et al.* (2015) The UK10K project identifies rare variants in health and disease. *Nature*, **526**, 82–89.
59. Guindo-martínez,M., Amela,R., Bonàs-guarch,S., Salvoro,C., Miguel-escalada,I., Carey,C.E., Cole,J.B., Rüeger,S., Atkinson,E., Leong,A. *et al.* (2021) The impact of non-additive genetic associations on age-related complex diseases. *Nat. Commun.*, **12**, 2436.
60. Geoffroy,V., Herenger,Y., Kress,A., Stoetzel,C., Piton,A., Dollfus,H. and Muller,J. (2018) AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics*, **34**, 3572–3574.
61. Fishilevich,S., Nudel,R., Rappaport,N., Hadar,R., Plaschkes,I., Iny Stein,T., Rosen,N., Kohn,A., Twik,M., Safran,M. *et al.* (2017) GeneHancer: genome-wide integration of enhancers and target genes in genecards. *Database (Oxford)*, **2017**, bax028.
62. Kumar,S., Harmanci,A., Vytheeswaran,J. and Gerstein,M.B. (2019) SVFX: a machine-learning framework to quantify the pathogenicity of structural variants. *Genome Biol.*, **21** , 274.
63. Cingolani,P., Platts,A., Lely Wang,L., Coon,M., Nguyen,T., Wang,L., Land,S., Lu,X. and Ruden,D. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80–92.
64. Cingolani,P., Patel,V.M., Coon,M., Nguyen,T., Land,S.J., Ruden,D.M. and Lu,X. (2012) Using drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, snpsift. *Front. Genet.*, **3**, 35.
65. Landrum,M.J., Lee,J.M., Benson,M., Brown,G.R., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Jang,W. *et al.* (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, **46**, D1062–D1067.
66. Rentzsch,P., Witten,D., Cooper,G.M., Shendure,J. and Kircher,M. (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic. Acids. Res.*, **47**, D886–D894.
67. Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A.R., Bender,D., Maller,J., Sklar,P., De Bakker,P.I.W., Daly,M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
68. Sutton,A., Crew,A. and Wysong,A. (2016) Redefinition of skin cancer as a chronic disease. *JAMA Dermatol.*, **152**, 255–256.
69. Chi,M. ju, Lee,C. yi and Wu,S. chong (2011) The prevalence of chronic conditions and medical expenditures of the elderly by chronic condition indicator (CCI). *Arch. Gerontol. Geriatr.*, **52**, 284–289.
70. Friedman,B., Jiang,H.J., Elixhauser,A. and Segal,A. (2006) Hospital inpatient costs for adults with multiple chronic conditions. *Med. Care Res. Rev.*, **63**, 327–346.
71. Fuller,Z.L., Berg,J.J., Mostafavi,H., Sella,G. and Przeworski,M. (2019) Measuring intolerance to mutation in human genetics. *Nat. Genet.*, **51**, 772–776.
72. Marchini,J. (2019) Haplotype estimation and genotype imputation. In: David,B., Ida,M. and John,M. (eds). *Handbook of Statistical Genomics*. John Wiley & Sons Ltd, Vol. **1**, 87–114.
73. Stelzer,G., Rosen,N., Plaschkes,I., Zimmerman,S., Twik,M., Fishilevich,S., Iny Stein,T., Nudel,R., Lieder,I., Mazor,Y. *et al.* (2016) The genecards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinforma.*, **2016**, 1.30.1–1.30.33.
74. Hara,Y., Balci-Hayta,B., Yoshida-Moriguchi,T., Kanagawa,M., Beltrán-Valero de Bernabé,D., Gündeşli,H., Willer,T., Satz,J.S., Crawford,R.W., Burden,S.J. *et al.* (2011) A dystroglycan mutation associated with limb-girdle muscular dystrophy. *N. Engl. J. Med.*, **364**, 939–946.

Downloaded from https://academic.oup.com/nar/article/50/5/2464/6527675 by eFADA user on 20 May 2023