



UNIVERSITAT^{DE}
BARCELONA

Comprehensive identification and characterisation of germline structural variation within the Iberian population

Jordi Valls Margarit

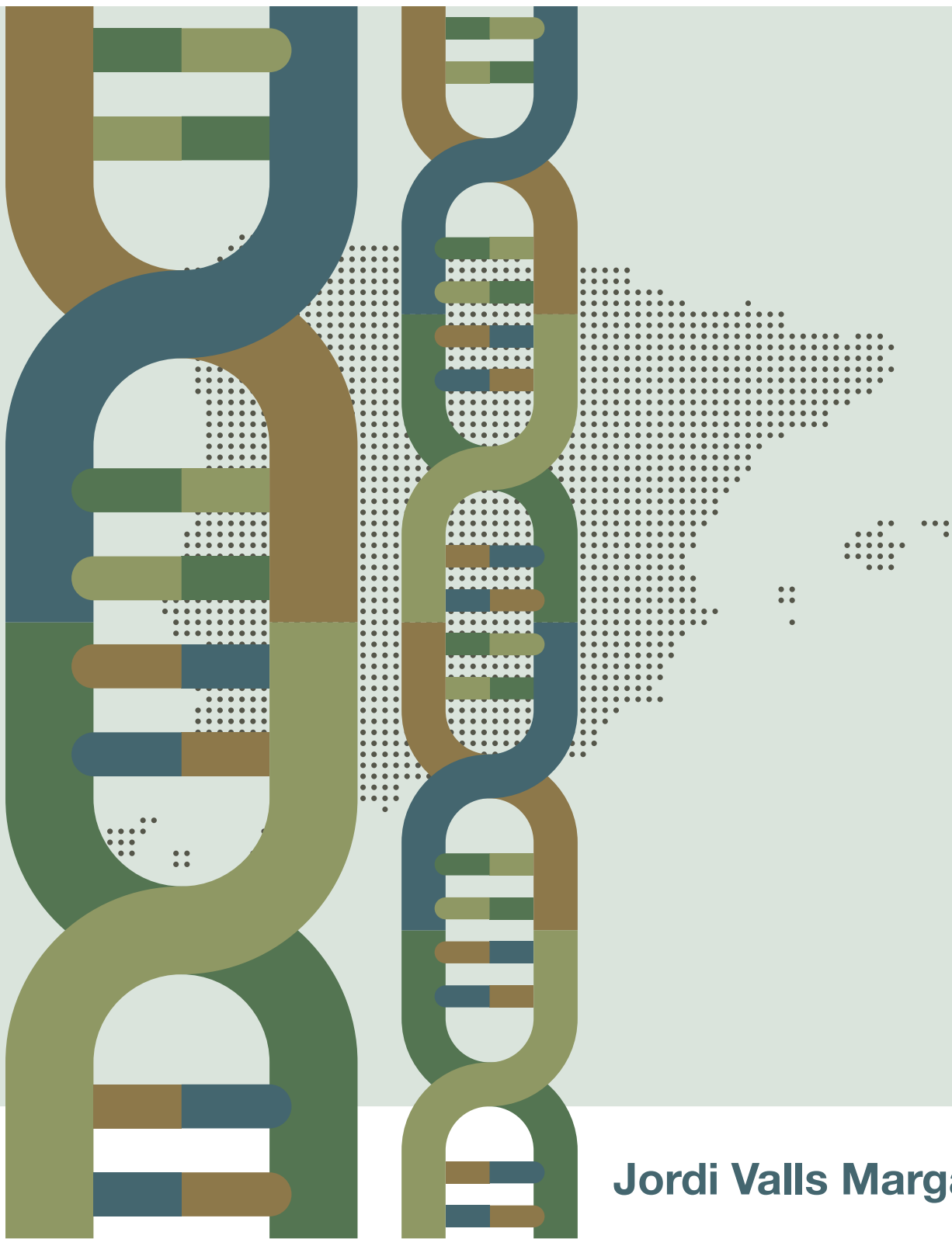


Aquesta tesi doctoral està subjecta a la llicència **Reconeixement 4.0. Espanya de Creative Commons.**

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

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

Comprehensive identification and characterisation of germline structural variation within the Iberian population



Jordi Valls Margarit

Facultat de Biologia, Universitat de Barcelona
Programa de Doctorat en Biomedicina (HDK05)
Línia de recerca 101114 - Bioinformàtica

Comprehensive identification and characterisation of germline structural variation within the Iberian population

Memòria presentada per Jordi Valls Margarit per optar al grau de doctor per la
Universitat de Barcelona

Tesis realitzada al
Barcelona Supercomputing Center (BSC)

Doctorand

Jordi Valls Margarit



Director

David Torrents Arenales



Tutor

Josep Lluís Gelpí Buchaca

Agraïments

Primer de tot, m'agradaria dedicar unes paraules a tota la gent que d'una manera o altre han fet possible que jo hagi arribat a fer aquesta tesi, sense ells no se quin camí hagués seguit.

En primer lloc, voldria agrair al meu tutor David Torrents per donar-me la possibilitat de poder fer una tesi en el seu grup, de creure en mi tot hi no tenir ni idea de bioinformàtica. Encara ara recordo al frase de "Amb l'energia que tens pots fer una tesi en bioinformàtica, i aprendràs molt", no s'equivocava. També voldria agrair-li la paciència i temps dipositats en mi, a mes de totes les trucades i consells, per així poder desenvolupar aquesta tesi i paper, moltes gràcies per tot.

No pensar en el "GCAT Crew" en aquesta tesi seria un delictes. L'Iván i el Dani han estat un pilar fonamental en el transcurs d'aquesta tesi, treballant en el projecte del panell de referència d'ibèrics. D'ells me'n porto la constància i perseverança per tal d'assolir l'objectiu de fer un projecte complex com aquest, no ens imaginàvem pas l'esforç que hem hagut d'invertir per tenir uns resultats dignes de "Nature!", i lluny de deixar-ho córrer hem seguit amb intensitat, no obstant crec que el resultat a valgut la pena. Part dels resultats presentats en aquesta tesi ha estat gràcies al seu esforç. Fer menció a la paciència de l'Iván, qui m'ha ajudat a aprendre a programar en R, a mes de tota la feina que tenia. Tots els equips haurien de tenir algú com ell. A més crec que me'n porto dues amistats, fruit de les hores que hem passat junts, discutint resultats, fent birres o fins hi tot partits de tenis. Moltes gràcies per aguantar-me, se que no ha estat fàcil.

També voldria dedicar unes línies a tota la gent que estan i han passat pel nostre grup de "Computational Genomics Group".

No podria començar amb ningú altre que la Luisa, amb qui he compartit la majoria d'alegries i decepcions. Recordo com amb paciència i comprensió em va ajudar a donar els primers passos en la bioinformàtica, entenent com funciona la seqüenciació, a més de començar a programar. A més, ella ha estat un pilar emocional, on poder compartir tot el que ens ha passat durant aquests anys. Espero que puguem seguir compartint experiències i anar al congost de Mont-Rebei a celebrar plegats el final d'aquesta etapa.

El mateix m'ha passat amb la Lorena, la súper matemàtica del grup. Durant aquests anys he pogut descobrir que darrera la seva timidesa, s'hi amaga una gran persona, amb un positivisme desenfrenat. Hem compartit molts moments, com alguna aposta... que no hem arribat a executar... discutir resultats (quan m'aixecava i venia a veure't la teva cara de "ai..." era memorable), com també ajudar-me a programar en python.

També el meu company de taula Ignasi, que ha tingut la paciència en ajudar-me en molts aspectes de la tesi, com escriptura, discutir resultats o revisar-la. Tot començava amb "Ignasi tinc una pregunta..." i acabàvem arreglant tota la tesi pràcticament. O quan fent una figura, tot hi esperar un "esta molt bé", sempre m'hi trobava millores. Moltes gràcies per dedicar-me el teu temps.

La Cecilia, una de les persones més recents en entrar al grup, i que he tingut la sort de conèixer. La seva passió per la ciència no te límits, tan de bo poguéssim haver treballat junts. M'ha encantat poder parlar de ciència, discutir resultats i anar a fer birres, com també donar-me consells i revisar aquesta tesi, moltes gràcies per tot.

Què dir de l'Álvaro, tan correcte i educat, i que li encanta la ciència i aprendre. Amb ell hem compartit moltes xerrades sobre com millorar les nostres respectives feines, i fent-nos preguntes constantment sobre resultats i aspectes de la bioinformàtica. Durant aquests anys m'he sentit identificat en ell en quant a "rizar el rizo", i a partir de discutir els resultats fer uns passos endavant. A part, també li agraeixo tot l'esforç que m'ha donat durant aquest temps.

Amb l'Ana tot hi no començar en bon peu, al final ens hem pogut conèixer i compartir bons moments. També li vull agrair l'ajuda que m'ha donat en poder alinear reads i explicar amb paciència com fer-ho de la manera correcte.

Voldria agrair especialment a la Montse, qui lluita cada dia perquè el grup pugui funcionar, vigilant l'espai que ocupem, o no ens carreguem el supercomputador. A part, vull fer menció a la seva paciència i donant-nos suport sempre que ho hem necessitat.

Després hi ha un grapat de gent que ha passat pel nostre grup, a qui també vull agrair la seva ajuda, com l'Àlex, qui amb molta paciència em va ajudar a programar en python, sense el seu temps, de ben segur que no hagués pogut aprendre a la velocitat que ho vaig fer. Al Flo, que vam compartir molts moments, a part de ensenyar-li les paraules claus del català. La Michelle, que amb la seva alegria, podia desconnectar dels problemes de la tesis. La Mercè, Elias i Marta, malgrat no haver pogut establir una relació propera, agrair-los l'ajuda que em van prestar al inici de tesis. I també, com si fos del grup la Romina, qui m'ha donat un cop de mà sempre que la he necessitat, acabant sempre amb un "Suerte en la vida", ara si, espero que tinguem sort!.

També donar gràcies al Rafa de Cid, per donar-me la possibilitat de poder treballar amb les dades del projecte GCAT, sense la seva col·laboració aquesta tesis no s'hagués pogut dur a terme.

Sortint de l'àmbit laboral, vull dedicar unes paraules al dermatòleg Dr. Ignacio Umbert, qui em va introduir en el món de la medicina personalitzada. Amb ell vaig aprendre molt en dermatologia, i entendre que cada persona pot desenvolupar una malaltia de forma diversa, sent la "Low-grade inflammation" i el estrès els causants de moltes malalties, i que cal personalitzar els tractaments per tal de millorar la salut de les persones. Moltes gràcies Nacho per el teu suport, sempre t'estaré agraït.

Aquesta tesis s'ha fet menys estressant gràcies als "amics de la festa". En especial a l'Albert i l'Isaac, que hem sortit de festa moltes vegades, per desconnectar dels nostres problemes. A més l'Albert, que a partir del seu criteri m'ha ajudat a millorar pòsters, figures de tesis i també aquesta portada, mostrant-me que un disseny atractiu pot vendre més que la pròpia informació derivada d'ella. I la Laura i el Gerard, que entre jocs de taula i sobretot l'ajuda que sempre m'han donat, han fet que tot sigui més fàcil. Moltes gràcies a tot@s per estar amb mi sempre que ho he necessitat.

Mai podré agrair lo suficient el que han fet els meus pares per mi. A partir del seu esforç, sacrifici i il·lusió he pogut arribar on sóc ara, ells han estat un suport incondicional, intentant, no ara sinó sempre que jo fos feliç i intentés el que volgués sense tenir por a fracassar. De ben segur que aquesta empena i força l'he après de casa, per això els meus dos germans Dani i Maria i jo hem arribat tan lluny. Us estimo molt.

També vull dedicar unes paraules al meu germà gran Dani i bessona Maria. La metamorfosi que va experimentar el Dani a primer de Batxillerat, on perseguir la feina perfecte era una

obsessió, o la Maria, sempre revisant i fent una feina impecable al llarg de tota la vida. Jo si hagués de definir un tret que ens englobés als tres es “intensitat”, sempre ho donarem tot per fer la feina lo millor possible. Us estimo molt i se que sempre us tindrè al meu costat.

Donar gràcies també als meus cunyats Ernest i Gemma, qui durant aquest temps han vist la meva evolució en la tesis, i donant-me ànims quan ho necessitava. I també els meus nebots Mireia i Jaume, que amb la seva alegria m’han permès desconnectar de la pressió de la tesis.

També vull agrair als meus sogres Sara i Toni tota l’ajuda i el suport que m’han donat, són com uns pares per a mi, i tenir un lloc on sempre hi ets benvingut és un tresor, moltes gràcies per la vostre ajuda.

Per últim, la Carla, la persona més important de la meva vida, qui a patit el dia a dia de la tesis, recolzant-me i ajudant-me en tot el que he necessitat sense demanar-ho. M’ha tret de la cova per poder desconnectar i m’ha donat alegria quan més ho he necessitat. Tot hi que encara ara no sap ben bé que faig, sense ella de ben segur que no hauria acabat, t’estimo.

No vull acabar sense recordar a tota la gent que a perdut gent estimada per la COVID, en el meu cas el Lluís i la Pepi, qui sempre tenien interès en la meva feina i volien venir a la meva defensa, desgraciadament la vida no ens ho ha permès, part de la tesis també va dedicada a vosaltres.

Summary

One of the central aims of biology and biomedicine has been the characterisation and understanding of genetic variation across humans, to answer important evolutionary questions and to explain phenotypic variability concerning the diseases. Understanding genetic variability, is key to study this relationship (through imputation and GWASs) and to translate the results into improved clinical protocols. Different initiatives have emerged around the world to systematically characterise the genetic variability of specific human populations from whole-genome sequences, usually by selecting geographical regions. Examples such as 1000 Genomes (1000G)¹, GoNL², HRC, UK10K³ or Estonian population⁴, have already identified and characterised millions of genetic variants across different populations. In combination with imputation analysis, these sequenced-based projects allow increasing the statistical power and resolution of Genome-Wide Association Studies (GWAS), identifying and discovering new disease-associated variants⁵. Additionally, genetic variability among population groups is associated with geographic ancestry and can affect the disease risk or treatment efficacy differently^{6,7}. For this reason, population-specific reference panels are necessary to characterise their genetic diversity and to assess its effect on human phenotypes, improving GWAS studies, as one of the cornerstones of precision medicine⁷.

Existing genetic variability panels include Single Nucleotide Variants (SNVs) and indels (<50bp) but are limited in large Structural Variants (SV) (≥ 50 bp). Technical and methodological limitations hindered the discovery of SVs using Next-generation Sequencing (NGS) technologies, as it produced False-Discovery Rates between 9-89% and recall 10-70%, depending on the SV type and size⁸. On average, the genomic variation between two human genomes is around 0.1%, but this difference increases to 1.5% with SVs⁸. The SVs also affect 3-10 times more nucleotides than SNVs⁹ (4M SNVs per genome¹⁰), showing their potential effect on human phenotypes. For this reason, including a complete catalogue of SVs in reference panels will increase the power in GWAS studies and provide opportunities to find new disease-associated variants.

To overcome these limitations, in this thesis, we have generated the first genome-wide Iberian haplotype reference panel, mainly focused on Structural Variants, using 785 samples whole-genome sequenced (WGS) at high coverage (30X) from the GCAT-Genomics for life project. We designed a complete strategy, including an extensive benchmarking of multiple variant calling programs and by building specific Logistic Regression Models (LRM) for SV types, as well as phasing strategies to come up with a high quality and comprehensive genetic variability panel. This strategy was benchmarked using different controlled sets of variants, showing high precision and recall values across all variant types and sizes.

The application of this strategy to our GCAT whole-genome samples resulted in the identification of 35,431,441 genetic variants, classified as 30,325,064 SNPs, 5,017,19 small indels (< 50bp), and 89,178 larger SV (≥ 50 bp). The latter group was further subclassified into 33,244 deletions, 6,269 duplications, 12,782 insertions, 10,115 inversions, 18,779 transposons and 7,989 translocations, covering all ranges of frequencies and sizes. Besides, 60% of the discovered SVs were not catalogued in any repository, thus increasing the insights of SV in humans. Additionally, 52.44% of common and 71.63% of low-frequency SVs were not included in any haplotype reference panel. Thus, new SVs could be used in GWAS, adding more value to the Iberian-GCAT catalogue.

The prediction of the functional impact of the SVs shows that these variants might have a central role in several diseases. Of all SVs included in the Iberian-GCAT catalogue, 46% overlapped in genes (both protein-coding genes and non-protein-coding genes), highlighting their potential impact on human traits. Besides, 92.7% of protein-coding genes were located outside low-complexity (repeated) genomic regions, expecting short-reads from NGS to capture the most interpretable SVs in humans¹¹. Moreover, 32.93% of SVs affected protein-coding genes with a predicted loss of function intolerance (pLI) effect, further supporting the potential implication of these variants on complex diseases and therefore enabling a better explanation of missing heritability.

Importantly, taking advantage of high coverage (30X), we accurately determine the genotypes of SVs, enabling to phase together with SNVs and indels, and increasing the SV phasing accuracy, in contrast to 1000G and GoNL. Besides, high coverage allowed to use Phasing Informative Reads (PIRs), increasing the phasing performance. The overall strategy enables the community to expand and improve the imputation possibilities within GWAS.

The Iberian-GCAT haplotype reference panel created in this thesis, imputes accurately common SVs, with near ~100% of agreement with sequencing results. Although the Iberian-GCAT haplotype reference panel can be used in all populations from different continental groups, due to closer ancestries, the imputation performance is high in European and Latin American populations, reflected in the amount of low-frequency ($1\% \leq \text{MAF} < 5\%$) and rare ($1\% > \text{MAF}$) variants imputed at high info scores. These results demonstrated the versatility of our resource, increasing their performance in closer ancestries. In general, we observed that when the allele frequency decreases, the imputation accuracy drops too, highlighting the necessity to include more samples in reference panels, to impute low-frequency and rare variants efficiently, which normally are expected to have more functional impact on diseases.

Finally, we compared the imputation possibilities of the 1000G and GoNL reference panels, with our Iberian-GCAT reference panel. We observed that the Iberian-GCAT reference panel outperformed the imputation of high-quality SVs by 2.7 and 1.6-fold compared to 1000G and GoNL, respectively. Also, the overall imputation quality is higher, showing the value of this new resource in GWAS as it includes more SVs than previous reference panels. The combination of different reference panels will improve the resolution and statistical power of GWAS, thus increasing the chances to find more risk variants in complex diseases, and ultimately, translated this insight to precision medicine.

Abbreviations and acronyms

1000G: 1000 Genomes

1KJPN: Japanese population reference panel

ACMG: American College of Medical Genetics and Genomics

AJ: Ashkenazi Jewish

ALT: Alternative Allele

AS: *De Novo* Assembly

BAM: Binary Alignment File

BND: Break-end

bp: Base Pair

BQSR: Base Quality Score Recalibration

BSC: Barcelona Supercomputing Center

BWA: Burrows-Wheeler Aligner

CDS: Coding Sequencing regions

CGH array: Array Comparative Genomic Hybridization Array

CI: Confidence Interval

CIGAR: Concise Idiosyncratic Gapped Alignment Report

chrY: Chromosome Y

CNV: Copy Number Variation

ddNTPs: Dideoxynucleotides

ENCODE: The Encyclopedia of DNA Elements Project

DEL: Large Deletions

DGV: Database of Genomic Variants

dNTPs: Deoxyribonucleotides

DR: Discordant reads

DUP: Duplications

EMBL-EBI: European Molecular Biology Laboratory-European Bioinformatics Institute

ERRBKP: Breakpoint-error

FDR: False-Discovery Rate

FN: False-Negative

FoSTeS: Fork Stalling and Template Switching

FP: False-Positive

GA4GH: Global Alliance for Genomics and Health

GATK: Genome Analysis Toolkit

GCAT: GCAT-Genomics for life

GIAB: Genome In A Bottle

GL: Genotype likelihood

GoNL: Genome of the Netherlands

GRC: Genome Reference Consortium

GSA: Genome Sequencing and Analysis

GTE_x: Genotype-Tissue expression

GWAS: Genome-Wide Association Studies

HGP: Human Genome Project

HGSVC: Human Genome Structural Variant Consortium

HI: Haploinsufficiency

HMM: Hidden Markov Model

HPRC: Human Pangenome Reference Consortium

HRC: Haplotype Reference Consortium

HWE: Hardy-Weinberg Equilibrium

IBD: Identity by Descent

IBS: Iberian ethnicity

IHGSC: International Human Genome Sequencing Consortium

Indels: Small Insertions and Deletions (size 1 to 30 bp)

INS: Insertions

INV: Inversions

LCRs: Low-copy repeats

LD: Linkage Disequilibrium

LRM: Logistic Regression Model

LTR: Long Terminal Repeat

MAF: Minor Allele Frequency

mCNV: Multiple Copy Number Variants

mtDNA: Mitochondrial chromosome

MEI: Mobile Element Insertion

ML: Machine Learning

MMBIR: Microhomology-Mediated Break-Induced Replication

MMEJ: Microhomology-Mediated End Joining

MNP: Multi-nucleotide Polymorphism

MNV: Multiple Nucleotide Variant

NAHR: Non-Allelic Homologous Recombination

NCBI: National Center for Biotechnology Information

NGS: Next-Generation Sequencing

NH: Non-Homologous

NHEJ: Non-Homologous End Joining

OEA: One End Anchored

OMIM: Online Mendelian Inheritance in Man

ORF: Open Reading Frame

PacBio: Pacific Biosciences

PANCAN: PanCancer Project

PAR: Pseudo-Autosomal Regions

PCA: Principal Component Analysis

PCR: Polymerase Chain Reaction

PIRs: Phase Informative Reads

PL: Phred-scaled likelihoods

pLOF: Predicted loss-of-function

pLI: Predicted loss-of-function intolerance

POPRES: Population Reference Sample

POPVAR: Population Variation

PSG: Pseudogenes

QC: Quality Control

RD: Read-depth

REF: Reference Allele

RG: Reference Genome
RO: Reciprocal-overlap
SAM: Sequence Alignment Map
SD: Segmental Duplication
SMRT: Single Molecule Real-Time
SNP: Single Nucleotide Polymorphism
SNV: Single Nucleotide Variants
SR: Split-reads
SV: Large Structural Variants
SVM: Super Vector Machine
TAD: Topologically Associating Domains
TGS: Third-Generation Sequencing
TopMED: The Trans-Omics for Precision Medicine program
TP: True-Positive
TRA: Translocations
TRP: Transposons
UCSC: University of California and Santa Cruz
UK10K: The UK10K project
UTR: Untranslated regions
VCF: Variant Calling Format
VIR: Viruses
VQSR: Variant Quality Score Recalibration
WGS: Whole Genome Sequencing

Thesis trajectory

Before exposing my thesis's content, I wanted to provide an overview of the trajectory and working environment of this PhD. I started this PhD with limited knowledge of bioinformatics, because I specialised in Ecology during my degree. Thus, during my first year, I invested most of my time learning genomics in general, as well as entering into programming with Bash, Python and R language. During this time, I have been involved in the study of transposon activity in cancer samples, within other ongoing projects in the group. First, I developed an *in-silico* sample genome for benchmarking, which enabled me to understand and apply my insights on python programming. Then, this *in-silico* sample was used to evaluate the benchmarking of variant callers. During this period, most of my groupmates, such as Alex Barberà and Montse Puiggròs, helped me with this learning curve, which was key for my real thesis project.

During my second year, I started my thesis project: The generation of the Iberian-GCAT haplotype reference panel using 785 WGS samples from GCAT biobank. The first part consisted of finding and collecting the methodology for analysis. Because little information about the different programs were available, I contacted with developers of these tools to try to learn how to best use them, or how to use them at all. This activity has helped me to understand and to use the existing information and the scientific environment. During this time, Iván Galván Femenía and Daniel Matías Sánchez incorporated into the project, contributing centrally in different parts of the study. Iván has been centred in the design of different Logistic Regression Models, with support for sample filtering and evaluation of the Iberian-GCAT reference panel's performance at imputation level. Dani has been mainly involved in the validation part, by comparing results with other panels and by analysing array validation data. Overall, without the Ivan Galván and Dani Matías collaboration, this project could not have been done. Besides, Jon Lerga-Jaso has also been involved in the validation of the inversions, and Montse Puiggròs helped with technical support and data managing issues when working with the Supercomputing MareNostrum4.

In conclusion, this thesis is the result of the efforts of many people, where under the supervision of David Torrents, I have been able to finish this thesis, obtaining the first Iberian reference panel.

Table of Contents

| | |
|---|-----------|
| SUMMARY | 4 |
| ABBREVIATIONS AND ACRONYMS | 6 |
| THESIS TRAJECTORY | 10 |
| TABLE OF CONTENTS | 11 |
| 1. INTRODUCTION | 16 |
| 1.1 THE GENETIC VARIABILITY CONTRIBUTION TO PRECISION MEDICINE | 17 |
| 1.1.1. <i>The importance of precision medicine in the healthcare system</i> | 17 |
| 1.1.2. <i>Genetic variability across populations</i> | 17 |
| 1.1.3. <i>The Human Genome Project</i> | 20 |
| 1.2 THE IDENTIFICATION AND CHARACTERISATION OF HUMAN GENOME VARIABILITY | 22 |
| 1.2.1. <i>The landscape of germline variation</i> | 23 |
| 1.2.2. <i>The generation of genomic data for variant detection</i> | 24 |
| 1.2.2.1. The array technology: SNP and Comparative Genomic Hybridisation arrays (CGH array) 24 | |
| 1.2.2.2. DNA sequencing technologies | 25 |
| 1.2.2.3. Read alignment using the reference genome | 27 |
| 1.2.3. <i>Variant calling in the sequencing era</i> | 28 |
| 1.2.3.1. The identification of SNV and small indels | 29 |
| 1.2.3.2. Structural Variant detection | 30 |
| 1.2.4. <i>Genotype the variants using sequencing reads</i> | 32 |
| 1.2.5. <i>Benchmarking variant identification of variant callers</i> | 33 |
| 1.2.6. <i>Variant caller integration: Improving the accuracy of variant detection</i> | 35 |
| 1.2.7. <i>The correlation of variants across the human genome</i> | 37 |
| 1.2.8. <i>The functional impact of variants on the human genome</i> | 38 |
| 1.3 GENETIC VARIABILITY PANELS (REFERENCE PANELS): AN INVALUABLE RESOURCE IN GENOME-WIDE ASSOCIATION STUDIES (GWAS) | 40 |
| 1.3.1. <i>The role of phasing in reference panel creation</i> | 43 |
| 1.3.2. <i>Genotype imputation in GWAS studies</i> | 44 |
| 1.4 THE RATIONALE OF THIS THESIS | 46 |
| 2. OBJECTIVES | 49 |
| 3. MATERIAL AND METHODS | 51 |
| 3.1 CREATION OF <i>IN-SILICO</i> SAMPLE | 52 |
| 3.1.1. <i>In-silico sample description</i> | 52 |
| 3.1.2. <i>Procedure to insert variants into the reference genome and create the in-silico sample</i> 56 | |
| 3.2 BENCHMARKING OF DIFFERENT VARIANT CALLERS | 57 |
| 3.2.1. <i>SNV and Indel calling</i> | 58 |
| 3.2.1.1. Haplotype Caller (GATK4 version 4.0.2.0) | 58 |
| 3.2.1.2. Deepvariant (version 0.6.1) | 58 |
| 3.2.1.3. Strelka2 (version 2.9.2) | 58 |

| | | |
|-----------|---|----|
| 3.2.1.4. | Platypus (version 0.8.1) | 58 |
| 3.2.1.5. | VarScan2 (version 2.4.3) | 59 |
| 3.2.2. | <i>Large Structural Variant (SV) calling</i> | 59 |
| 3.2.2.1. | Delly2 (version 0.7.7) | 59 |
| 3.2.2.2. | Manta (version 1.2) | 60 |
| 3.2.2.3. | Pindel (version 0.2.5b9) | 60 |
| 3.2.2.4. | Lumpy (version 0.2.13) | 60 |
| 3.2.2.5. | Whamg (version v1.7.0-311-g4e8c) | 61 |
| 3.2.2.6. | SvABA (version 7.0.2) | 61 |
| 3.2.2.7. | CNVnator (version v0.3.3) | 61 |
| 3.2.2.8. | Popins (version damp_v1-151-g4010f61) | 61 |
| 3.2.2.9. | MELT (version 2.1.4) | 61 |
| 3.2.2.10. | ViFi (no version reported) | 61 |
| 3.2.2.11. | VERSE (VirusFinder2) (version 2.0) | 62 |
| 3.2.2.12. | Genome Strip (Version 2.0) | 62 |
| 3.2.2.13. | Pamir (version 1.2.2) | 63 |
| 3.2.2.14. | AsmVar (version 2.0) | 63 |
| 3.2.3. | <i>Recall, Precision, and F-score</i> | 63 |
| 3.2.3.1. | Evaluation of breakpoint-error for Indels | 63 |
| 3.2.3.2. | The categorisation of <i>in-silico</i> variants | 64 |
| 3.2.3.3. | Determination of SV breakpoint-error | 65 |
| 3.2.3.4. | Evaluation of variant caller metrics | 65 |
| 3.2.3.5. | Evaluation of genotype errors | 66 |
| 3.2.3.6. | Selection of a strategy to construct all the BAM files of GCAT samples | 67 |
| 3.3 | GENOME IN A BOTTLE SAMPLE | 67 |
| 3.4 | INCREASING ACCURACY DETECTION USING A MACHINE LEARNING ALGORITHM | 68 |
| 3.4.1. | <i>Logistic Regression Model for SNVs and small Indels</i> | 68 |
| 3.4.1.1. | Training and Testing of the model | 68 |
| 3.4.1.2. | Genotype reported by LRM for SNVs and small Indels | 68 |
| 3.4.2. | <i>Logistic Regression Model for SVs</i> | 68 |
| 3.4.2.1. | Training and Testing of the LRM for SVs | 68 |
| 3.4.2.3. | Genotype reported by LRM for SVs | 70 |
| 3.4.2.4. | Filtering out the SVs following the GoNL strategy | 71 |
| 3.5 | THE GCAT PROJECT | 71 |
| 3.5.1. | <i>Genomic data features</i> | 71 |
| 3.5.2. | <i>BAM file generation</i> | 72 |
| 3.5.2.1. | Selection of the Reference Genome, based on the sample gender | 72 |
| 3.5.2.2. | Data structure of the WGS samples of the GCAT project | 72 |
| 3.5.2.3. | The construction of BAM files in the GCAT project | 73 |
| 3.5.3. | <i>Quality Control (QC) of the BAM files and sample ancestry analyses</i> | 73 |
| 3.5.3.1. | Alignment quality | 73 |
| 3.5.3.2. | Contamination analysis | 75 |
| 3.5.3.3. | Population structure using reference ancestries | 75 |
| 3.5.3.4. | Identity by Descent analysis (IBD) | 76 |
| 3.5.4. | <i>The impact coverage on SV calling</i> | 76 |
| 3.6 | VARIANT CALLING IN THE GCAT SAMPLES | 76 |
| 3.6.1. | SNV AND INDEL CALLING | 76 |

| | | |
|-----------|---|-----------|
| 3.6.1.2. | Strelka2..... | 77 |
| 3.6.1.3. | Deepvariant..... | 78 |
| 3.6.1.4. | SNV/Indel VCF Normalisation | 78 |
| 3.6.2. | <i>Mid-size and Structural Variant calling</i> | 78 |
| 3.6.2.1. | Delly2..... | 78 |
| 3.6.2.2. | Manta..... | 79 |
| 3.6.2.3. | Pindel..... | 79 |
| 3.6.2.4. | Lumpy | 79 |
| 3.6.2.5. | SvABA..... | 80 |
| 3.6.2.6. | Whamg..... | 80 |
| 3.6.2.7. | CNVnator | 80 |
| 3.6.2.8. | Popins..... | 80 |
| 3.6.2.9. | Melt | 80 |
| 3.7 | VARIANT CALLING INTEGRATION | 80 |
| 3.7.1. | <i>VCF pre-processing</i> | 80 |
| 3.7.2. | <i>Merging all VCFs per sample and per variant type</i> | 81 |
| 3.7.3. | <i>Combining all samples in a single VCF</i> | 81 |
| 3.7.4. | <i>Variant Quality Control</i> | 81 |
| 3.8 | NEW DISCOVERIES AND VALIDATION OF THE GCAT VARIANTS | 82 |
| 3.8.1. | <i>Comparative studies with different datasets</i> | 82 |
| 3.8.1.1. | SNVs and indels | 82 |
| 3.8.1.2. | Structural Variants | 82 |
| 3.8.2. | <i>Experimental validations</i> | 82 |
| 3.8.2.1. | Validation of SNVs and indels using the GCAT SNP-array..... | 82 |
| 3.8.2.3. | Inversions validation using a verified dataset | 83 |
| 3.9 | CREATION AND INTEGRATION OF HAPLOTYPES SETS | 83 |
| 3.9.1. | <i>Benchmarking of different phasing strategies</i> | 84 |
| 3.9.1.1. | Sample pre-processing..... | 84 |
| 3.9.1.2. | Phasing strategies | 84 |
| 3.9.1.3. | Imputation using different phasing strategies | 86 |
| 3.9.1.4. | Selecting the strategy to phase the GCAT samples..... | 87 |
| 3.9.2. | <i>Pipeline to construct the Iberian-GCAT haplotype panel</i> | 87 |
| 3.10 | IMPUTATION USING THE IBERIAN-GCAT REFERENCE PANEL | 87 |
| 3.10.1. | <i>Imputation analyses using the GCAT SNP-genotyping array</i> | 88 |
| 3.10.2. | <i>Imputation quality using the array data of 1000G</i> | 88 |
| 3.10.2.1. | Sample filtering and Quality control of the 1000G array | 88 |
| 3.10.2.2. | Imputation of non-Iberian samples with the Iberian-GCAT panel | 89 |
| 3.11 | BENCHMARKING THE REFERENCE PANELS..... | 90 |
| 3.12 | BIOLOGICAL IMPACT OF STRUCTURAL VARIANTS..... | 91 |
| 3.12.1. | <i>Structural Variant distribution in the worldwide populations</i> | 91 |
| 3.12.2. | <i>Functional impact of Structural Variants</i> | 92 |
| 3.12.2.1. | Structural variant annotation using AnnotSV | 92 |
| 3.12.2.2. | Evaluation of SVs using the GWAS catalog..... | 93 |
| 4. | RESULTS | 95 |
| 4.1 | IDENTIFICATION AND CLASSIFICATION OF VARIANT CALLERS USING THE GENOME IN A BOTTLE (GIAB) AND <i>IN-SILICO</i> SAMPLES | 96 |

| | | |
|-----------|---|------------|
| 4.1.1. | <i>The software selected to perform the variant detection in GCAT samples ...</i> | 97 |
| 4.1.2. | <i>Variant classification according to size</i> | 98 |
| 4.1.3. | <i>Benchmarking of variant callers and the Logistic Regression Model (LRM) .</i> | 99 |
| 4.1.3.1. | <i>Benchmark analyses of SNVs and small Indels</i> | 100 |
| 4.1.3.2. | <i>Measuring the breakpoint-error of each variant caller in SV discovery</i> | 102 |
| 4.1.3.3. | <i>Accuracy of detection of SVs by size.....</i> | 102 |
| 4.1.3.4. | <i>Benchmarking analyses of SVs between variant callers, GoNL strategy and Logistic Regression Model.....</i> | 104 |
| 4.1.3.5. | <i>Benchmarking of genotyping between variant callers and Logistic Regression Model.....</i> | 105 |
| 4.1.3.6. | <i>Evaluation of the strategy used to generate the BAM files of the GCAT samples</i> | 106 |
| 4.2 | CHARACTERISATION OF THE GCAT SAMPLES..... | 108 |
| 4.2.1. | <i>Filtering of GCAT samples and features of BAM files</i> | 108 |
| 4.2.1.1. | <i>Filtering the non-Iberian representative samples.....</i> | 109 |
| 4.2.1.2. | <i>Importance of coverage in variant detection.....</i> | 110 |
| 4.2.1.3. | <i>Improving variant detection in chromosome X.....</i> | 111 |
| 4.2.2. | <i>A general description of the variants recovered after applying the merge strategy and the Logistic regression model.....</i> | 112 |
| 4.2.3. | <i>A detailed description and characterisation of Structural Variants into the Iberian cohort.....</i> | 114 |
| 4.2.3.1. | <i>Comparison of variants detected against different repositories.....</i> | 115 |
| 4.2.3.2. | <i>Variant size ranges detected by our methodology.....</i> | 116 |
| 4.2.3.3. | <i>Structural Variant distribution in the Iberian cohort.....</i> | 117 |
| 4.2.3.4. | <i>Functional impact of Structural Variants</i> | 118 |
| 4.2.4. | <i>Validation of Iberian dataset</i> | 123 |
| 4.2.4.1. | <i>Validation of SNVs and indels using the GCAT genotyping array data</i> | 123 |
| 4.2.4.2. | <i>Experimental validations of Structural Variants</i> | 124 |
| 4.3 | A HAPLOTYPE-RESOLVED PANEL OF THE IBERIAN COHORT | 126 |
| 4.3.1. | <i>Evaluating different phasing strategies to create the Iberian-GCAT reference panel</i> | 127 |
| 4.3.2. | <i>Imputation performance using the Iberian-GCAT reference panel.....</i> | 128 |
| 4.3.2.1. | <i>Evaluation of imputation on the GCAT genotyping array.....</i> | 128 |
| 4.3.2.2. | <i>Evaluation of imputation on the 1000G genotyping array.....</i> | 131 |
| 4.3.2.3. | <i>Structural variant worldwide distribution</i> | 133 |
| 4.3.3. | <i>Comparing imputation performance of multiple reference panels</i> | 135 |
| 5. | DISCUSSION | 138 |
| 5.1 | VARIANT CALLER BENCHMARKING..... | 139 |
| 5.1.1. | <i>Benchmarking of SNVs and indels</i> | 140 |
| 5.1.2. | <i>Benchmarking mid Deletions and large Structural Variants</i> | 141 |
| 5.2 | PROCESSING 808 WHOLE-GENOME SEQUENCING SAMPLES FROM GCAT BIOBANK .. | 143 |
| 5.3 | THE IBERIAN-GCAT CATALOGUE DESCRIPTION | 144 |
| 5.3.1. | SNV AND INDEL DESCRIPTION IN THE IBERIAN-GCAT CATALOGUE..... | 144 |
| 5.3.2. | STRUCTURAL VARIANT DESCRIPTION IN THE IBERIAN CATALOGUE | 145 |
| 5.3.3. | THE IMPACT OF THE STRUCTURAL VARIANTS ON HUMAN TRAITS | 148 |
| 5.4 | THE IBERIAN-GCAT HAPLOTYPE PANEL..... | 150 |

| | | |
|-----------|--|------------|
| 5.4.1. | <i>Performance of the Iberian-GCAT haplotype reference panel</i> | 150 |
| 5.4.1.1. | Imputation performance using non-Iberian samples..... | 151 |
| 5.4.2. | <i>Benchmarking of multiple haplotype reference panels</i> | 153 |
| 6. | CONCLUSIONS | 155 |
| 7. | SUPPLEMENTARY MATERIAL | 157 |
| 8. | REFERENCES | 169 |

1. INTRODUCTION

1.1 The genetic variability contribution to precision medicine

1.1.1. The importance of precision medicine in the healthcare system

The human life expectancy has increased over the years, being in 2,020 on average of 72.63 years (Figure 1A), thanks to the advances in science, medicine and technology and the promotion of healthy lifestyles. Living longer lives also implies that age-related diseases, such as cancer, heart failure, and other complex and degenerative diseases, will increase¹², making it unsustainable to maintain the current healthcare system with continuously growing elderly populations. Estimations point to a 29% increase of European people at ages of 65 years in 2,070¹³. For this reason, particular attention has to be focused on identifying ways of predicting and understanding these diseases, ultimately generating new prognosis and treatment protocols, and lowering the personal and economic burden of complex diseases in developed countries.

In this context, precision medicine is opening new insights on disease prevention¹², helping to reduce the treatment costs without losing quality. Current medicine relies on treating the diseases after their clinical intervention, being expensive and hard to maintain by healthcare systems. For this reason, a transition to modern medicine by using genomic, metabolomic, proteomic, and epigenomic data is of paramount interest in order to prevent, delay, or predict disease offset and development^{12,14}. In the last decade, the interest in precision medicine on humans has increased exponentially in the scientific community, publishing in 2,019 nearly 5,000 articles on this field (Figure 1B). In Spain, this interest rose, publishing 255 articles in the same year. However, in 2,020, the publications dropped dramatically during the COVID-19 pandemic, truncating this progress (Figure 1B). A key concept in precision medicine is the individual's genetic background, which can increase or decrease the relative risk to develop particular diseases, improving the diagnosis, and treatments¹⁵. For this reason, understanding the relationship between genetic variability and phenotype is one of the central goals in molecular biology and biomedicine.

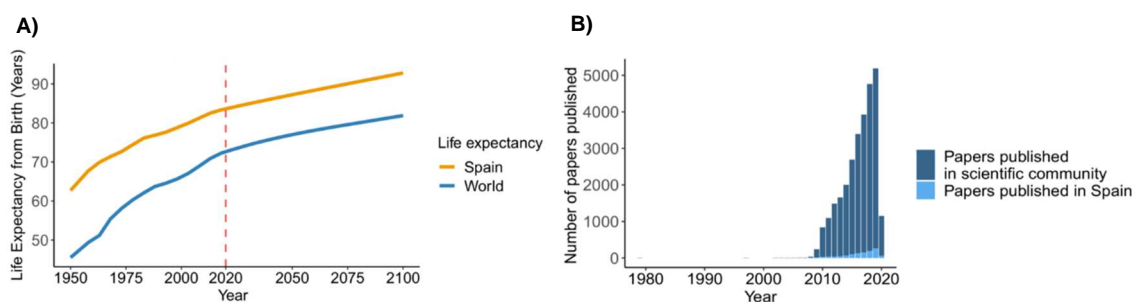


Figure 1. Evolution of life expectancy and Precision medicine. A) Life expectancy evaluation across the years. Data from <https://www.macrotrends.net>. **B)** Papers published about precision medicine in humans. Data from Pubmed/MEDLINE, using the topics “precision medicine”, “human” and “Spain” filters.

1.1.2. Genetic variability across populations

The human genome is a diploid organism composed of 3.2 billion nucleotides, codified by four bases A, T, C, and G (Adenine, Thymine, Cytosine, and Guanine). Each position is determined as a locus, and each form of this locus is referred to as an allele. The genome between two individuals is not identical due to aberrant rearrangements produced in germ cell line lineage, contributing to population genetic variability. This variability is usually classified according to their

sizes: Single Nucleotide Variants (SNVs) which are the exchange of one nucleotide by another, small Insertions and Deletions (<50bp) commonly referred as indels, and larger aberrant rearrangements referred as Structural Variants (SVs) (normally larger than 50 bp). The majority of SNVs and indels derive from DNA replication errors^{16,17}, due to incorporating an incorrect base (SNVs), generating mutations at 1.16×10^{-8} per site per generation¹⁶. The slippage of the DNA polymerase, with a rate of 0.20×10^{-9} per site per generation¹⁶, is the origin of indels.

On the other hand, SVs are generated through different mechanisms¹⁸: 1) Errors in DNA recombination, such as Non-Allelic Homologous Recombination (NAHR) in meiosis or mitosis, where highly-homologous sequences (10Kb or higher with more than 95% of homology) are misaligned from different genome regions, producing SVs such as deletions, duplications, inversions, and translocations. For this reason, some SVs are located in highly repetitive regions. 2) Errors in the DNA repair mechanisms, such as Non-Homologous End Joining (NHEJ) or Microhomology-Mediated End Joining (MMEJ). The NHEJ is the DNA repair mechanism preferred in mammals, which does not require homologous sequences between break ends to fuse the double-strand breaks (DSBs), generating short insertions and deletions in the breakpoint junction. The difference between NHEJ and MMEJ repair methods mainly resides in microhomology sequences' usability to repair the DNA. MMEJ generates more SVs than NHEJ, such as deletions and translocations. Finally, the other mechanism which produces SVs is related to 3) Errors in DNA replication, such as Fork Stalling and Template Switching (FoSTeS) or Microhomology-Mediated Break-Induced Replication (MMBIR). During the replication, the active polymerase is stalled and switches templates by microhomology of another active replication fork. This polymerase switch can affect from a few kilobases to megabases, generating complex rearrangements, as well as inversions, tandem duplications and translocations.

Although most of these variants are expected to be functionally silent, a fraction of them could affect some traits, such as eye, hair, or skin colour, and even be related to disease, such as cancer, diabetes, neurodegenerative, heart diseases, among others. For this reason, the identification of the genetic variation behind diseases, as well as the interpretation of its functional impact within each pathology, is key to identify markers to improve diagnosis and treatment protocols.

1.1.2.1. The heritability of variants and traits

Since Gregor Johann Mendel (1822-1884) published in 1866, his study of pea plants (*Pisum sativum*) "Experiments on Plant Hybridization" and postulated "*Mendelian Laws of Inheritance*," where specific traits are inherited to offspring following certain rules, much research has been done. In 1936, J.L.Lush used the term "heritability" to formally describe the proportion of variation in a particular trait attributable to genetic factors¹⁹. For this reason, following the genetic inheritance rules, the offsprings tend to be phenotypically similar to progenitors (mostly driven by additive genetic variance), where each half of parents genetic material is passed to the offspring, acquiring their genetic variability (germinal variants) in 23 pairs of chromosomes and the maternal mitochondrial chromosome (mtDNA).

In this direction, as the genetic variability is transmitted, the variants associated with diseases (i.e., complex diseases) could be passed across the generations and expanded in populations, making it crucial to understand the genetic background of populations in order to establish the bases of precision medicine. In addition, the environmental factors also contribute to human traits; for this reason, the phenotype has to be considered as a combination of genotype and

environment factors, where the heritability allows to compare the genetic contribution in different human traits¹⁹.

1.1.2.2. Population genetic variability

Variant distribution is not equal across all populations. The African ancestry populations are the groups with the highest number of variant sites^{1,11,20}, coincident with the out-of-Africa model of human origin, and additionally, they conserve genetic substructures correlated to geography, language, and culture²⁰. When humans spread within Africa and subsequently over the world, the adaptation to new environmental conditions and the local pathogenic environment, produced a selective pressure on individuals^{20,21}. Additionally, new cultural innovations, such as animal domestication and fishing, helped to expand and establish humans in different world regions (founding events), losing genetic variability compared to those in Africa. These migrations and posterior environmental adaptations directly affect genome variability, where selective pressures stratified the variants across populations^{20,21}. For example, the skin colour is an adaptation to sun exposure and is expressed according to the geographical region²⁰. The dark skin protects from ultraviolet rays in African populations, in contrast to populations established at high latitudes where the sun exposure is lower, selecting variants favouring lighting the skin²⁰. In addition, the effect of some population-specific variants can vary across populations, such as a deletion detected in American populations, which removes an exon of the *MS4A1* gene and is associated with lymphoma, leukaemia, and autoimmune treatment response²¹, or a duplication in the *HCAR2* gene present in Asians, which has been proposed as a therapeutic target for mediating anti-inflammatory effects in diseases²¹. These particularities indicate that one treatment could be effective in one ethnic group and not in others, demonstrating the relevance of populations' genetic background in precision medicine.

In this direction, Sirugo et al.⁶ highlighted the necessity to study the genetic variability of different populations because the under-representation of different ethnic groups impedes the understanding of the genetic architecture of human diseases fully. For example, the allele frequencies of structural variants generated by NAHR fluctuate across populations¹⁶, limiting the complete understanding of genetic variability and diseases.

The genetic particularities across ethnic groups could tag the causative variants of a disease differently, highlighting the necessity to include different ethnic groups in association tests⁶. Nowadays, predominantly European ancestry populations are used to characterise human genetic diseases. This factor could drive to misinterpretations in identifying the genetic risk factors in rare or complex diseases in non-European populations due to incomplete genetic information. Additionally, different populations can have different responses to treatments. For example, the warfarin dosage varies considerably between patients due to variants in *CYP2C9*, *VKORC1*, *CYP2C*, and *CYP2C* genes. In Europeans, these variances explained up to 30% of warfarin's metabolism, but in African descendants, the same variants explained less variance⁶.

In summary, most of the genetic differences between individuals and populations are a combination of genetic rearrangements and adaptive pressures to the environment. The heritability plays an important role in transmitting the variants across generations, maintaining the background genetic in populations^{19,20}. The majority of these variants are neutral and do not affect human traits or diseases²². However, some variants could be related to diseases susceptibility, where the genetic variability of populations plays a key role in diagnostic or treatment efficiency^{6,7}. For this reason, studying the genetic background of populations will be determinant in precision

medicine, allowing to understand the genetic architecture of diseases and determine the best prevention, diagnostic or treatment for each patient, and ultimately improve the healthcare system.

1.1.3. The Human Genome Project

The irruption of the Human Genome Project (HGP) has been crucial in biomedicine, as it has established a new paradigm of research, based on large scale genome sequencing and analysis. The possibility of understanding the landscape of our genome and having a clear functional map of the different regions, allows translating this knowledge into medical decisions, which constitutes the basis for personalised medicine. The first complete version of the reference genome (NCBI Build 34 or UCSC hg16) was presented in 2003 by the International Human Genome Sequencing Consortium (IHGSC), after an investment of around ~\$450 million (<https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost>). This version was obtained by sequencing several volunteers with hierarchical shotgun sequencing, enabling the decodification of the human genome, given that more than 50% of the genome consist of repeat sequences. The following release in 2004 (NCBI Build 35 or UCSC hg17) contained 2.85 billion nucleotides interrupted by 341 gaps, covering 99% of the euchromatic genome and predicting around 20,00-25,000 protein-coding genes²³.

Since then, three new versions have been generated, improving and completing further the previous version, mostly in repetitive regions such as segmental duplications (SD), centromeres, and telomeres. In 2009 the Genome Reference Consortium (GRC) launched the version GRCh37 (UCSC hg19), derived from 13 people²⁴, at the same time that the Illumina high throughput sequencing technology started to be used in biomedical research²⁵, including 3.32 billion nucleotides, 20,805 protein-coding genes, 22,966 non-protein-coding genes, and 14,181 pseudogenes (https://grch37.ensembl.org/Homo_sapiens/Info/Annotation), and only 250 gaps. As a result, the GRCh37 is a mosaic haploid genome, where the common alleles were included in the consensus reference genome²⁴. Usually, the diploidy of the human genome was not represented in the reference sequence until GRCh37, which included nine alternate loci in six haplotypes on the MHC region of chromosome 6, better representing the extent of structural variation and population genomic diversity in the locus²⁶ (<https://grch37.ensembl.org/info/genome/genebuild/assembly.html>).

Additionally, the GRCh37 reference genome has different versions, according to the additional sequences included, which can be used to reduce the sequence misalignments. In this context, the hs37d5 reference genome generated by 1000 Genomes Phase II is the most used, including GRCh37 primary assembly, the rCRS mitochondrial sequence, Human herpesvirus 4 type1, and decoy sequences⁵. Decoy sequences are repetitive regions, of which 50% are satellite or tandem repeats, and 23% are interspersed repeats (SINE/LINE/Long Terminal Repeats (LTR)); it includes BAC/fosmid clones, HuRef contigs, NA12878 ALLPATH-LG assembly, and Epstein-Barr Virus genome. Thus, the inclusion of the decoy sequences allows for a better SNV discovery and SV as well⁵.

Finally, the GRCh38.p13 (UCSC hg38) is the newest reference genome version constructed from many donors in 2,013. This version modified the genome coordinates since 2009²⁶, hindering making it difficult to compare variant coordinates between projects. The GRC determined that this version is the most complete and accurate compared to GRCh37, correcting different

misassembled regions, decoding sequences from centromeres and telomeres, filling in gaps, and including more diversity in the reference genome, with 261 alternate loci across 178 regions^{25,26}.

Nowadays, hs37d5 is still in use, such as in the gnomAD¹¹ or Pan-cancer²⁷ projects. There are different reasons for the hesitation to switch to the latest reference build; for example, all tools used to detect genome variability are tested with this genome or the resistance to altering the working pipelines by researchers²⁵. Besides, changing the reference genome version implies additional work and efforts to update coordinates of previous projects. For example, liftOver tools such as liftOverPlink, allowed to switch the variants coordinates to reference genome required. However, liftOver tools had problems to convert the coordinates of GRCh38 to hs37d5, losing 5% of genetic information in SNVs²⁸. In addition, 1.5% of discrepancies between successfully converted variants and sequencing information of those variants were observed in SNVs, suggesting caution when converting genomic variants between assembly versions²⁸.

Despite these advances, all versions of the reference genome are still incomplete because they are generated using a few individuals, limiting the representation of genetic variability across populations^{21,29-31}. Nearly 10% of the total genome is not represented in the reference genome³². Besides, there is evidence of insertional sequences from human populations not represented in the reference genome (GRCh38)²⁹⁻³¹, with potential functional population-specific implications. For this reason, adding these sequences in the reference genome will improve the sequencing alignments and population-specific variant detection^{29,32}. In this direction, the Human Pangenome Reference Consortium (HPRC) (<https://humanpangenome.org/>), is borne with establishing a human reference genome, including the genetic variability of 350 individuals from different populations, in order to capture the whole genetic variability and generate a reference pan-genome. Thus, generating catalogues of genomic variation across populations and assembling them in a human pan-genome will be relevant in precision medicine, improving variant detection, and performing accurate treatments^{29,32}. However, technical challenges will emerge due to alternative loci of the reference pan-genome³³. The current tools, such as alignment or variant discovery programs, expect sequences to have a single location in a haploid assembly model³³, so new strategies will be necessary to facilitate the use of alternative loci in many bioinformatic tools.

1.1.3.1. Annotation of the human reference genome

The availability of a reference genome, helped the scientific community to annotate functional regions and to share these annotations with the community. A complete gene mapping in the human genome will facilitate the interpretation of genome variability on the human phenotype. The latest release from the UCSC refGene database harboured 19,412 protein-coding genes, representing 1-2% of the genome³⁴ and 11,579 non-protein-coding genes. However, ~90% of variants identified in disease association tests (Genow-Wide Association Studies (GWAS)) are located outside of protein-coding regions³⁵, such as regulatory or intergenic regions, hindering the understanding of their functional impact.

Besides, identifying gene regulatory elements, such as enhancers and promoters, is paramount of interest, due to their role in gene expression modulation. The alteration of these regulatory elements has been involved in several diseases^{36,37}, such as thalassemia, highlighting the importance of annotating these non-protein-coding regions to facilitate the functional interpretation of genome variability. It is estimated that more than 399,124 regions in the genome have enhancer-like features, and 70,292 regions have promoter-like features³⁸, more than the

number of genes. Thus, connecting the regulatory regions to their target genes is a challenging task. Generally, promoters are located between 1-2 Kb of the transcription start site³⁷. However, enhancers are located dozens of Kb away from genes, and can even influence multiple genes, making it difficult to determine the link to their target gene. For example, the GeneHancer³⁷ database integrated human known enhancers with associated genes, facilitating the enhancer functional interpretation.

On the other hand, nearly 50% of the human genome is covered by repetitive regions that are involved in numerous processes, such as the generation of rearrangements (ex: Structural Variants) and variation in general³⁹. The repetitive sequences can be classified as Mobile Element Insertions (MEIs), pseudogenes, tandem repeats (particularly in centromeres, telomeres, ribosomal gene clusters, and the short arm of acrocentric chromosomes), simple repeats, and low-copy repeats (LCRs) such as segmental duplications (SD), characterised by DNA sequences ≥ 1 Kb with 90-95% of identity in the reference haploid genome, constituting nearly 4-5% of the human genome³⁹.

In addition, many more annotations are included in the reference genome, such as DNA methylation, chromatin structure information, such as Topological Associating Domains (TAD), expression analyses, among others, showing the complexity and the high amount of data generated in human research. Different initiatives emerged to generate different genomic annotations, such as the Encyclopedia of DNA Elements Project (ENCODE)³⁸, Genotype-Tissue expression (GTEx)⁴⁰, Roadmap Epigenomics⁴¹ or GeneCards⁴², resulting in an invaluable resource for variant interpretation. Each of these databases has its particularities. The ENCODE project grouped all functional elements encoded in the human genome³⁸ or GTEx, provided relevant information about the effects of genetic variation on gene expression in multiple human tissues⁴⁰. Different web browsers, such as the UCSC Genome Browser⁴³ or EMBL-EBI ensemble, manage to display this amount of data efficiently, facilitating the observation and analysis of relevant genomic information, for example, the functional impact of variants in human traits.

1.2 The identification and characterisation of human genome variability

Genome variability can be classified into two main groups, depending on the cell type they affect: 1) Somatic variants accumulated in somatic (non-sexual) cells during life which are not transmitted to the offspring, and 2) Germline variants, occurring in germline (sexual) cells, which are passed onto the offspring, as part of their genetic background. In terms of disease implications, somatic variants are mostly related to the onset of non-familial tumours, covering around 80-90% of total cancer cases, while germline variants can be implied in rare and complex diseases such as Cystic fibrosis or diabetes, respectively. Rare diseases are associated with few low-frequency variants with a high penetrance⁴⁴, affecting mainly individuals and families. On the other hand, complex diseases are the result of multiple factors, and each genomic variant, generally with high prevalence in the population, contributes a small fraction to overall disease risk⁴⁴. Besides, the interaction between some germline and somatic variants and its effect on disease has also been described⁴⁵; for example, the penetrance of some germline variants in the BRCA1 gene increases breast cancer risk. For this reason, the characterisation of genetic background in humans by detecting risk variants (germline variants) could improve the diagnosis, prevention, and treatment of diseases.

1.2.1. The landscape of germline variation

As previously described, germline variants are catalogued according to their size: A) Single Nucleotide Variants (SNVs), which change one nucleotide by another, B) Insertions and Deletions smaller than < 50 base pairs (bp) (INDELs), and C) variants ≥ 50 bp, classified as Structural Variants (SVs)^{5,9,11} (Figure 2).

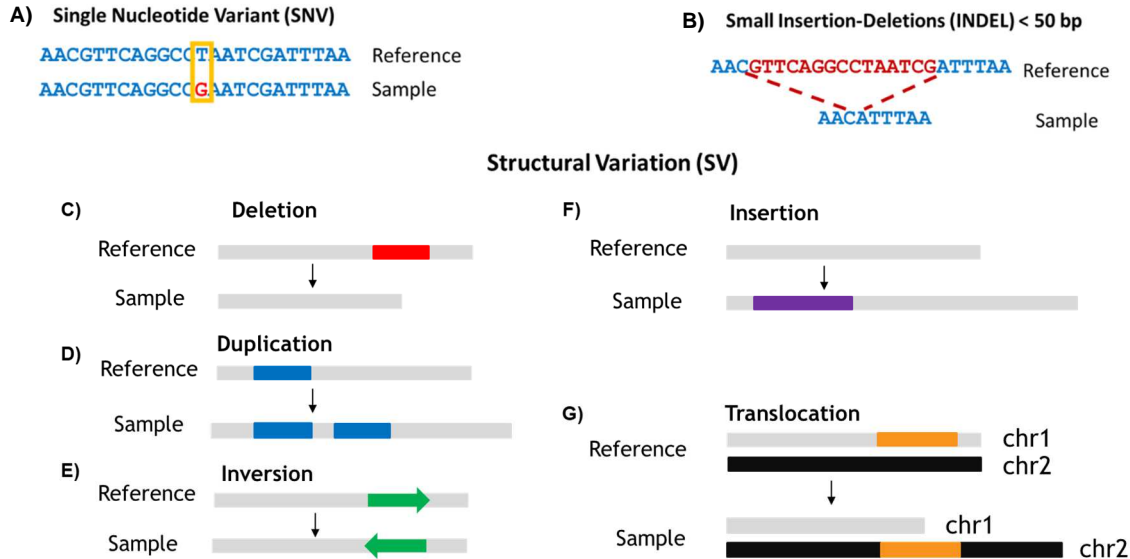


Figure 2. Classification of germline variants according to their size. A) Single Nucleotide variants. B) Small insertions and deletions (INDELs) < 50 bp. The structural variants (SVs) ≥ 50 bp are the largest genomic rearrangements in the genome. They can be unbalanced, such as C) Deletions, D) Duplications and F) Insertions, or balanced, such as E) Inversions and G) Translocations.

The SVs can be classified into two categories^{18,46}, 1) Unbalanced SVs if they modify the copy number of the genome, such as deletions (loss of genetic material) (Figure 2C), duplications (copy a variable number of DNA segments) (Figure 2D), which can be in tandem (segment inserted contiguous to the original one) or interspersed (segment inserted in elsewhere of the genome), and the insertion of novel sequences (Figure 2F). Then, 2) balanced SVs are those in which the genetic material is not altered, such as inversions (the DNA segment change the orientation) (Figure 2E) or translocations (exchange the genetic material between different chromosomes) (Figure 2G). Additionally, the complexity of SVs is increased if different SVs types are combined in a single event^{11,47,48}, such as inverted duplications, inversions flanked by duplications or deletions, insertion of new segments interposed at the breakpoints (genomic shards)⁴⁹, with chromothripsis (large chromosome segments shattered and imprecise repaired and combined) as one of the more extreme cases of complex SVs.

As previously mentioned, two human genomes are never equal. Considering SNVs, the genomic variation between genomes is around 0.1%, and the proportion increased to 1.5% when also considering SVs^{8,50}. For this reason, the allele frequencies of variants fluctuate across populations, providing a way to classify the genetic variability at the population level. In population genetics, when each allele is above 1%, it is considered as a polymorphic, cataloguing the SNV variants as Single Nucleotide Polymorphisms (SNPs). Besides, according to the minor allele frequency (MAF), the variants are classified as common (MAF $\geq 5\%$), low-frequency ($1\% \leq \text{MAF} < 5\%$), rare ($0.1\% < \text{MAF} \leq 1\%$), and when the allele is shared in two or one individuals, as doubletons and singletons respectively. The allele frequency provides relevant information about

populations' genetic architecture, where the low-frequency and rare variants could have arisen recently in populations, making them population-specific, or in the case of the natural selection affecting the variant negatively in a specific population region^{4,50,51}. Besides, low-frequency and rare variants could increase the risk of rare or complex diseases in specific populations^{6,7,52}, showing the necessity to improve each population's genetic characterisation individually.

1.2.2. The generation of genomic data for variant detection

The strategies to detect and classify genome variability have evolved during the past decades, gaining accuracy, sensitivity and decreasing the costs. Considering the variant detection as one of the final goals of genome analysis in biomedicine, there are two major strategies to approach and read the genome: 1) Genotyping arrays and 2) Sequencing technologies. The genotyping arrays provide a cheap, fast and direct way to characterise specific variants within large cohorts. Although the genotyping arrays are expected to capture major variability positions (through Linkage Disequilibrium (LD) (section 1.2.7)), they still miss a large fraction of variants, such as the low-frequency and rare variants, as well as SVs. This technology has been the basis for practically all the genetic studies, where combined with imputation, attempts to predict and populate the samples with inferred variants (section 1.3.2). Here we describe the strengths and weaknesses of each methodology in the context of genome analysis and variant detection and classification.

1.2.2.1. The array technology: SNP and Comparative Genomic Hybridisation arrays (CGH array)

The SNP array is the most popular and cheapest technology to detect genome variability across multiple samples. This technology consists of a solid support with hundreds of thousands of oligonucleotides (probes), where the DNA sample is hybridised with probes to detect SNPs or copy number variants (CNVs)^{53,54}. The array is then subjected to laser confocal scanning, where the intensities of fluorescence signal determine the variant detection. The main difference with the CGH array is the comparison of the problem and reference samples' intensities to detect the CNVs. When the probe intensity of the problem sample indicated a gain, the SV is a duplication; otherwise, it is a deletion.

An important characteristic of this technology is that researchers have to predefine the genomic positions and the type of variation to be interrogated. Different commercially available arrays are restricted to a limited preselected number of variants (in the order of 10K to 5M per array), mainly focused on common variants, limiting the detection of whole genome variability. Particularly, in SV detection, the SNP array data is restricted to large CNVs (>25 Kb) in non-repetitive regions, with a bias towards deletions, avoiding the identification of balanced SVs, such as translocations or inversions⁵⁴. Nevertheless, the SNP arrays are still in use, due to a reported > 99% genotype accuracy⁵⁵, and a price substantially cheaper than sequencing technologies, allowing the analysis of thousands of individuals, required for instance Genome-Wide Association Studies (GWAS).

HapMap project phase II constitutes an example, which genotyping with SNP array 270 samples, obtained 3.8M SNPs from different populations, collecting the common variants across different geographic zones⁵⁶. Despite their importance in GWAS studies, the HapMap project only included SNPs in their catalogue, overlooking the inclusion of indels and SVs. Thus, a different

technology is necessary to analyse the full spectrum of genome variability, to create a complete genetic variability map of humans, which would help to improve the GWAS resolution.

1.2.2.2. DNA sequencing technologies

The irruption of sequencing technologies to decodify the whole genome revolutionised biomedicine, allowing us to evaluate humans' whole-genome variability, including large genome rearrangements (SVs). This approach expands the possibilities of finding associations between genetic variants and diseases. However, computational and bioinformatic challenges, such as the large space requirements to store all genetic data or the software able to detect the genetic variability produced with this technology, brought new lines of research.

The sequencing methodologies have evolved over the years, resulting in faster, cheaper, and more accurate DNA sequencing. The first generation of DNA sequencing appears in 1977, where the Sanger sequencing methodology based on chain-termination technique improved all previously designed strategies⁵⁷. This technique used the analogues of deoxyribonucleotides (dNTPs) and the DNA polymerase enzyme to synthesise the DNA chain. The particularity of Dideoxynucleotides (ddNTPs) lacks 3'hydroxyl, necessary for elongation of DNA chain, besides are tagged by fluorescent dyes. The key feature was the inclusion of the dNTPs and ddNTPs in DNA synthesis. When the DNA polymerase incorporated ddNTPs, it inhibits the strand extension, obtaining DNA fragments at different lengths. Then, these fragments were then distributed by size using electrophoresis and revealed, resolving the DNA sequence.

The main limitation of Sanger sequencing was the amount of DNA sequenced. This method needed to amplify and sequence each DNA fragment individually, making it hard to detect the whole variants from a genome⁵⁸. In addition, the high costs of this method and its high manual labour requirements, limited its scalability to a high number of samples. However, Sanger sequencing is still being used, for example, to sequence single genes or for validation purposes.

A significant breakthrough was able to overcome the Sanger sequencing limitations, with the emergence of the high-throughput technologies named Next-Generation Sequencing (NGS). NGS could generate a high quantity of data in parallel from a small DNA amount. For example, the HiSeq4000 machine from Illumina could sequence six human genomes, producing 250-400 M of short sequences (reads) of 150 bp per run in three days. This improvement allowed the sequence of more genomes at lower costs, favouring and making possible the analysis of multiple genomes in different types of studies. This technology has also been key to improve genetic studies by exploring millions of variants throughout thousands of individuals, improving genetic variants' insights on diseases.

The first steps of NGS consisted of preparing the library, by randomly fragmenting the DNA into short fragments, selecting the size of the DNA fragment to be sequenced (insert size). Then, the insert size is anchored in solid support by adaptors and amplified through Polymerase chain reaction (PCR). The library properties are of interest because the tools which detect genome variability (variant callers) are influenced by coverage, insert size, and read length. Once the library is prepared, the sequencing reaction is performed into cycles in a flow cell and parallelised massively. The sequencing reaction terminates when the whole DNA fragment is read, generating millions of short sequences (reads)^{57,59}.

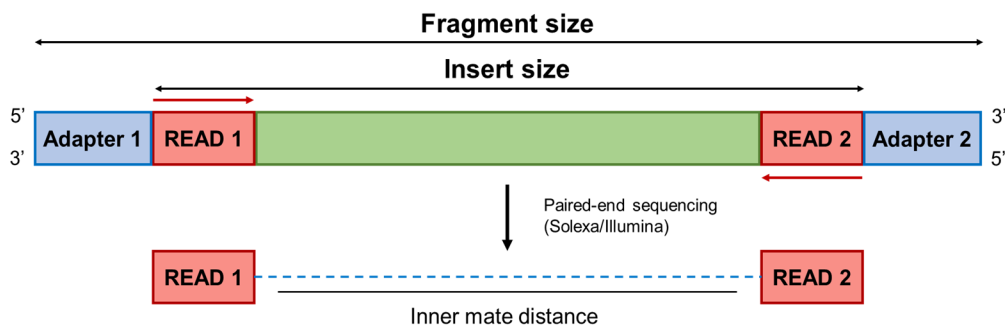


Figure 3. Structure of pair-end reads. The insert size is a single strand of DNA (ssDNA) selected to sequence. The fragment size includes the adapters, needed to anchor the insert size in the solid support. The paired-end reads are the terminal sequences of the insert size, which depending on the number of cycles, has a read length ranging from 100 to 300 bp. Read1 is sequenced in the forward strand, and their mate is synthesised in the reverse strand. The distance between reads is referred to as inner mate distance.

Moreover, the Solexa/Illumina platform had the advantage that it could sequence both ends of an insert size by generating paired-end reads (Figure 3). The paired-end reads information is crucial to detect structural variants and determine their type^{46,60}. For example, the distance between two reads, named "inner mate distance," allows detecting the structural variants such as deletions or insertions. When the inner mate distance is shorter than expected, the SV is an insertion, and the opposite is a deletion. For this reason, using NGS with paired-ends allows uncovering large genomic rearrangements. Nowadays, most software to detect SVs uses paired-end information to improve the performance of variant discoveries. However, NGS still has limitations, such as the length of the sequencing reads, from 100 to 300 bp, since the sequencing errors rise with increasing read length. For this reason, the detection of SVs using short reads is a challenging problem due to (i) the read mapping issues, which cannot detect SVs in repetitive regions, extreme cytosine-guanine (CG) content or homologous regions in the genome⁶¹, (ii) the sample coverage, which fluctuates across the genome, and (iii) chimeric molecules in the library preparation⁶².

Different projects, such as 1000G or GoNL, used NGS to sequence cohorts of individuals, by generating the first panels of haplotypes, including the genetic variability of a wide range of individuals. However, the low sequencing depth (coverage) of 7.4X and 14.5X challenges the detection of SV efficiently. Particularly, with higher coverage, more SVs can be detected, ameliorating the detection of heterozygous variants^{9,46,63,64}. Nowadays, sequencing at high coverage (30X) is cost-effective in large populations, enhancing the odds of SV discoveries. For example, there are new catalogues, such as Abel et al.⁶⁵ and gnomAD-SV¹¹, which detect SVs in large cohorts, improving SVs' insights in the human genome.

Finally, the Third-Generation Sequencing (TGS) approaches allowed an increase in the length of reads to 10-20 Kbps (long-reads), allowing the analysis of low complex genome regions, such as repetitive or segmental duplications, improving the discovery of SVs. The TGS leading platforms are Pacific Biosciences (PacBio), and Oxford Nanopore, which uses different approaches to generate reads with similar length. PacBio, for example, uses the Single Molecule Real-Time (SMRT) approach, where the ligation of both ends circularises the dsDNA fragment, generating the SMRTbell. Then, the SMRTbell is loaded in an SMRT cell, which contains a polymerase immobilised in a bottom, allowing to perform the DNA synthesis. On the other hand, the Oxford Nanopore platform uses a Nanopore sequencing approach. When the DNA fragment moves through the nanopore by a current channel, it produces changes in this current, which are

measured by a semiconductor sensor. Each nucleotide disrupts the electric field differently, allowing to codify of the DNA fragment⁵⁹.

The high expectations of both platforms to improve variant detection in low-complexity genome regions and SVs are promising. However, the costs and high sequencing error rate (1 error every 10 nucleotides) still limits the use of this technology at a large scale^{8,59}. For example, the Human Genome Structural Variant Consortium (HGSVC) sequenced 15 samples with this technology, demonstrating the capacities of long-reads on SV detection^{66,67}.

1.2.2.3. Read alignment using the reference genome

NGS and TGS sequencing technologies generate a high amount of genomic data, saving all reads in FASTQ files without a priori information of the position that they represent in the genome. The alignment or mapping tools of these sequences onto a human reference genome must be interpreted and evaluated in the right genomic context. BAM files included all reads aligned in the reference genome. The information of these alignments can be translated into genome variability, using complex bioinformatic methods that aim to identify differences within sequencing reads. For example, an SNV can be detected if a nucleotide position in the read does not match with the reference sequence, or an SV can be identified using the paired-end read information after mapping step⁸ (Figure 4). There are different software to align the DNA data; the most used in NGS are based on index-based algorithms, such as Burrows-Wheeler Aligner (BWA) and Bowtie2⁶⁸.

However, mapping algorithms are not exempt from errors. One of the main limitations is the correct alignment of short reads (NGS) in repetitive or complex polymorphic regions, such as regions with high CG content, or the impossibility to align reads in absent regions of the reference genome, such as gaps or structural variants^{8,69}. In such cases, short reads cannot map or map incorrectly to the reference genome, leading to false-positive detections by software (variant callers). For these reasons, different tools, such as Biobambam2⁷⁰, Picard, or Alfred⁷¹, are used to evaluate the quality of the alignment. 1000G and GoNL, for example, evaluated different parameters, such as the fraction of reads aligned in pairs or mean insert size, among others, providing an idea about the alignment quality^{1,2}.

In this direction, the Genome Analysis Toolkit (GATK) development team, and the Genome Sequencing and Analysis (GSA) group of the Broad Institute, developed Best Practices recommendations in order to improve the BAM file construction to decrease the False Discovery Rate (FDR) derived from (i) duplicated reads in PCR amplification and (ii) systematic errors produced by sequencing machines to calibrate base quality scores⁷². These recommendations consisted of marking the duplicated reads and recalibrating the base quality scores (BQSR) produced by the sequencing machine. Besides, selecting a reference genome for the alignment step is also key to deplete read misalignments or artefact reads. For example, the hs37d5 reference genome allows the filtering of conflictive reads by including decoy sequences, such as BAC/fosmid clones, HuRef contigs, Epstein-Barr Virus genome, and microsatellites, facilitating the variant detection by variant callers.

Additionally, the alignment of short reads onto sex chromosomes is a challenge due to the high similarity between X and Y chromosomes, producing technical artefacts and affecting downstream analyses on variant calling⁷³. Nowadays, all reference genomes include both sex chromosomes in sequencing studies, making this approach not accurate to detect variants in the

X chromosome due to the scavenger effect of the Y chromosome in female samples^{73,74}. For this reason, generating two reference genomes based on sample gender will improve the alignment and variant detection in these chromosomes.

Overall, Genotyping or CGH arrays identify SNPs, indels and CNVs. In combination with imputation analysis, this approach improved the statistical power of GWAS. On the other hand, high through sequencing methodologies can detect the whole genome variability, increasing the computational requirements, due to space and aligning/mapping steps. However, sequencing approaches overcome the limitations of genotyping arrays, enabling to detect all variant types, and increasing human genome variability insights.

1.2.3. Variant calling in the sequencing era

In variant calling using whole-genome sequencing (WGS) data, BAM files play a key role in the detection of genome variability. The accuracy of variant detections is influenced by variant properties, such as the type and size^{8,9,34,46} or library particularities, such as coverage, read length, insert size, among others^{9,34,46}. These particularities also affect the accuracy and the possibilities of variants characterisation (i.e. the resolution of variant position, variant size and type). Nowadays, more than 150 variant identification programs (variant callers)⁷⁵, having a wide range of possibilities to detect the genome variability.

Few variant callers are able to detect SNVs, indels, and SVs. For SNVs, the strategy followed consists of looking for discordant sequences when comparing them with the reference genome. The read and paired-end information are key to detect indels (< 50 bp) and large structural variants (SVs). Based on the anomalously mapped reads and coverage, five approaches are developed to detect the genome rearrangements^{18,46,76,77} (Figure 4). (1) Split read strategy (SR) detected all SVs at base-pair resolution by using the reads broken in multiple parts and aligned to both breakpoint sides. Nevertheless, there are difficulties in aligning these reads to the reference genome, resulting from the split read, limiting the detection of large SVs. To solve this barrier, sequencing the samples at high coverage (30X) could improve the SR signals, allowing them to detect large SVs. 2) Discordant-read strategy (DR) uses the inner mate distance and paired-end information to detect the genome rearrangements. All anomalous mapped paired-end read information, such as the distance between reads, the read orientation, the order of reads, or the mapped reads in different chromosomes. The breakpoint resolution is influenced by the coverage, the insert size mean, and its standard deviation. In contrast to SR strategy, this method is less accurate to report the exact breakpoint. The DR strategy limitations are related to the read mapping issues, such as the mapping in repetitive regions or the impossibility to map reads in *de novo* insertions larger than insert size average. 3) Read-depth (RD) strategy divides the genome in bins and uses the sample coverage to detect abnormal RD, where higher and lower RD are then classified as DUP and DEL, respectively. Despite the breakpoint resolution being poor, this method is able to detect large CNVs in detriment of small ones. 4) *de novo* Assembly strategy (AS) is a sophisticated method of building long DNA chains by assembling contiguous reads to generate contigs. These contigs are then aligned to the reference genome to detect all SV types. This approach tries to solve the alignment deficiencies by constructing longer stretches than read lengths, improving the mapping ambiguities near SVs. However, this approach is highly computationally demanding and requires high-coverage data to construct contigs accurately, avoiding assembly errors^{18,76,78–80}. Finally, 5) the new generation of variant callers apply Machine Learning algorithms (ML) to detect genome variants. One of the most

popular is Deepvariant⁸¹, a new TensorFlow machine learning-based, used to improve the detection of detect SNVs and indels.

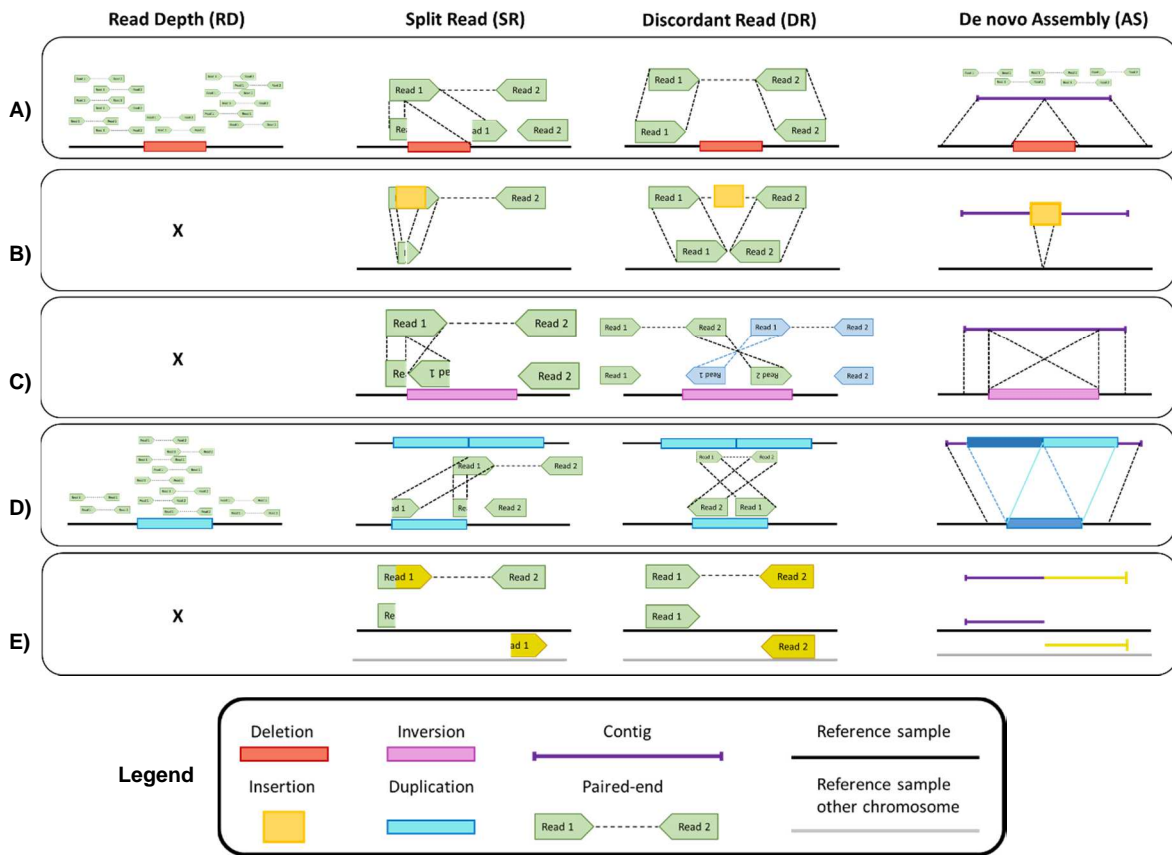


Figure 4. Strategies to detect structural variants and indels. RD uses the sequencing coverage to detect low or high depths, after coverage normalisation. In SR, the read covers the breakpoint junction. When this read is mapped to the reference genome, both read segments are aligned flanking the breakpoint. In DR, the orientation distribution, and inner mate distance of paired-ends allow for the detection of structural variants and their type. In AS, after constructing contigs, they are aligned to the reference genome. The disposition of contig segments in the reference genome allows the detection of SVs or indels. Each block of the figure illustrates which features are used to discover **A)** Deletions, **B)** Insertions, **C)** Inversions, **D)** Tandem duplications and **E)** Translocations.

Figure adapted from Tattini et al⁷⁷.

However, each of these strategies has its own strengths and weaknesses, limiting the whole genome variability detection efficiently^{18,76,77}. For example, the AS and SR can detect the breakpoints and all SV types accurately, as well as SNVs and indels, while DR gives approximate breakpoint positions, and it is appropriate for variants with median sizes. RD is recommended for large deletions and duplications, detecting the breakpoint positions with poor resolution, where AS can be used to detect *de novo* insertions longer than insert size. For these reasons, the new variant caller updates combine different detection signals in a single caller in order to improve the recall and precision.

1.2.3.1. The identification of SNV and small indels

SNVs and indels (< 50bp) are the most prevalent across the human genome, estimated to reach a median of ~3.3-4M SNVs and ~492K-851K indels per genome^{10,82,83}. The majority of

these variants do not likely affect any function and are not reflected in specific phenotypes (i.e. they are neutral). However, some of them are clinically relevant. For example, 3% of all cancer cases have a hereditary component⁸⁴. SNPs close to the TERT gene confer low risks in some cancers, such as breast, colorectal, and testicular. Besides, the SNVs and indels are key in GWAS studies in order to find associations of variants to complex diseases, such as diabetes. The resolution of GWAS studies could improve by using panels of genetic variability, where the samples are sequenced with NGS technologies, detecting the whole genome variability of a sample.

Variant callers focused on SNVs detection are highly accurate, thanks to small sizes, facilitating read mapping. Besides, the lower costs of sequencing at high coverage provide more signals to detect those events correctly^{8,85,86}. On the other hand, the indel detection presents more difficulties due to low concordance between sequencing platforms, alignment errors in repetitive regions, indel size, or different representations between variant callers^{87,88}. For example, GoNL reported that the power to detect the alternative allele correctly decreases with indel length increases, concluding as short indels the insertions and deletions ≤ 20 bp². Besides, HRC constructed the reference panel using SNPs since they estimated indels to be very inconsistent across projects. For this reason, the normalisation of indels is required. The Global Alliance for Genomics and Health (GA4GH) designed a pipeline to standardise the SNP and indel representations, allowing the comparison of the outputs of different variant callers⁸⁸.

Several variant callers detect SNVs and indels. The most popular is Haplotype caller⁸⁹, which uses a combination of SR and AS strategies to detect genetic variants. Others widely used are Freebayes⁹⁰, Platypus⁹¹, VarScan2^{92,93}, Strelka2⁹⁴, or Deepvariant⁸¹. These tools used different approaches to detect genetic variants; for example, Strelka2 and Platypus use AS strategy, Deepvariant use a deep learning model or VarScan2, use the SR. However, not all variant callers exposed have the same capabilities to detect indels, for example, Strelka2 is able to detect indels up to 50 bp, and Haplotype caller can detect indels bigger than 100 bp, indicating that the variant caller selection is of paramount interest, in order to cover all indel sizes accurately.

1.2.3.2. Structural Variant detection

Structural Variants (SVs) are the major source for biological variability within populations^{9-11,75,95}. These variants are largely responsible for the evolution as well as numerous phenotypes in humans⁹. Besides, SVs can modify gene expression, topological associating domains (TAD), or disrupt protein-coding genes, producing an impact on gene function or resulting in different rare or complex diseases^{5,11,65,66,95}. Notwithstanding its importance on human diseases, SV analysis compared to SNV is lagging, mainly because of technical limitations. Since the first Copy Number Variant (CNV) detected in the 2000s by genotyping array, the SV discovery has evolved in conjunction with sequencing technologies⁹⁵. A substantial improvement appeared in 2007 with the Next-Generation Sequencing (NGS), allowing for the detection of SVs in whole genomes, with a bias towards in non-repetitive regions. Finally, the Third-Generation Sequencing (TGS) appeared in 2015 and allowed the detection of SVs in all the genomic genome^{8,95} (Figure 5). During the NGS and TGS period, different public databases such as dbVar⁹⁶ or Database of Genomic Variants (DGV)^{96,97} collected validated SVs, gathering 18,366,594 SVs in DGV and 36,126,123 SVs in dbVar. Recently, new SV catalogues have appeared, such as gnomAD-SV¹¹ and one from the Ira M.Hall lab⁶⁵. In addition, some samples have been characterised using long-reads, such as in the Human Genome Structural Variant consortium (HGSV)^{66,67}.

Current challenges in SV detection by NGS have raised multiple questions regarding the number of SVs per genome. From the first estimation done in 2015 by 1000G, which consisted of 2,100-2,5000 SVs per genome^{1,95}, to more recent results of the gnomAD-SV project that reaches 7,439 SV per individual¹¹, the sequencing technologies have evolved (higher coverages and long-read sequencing, for example), allowing the discovery of more SVs. The TGS studies estimated that a typical genome has > 20,000 SVs per genome^{66,67,95}, showing the NGS limitations for accurate SV detection.

Nowadays, NGS technologies contributed vastly to SV detection progress due to their lower costs as compared to TGS. However, variant callers used to detect SVs with NGS vary in recall and precision due to SV features, library properties, or the genomic context^{9,34,46}, generating a wide range of 9 to 89% False Discovery Rate (FDR), and a recall rate between 10 and 70% for some SV types and sizes^{48,64,86,87}.

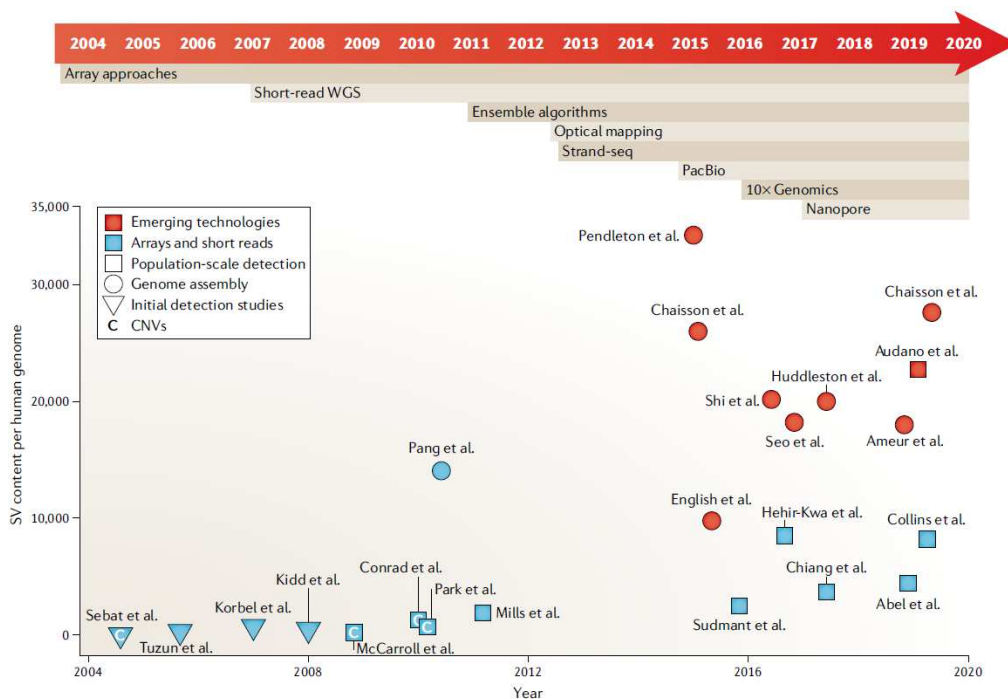


Figure 5. Evolution of Structural variant detection across sequencing technologies. The SV detection is correlated with technical improvements, where deeper coverages (30X) of Next-Generation Sequencing and higher lengths of Third-Generation Sequencing enabled to discover more SVs per genome. *Figure from Ho et al.⁸⁵.*

Currently, no single variant caller is considered standard because they cannot cover all SV types and sizes^{9,46,76}. In this direction, multiple variant callers have been developed in the scientific community, that detect different types of variants, including SVs subtypes such as deletions, duplications, insertions, inversions, translocations and transposons, using paired-end read and alignment information⁴⁶ (Figure 4). For example, one of the most popular is Delly²¹⁰¹, which combines SR, DR, and RD strategies to detect the whole SV spectrum. Other variant callers used multiple SV detection signals, such as Lumpy¹⁰², SvABA⁷⁹, Pindel¹⁰³, or Whamg⁷⁸, taking advantage of different combinations in SV analysis. However, other variant callers such as Manta¹⁰⁴ or Popins⁸⁰ only used the AS strategy, since it is one of the most accurate strategies to detect SVs, thanks to constructing contigs based on paired-end reads. Nevertheless, this approach requires high computational resources and coverage, limiting their use in specific projects¹⁰⁵. Besides, some variant callers are specialised to detect specific SVs, such as

CNVnator¹⁰⁶, mainly focused on deletions and duplications, Popins, in *de novo* insertions, MELT¹⁰⁷, in transposon detection or ViFi¹⁰⁸ and VERSE¹⁰⁹, designed to detect viruses (detailed description in Table 13).

Despite the wide variety of variant callers and combinations of multiple strategies, SV detection is still not accurate. For this reason, there has been an increase in interest in integrating the outputs of individual variant callers, improving their strengths, and reducing the false-positive detections, without losing recall^{5,9,46,76,95,110} (further details in section 1.2.6).

1.2.4. Genotype the variants using sequencing reads

Most of the studies are focused on variant detection of variant callers, but limited information is about the process to determine the genotype of those variants. The variant callers perform two processes: 1) The variant discovery (see section 1.2.3), and 2) the genotyping of these candidate variants (genotype calling), which corresponds to their characterisation as to their homo or heterozygous state. For this reason, it is important to note the distinction between variant and genotype calling^{54,86,111–115}. For example, hard filters are applied in variant calling because it has to reduce the false-positive detections^{86,114}. In contrast, genotype calling characterises the allelic state of the variant, having more relaxed filters. Genotype calling is thus key in population-genomic studies; for example, an accurate genotype can improve the statistical power of GWAS studies by finding more variants in linkage disequilibrium (LD), especially for rare variants^{116,117}. Besides, improving the genotyping provides more realistic information about the variant allele frequency in a population, correcting the estimation of population size and allowing to determine the relatedness grade between relatives. For these reasons, accurate genotypes are necessary to increase the performance of genetic variability panels for GWAS studies¹¹⁷.

The uncertainty to report the genotype accurately is correlated to coverage^{86,118}. Based on this factor, there are two approaches to genotype the variants using NGS: the genotyping can be classified as 1) low-medium coverage (7-12X) methods, which use a probabilistic approach to report the genotypes, and 2) high coverage methods (>20X), which count the reads and determine with high precision the ploidy of the sample¹¹⁸.

The genotype likelihood (GL) is used to genotype variants from samples sequenced at low-medium coverage (7X-14X). The GL is computed by the $p(X|G)$ formula, where G is the genotype, and X are all reads of an individual at a particular site. Then, in conjunction with genotype priors (a group of genotypes from databases), the posterior genotype probabilities are calculated, providing for each variant three probabilities, regarding each allele state (0/0, 0/1, or 1/1). The highest probability is considered the correct state. The genotype priors are necessary to improve the genotype accuracy, correcting the genotype probabilities with the allele frequencies of populations or multi-samples. For this reason, to increase the genotype accuracy, more samples are necessary for the genotyping step. However, when the coverage is increased to >20X, the uncertainty to report the genotype decreases because more reads are evaluated, determining a heterozygous variant if the proportion of non-reference reads is between 20% to 80%; otherwise, the genotype is deemed homozygous⁸⁶. For example, in SNV detection, to call a homozygous variant, the coverage required is 15X; however, the heterozygous requires 33X, showing the importance of coverage to genotype the variants correctly^{63,64}.

Currently, the strategy to report the genotype between SNVs and indels differs slightly from SVs. The SNVs and indels record in a genome Variant Calling Format (gVCF) file the whole

genome positions, and whether there is a variant or not (ex: Haplotype caller, Deepvariant)^{8,115}. All samples can then be combined and re-genotyped, recovering some variants missed in the variant calling step, due to the hard filters of variant discovery. However, this approach is not followed in structural variants due to the imprecision to report the breakpoint position, which does not allow the combination of different samples¹¹⁵. Some variant callers such as Delly or Lumpy (SVTyper¹¹⁹), after the variant calling step, are able to combine all samples in a multi-sample VCFs, enabling the posterior re-genotyping of whole samples, using the positions reported in the VCF, allowing to recover SVs not accepted in variant calling step. Nevertheless, because this strategy is computationally demanding, it is challenging to implement for most studies.

On the other hand, a new generation of tools emerged to genotype SVs, such as BayesTyper¹²⁰, SVJedi¹¹⁴, or Vg¹²¹. These tools require a VCF file with all the SV candidates. Then, the genotyping is performed using the BAM file of each sample and evaluating the sample ploidy using the VCF. In this direction, different new catalogues can be used to genotype SVs, such as gnomAD-SV, Ira M Hall resource, or even the dbVar catalogue. This strategy allows the SV genotyping in whole samples from VCF, obtaining the allele frequencies for each SV in the cohort. However, it can not detect new SVs, limiting the insights into human genome variability and losing the rare and low-frequency variants from specific populations.

In summary, genotype calling is of paramount interest in population studies, where an accurate genotype could ameliorate the quality of genetic variability panels, improving the imputation analysis for GWAS. Sequencing the samples at high coverage enabled the genotyping of the variants without GL, improving the genotyping resolution. Finally, genotype the variants using an SV catalogue as a template, could evaluate the SV candidates in each sample, obtaining the allele frequencies in the cohort. However, this approach limits the detection of new SVs.

1.2.5. Benchmarking variant identification of variant callers

Although NGS technology can be used to detect any germline variant, no single variant caller is able to detect the whole variability landscape accurately, having different strengths and weaknesses, when facing different variant types⁹. Many factors influence the accuracy of variant calling, such as coverage, insert size, and read length from library sequencing or variant features, for example, type, length, and frequency⁴⁶. For this reason, golden samples (reference samples) are necessary to compare the efficiency and accuracy of variant callers.

Currently, the Genome In a Bottle (GIAB) consortium is investing efforts in generating different reference samples to benchmark variant callers, with NA12878 sample the first reference sample launched by GIAB¹²² and one of the most used. This consortium applied multiple sequencing platforms, aligners, and variant callers to correct the variant detection errors, producing an accurate data set for SNV and indel¹²². In 2016, GIAB provided six more samples (trios from Ashkenazi Jewish (AJ) and Han Chinese ancestry) sequenced with different platforms, covering NGS to TGS technologies¹²³. Besides, the last updates provide 17% new SNVs and 176% new indels, compared to older releases, generating a complete catalogue for benchmarking analyses¹²⁴. Recently, GIAB has focused on including an SV golden set of large deletions and insertions (≥ 50 bp), which allows, for the first time, the benchmarking of germline SVs¹²⁵. However, these reference samples have some limitations. The variants recommended for benchmarking are in high evolutionary sequence conservation regions (conservative regions) that may be easier to detect and genotype¹²⁰, overestimating variant detection and genotyping

accuracy. Besides, the SV dataset does not cover the broad spectrum of SVs, being thus insufficient to perform a complete benchmark.

Given that no real sample is thoroughly characterised, it is necessary to generate artificial samples (*in-silico*) that can cover these limitations, controlling the variants included in the genome, and allowing the evaluation of the variant caller accuracy under the desired conditions¹²⁶. The majority of simulators require a reference sequence to create an *in-silico*. Then, to generate the sequencing library, many parameters can be included, such as the insert size, the read length, coverage, the sequencing platform, even the sequencing errors allowed by the sequencing machine. This versatility can lead to simulated data that can then be used to evaluate the performance of variant calling closer to reality. Alternatively, several simulators can generate synthetic variants directly into the BAM file, such as BAMSurgeon. However, this strategy cannot evaluate the variant effect of the alignment step. In contrast, other simulators, such as ART¹²⁷, can create the *in-silico* samples by sequencing. This tool requires a reference sequence with all variants inserted; then, ART emulates the sequencing machine generating FASTQ files. This approach controls all sequencing parameters and includes all variant types, from SNVs to all SV types and sizes.

Independently of the reference sample used, the variant caller benchmark is performed evaluating the following metrics: recall, precision, and F-score. True-Positive (TP), False-Negative (FN), and False-Positive (FP) variant detections are used to calculate these metrics. The recall is the fraction of TP variants detected among all sample variants. Then, the precision is the proportion of TP that are positives. Finally, the F-score is a harmonic mean of precision and recall.

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

Different benchmark analyses have been conducted, demonstrating the capabilities of different variant callers in many variant types. In SNVs, the F-scores are similar in medium (15X) and high (30X) coverages⁸³. For example, using the NA12878 sample sequenced at 30X, the F-scores obtained from Deepvariant, Haplotype caller, and Strelka2 were similar, at 0.98 and 0.98 and 0.99 (using Illumina Hiseq4000 machine), respectively, demonstrating that high coverage improves the in SNV detections^{83,128,129}. However, with variant size increase, the accuracy of variant detection decreases. For example, in indels (< 50bp), the F-scores of previous variant callers were 0.94, 0.90, and 0.90^{128,129}, respectively, showing a reduction between 4-8%. These results indicated that variant detection of small variants was highly accurate; however, they were not exempt from a small proportion of false discoveries.

On the other hand, compared to SNVs and indels, the accuracy of SVs discoveries manifested larger divergencies. The accuracy is influenced by many factors, such as the SV type, size, coverage, among others^{9,130}. For example, the precision and recall in the deletion discovery of Manta are 95.9% and 83.1%, respectively. These values change in the case of inversion discovery, at 97.6% and 80.9%, or in the detection of large insertions, where the precision is 96.5% and recall 11.9%⁹. These divergences increased across different variant callers, finding more variability in their metrics. Kosugi et al.⁹ performed a benchmark of 69 algorithms,

determining that Delly, Manta, Lumpy, SvABA, Pindel, and Wham, were among the most accurate callers to detect SVs. However, size is an important factor in detecting SVs. With increases in the SV size, the mapping quality of reads decreased, leading to misinterpretations and increasing false-positives⁸. Besides, it is known that all variant callers do not detect all variant size ranges equally⁹, varying their recall and precision depending on the SV size, or detection method, among other features. For example, CNVnator is useful to detect large CNVs (>1Kb)¹¹², demonstrating that many factors have to be considered to perform an accurate SV calling.

Besides, the reference samples can be used in innumerable analyses, for example, in the correct determination of the SV size, which is crucial to reach consensus and consider an SV the same across different variant callers. In this direction, Lumpy, Manta, and SvABA are able to report accurately deletions and duplications, thanks to using the DR strategy⁹. On the other hand, to determine the position where the SV has occurred is another relevant feature, because it allows improving: 1) the variant classification and annotation, 2) the evaluation of the functional impact of the SV, 3) the construction of personal diploid genomes¹³¹, and 4) the merging of redundant SVs between different samples or across variant callers. For example, Delly and SvABA are highly accurate to report the breakpoint in deletions and duplications, and Manta and Wham for insertions⁹. These particularities indicate that no single variant caller can accurately detect whole SVs and sizes, highlighting that more than one variant caller has to be used to detect all SVs accurately.

In conclusion, reference samples are needed to evaluate the variant caller performance, to adapt the variant detection with the sequencing data, to improve the strengths of each approach, and to reduce false-positive detections.

1.2.6. Variant caller integration: Improving the accuracy of variant detection

Given that, as previously discussed, the variant discovery has an FDR between 9-89% and recall 10-70%, depending on the SV type, having ample room to improve. Until the NGS improvements, offering, for instance, a decrease in sequencing errors, or an increase in the coverage and read length, or until the integration of different read signals (RD, DR, SR, or AS) is ingrained into the variant callers, a different approach is needed to improve variant discovery accuracy. For this reason, recently, there has been an increased interest in integrating the outputs of individual variant callers to improve the strengths of different approaches reducing the false-positive detections and without losing recall^{8,9,46}.

The difficulty of integrating different variant callers increases proportionally with variant size and complexity. For example, for SNVs, merging the outputs of different variant callers is not a challenge, because the position and alternative allele (ALT) have to be the same, filtering the redundant SNVs. In the case of indels, the size and breakpoint resolution increase the complexity. GoNL classified the indels by small < 20bp or mid-deletions between 20 and 100bp. The merging strategy for small indels was the same as for SNVs. In contrast, the mid-deletion positions did not have a base-pair resolution, so a breakpoint error of ± 10 bp had to be included to consider mid-deletions the same across variant callers¹¹⁰. However, if the variant is included within the read size, the errors due to variant length only affect breakpoint resolution and thus are not critical.

The main challenge is then to integrate the variant caller outputs for SVs, where the variant size and type, breakpoint, or even the strategies to detect the SV, are key to consider an

SV as the same across different variant callers^{9,110,132}. Firstly, the SVs are classified by their type. Secondly, the position is determined by the most accurate strategy (for example, SR and AS), ensuring that the final breakpoint is as precise as possible. However, there needs to be a breakpoint-error margin in order to group the redundant SVs. Finally, the SV size is used to consider an SV as the same across different variant callers. Depending on the study, if the lengths are between 50-80% in reciprocal overlap (RO), the SV is redundant, allowing their merge through variant callers^{5,9,110}. Besides, combining algorithms based on different detection methods could improve the SV discovery, increasing the precision and recall instead of combining just the same methods due to the incorporation of different signals⁹.

In this direction, independent integrating tools such as SVmerge¹³³, MetaSV¹³⁴, SURVIVOR¹³⁵, or Parliament¹³⁶, have demonstrated an improvement of recall and precision compared to single callers. However, these tools combine the variant caller outputs using different heuristic decision rules and validate the SVs applying an assembly-based method⁹⁵, which is not the optimal way to obtain the best merge results because the particular strengths of each variant caller are not taken into account, losing power in recall and precision¹³⁰. GoNL is an example of combining variant callers using heuristic rules, in which an SV is true-positive if at least two variant callers detect it. This approach is inaccurate because, depending on the precision of the pair of algorithms, it could include more false-positives for a determined size⁹.

To overcome the heuristic methods limitations and provide specific relevance in each variable, such as different variant callers, sizes, number of strategies, among others, creating sophisticated machine learning models could improve the precision and recall values. There are different machine learning strategies, such as Logistic Regression Model (LRM), Support Vector Machine (SVM), random forest, or convolutional neural networks. For example, LRM is one of the basic machine learning models, where the result of the outcome variable using multiple predictor variables is a binomial response (ex: PASS/NO PASS). LRM will model the chance of an event based on different factors¹³⁷. The following formula is applied in the LRM model:

$$\log\left(\frac{\pi}{\pi-1}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

where π is the event probability (ex: variant is true-positive or false-positive), β are the coefficients associated with the golden group (in our case the *in-silico*) and X are the explanatory variables of each query. In this direction, FusorSV^{132,95} applies a random forest model which uses the size, and SV type as discriminative variables, maximising the SV discovery without losing precision. Compared to MetaSV, FusorSV outperformed SV detection, demonstrating that machine learning algorithms increase SV discovery performance. However, FusorSV just includes two variables in their model, suggesting that incorporating more signals could maximise the variant detection accurately¹³⁰. In the currently published reference panels, only the 1000 Genomes used a machine learning algorithm (Support Vector Machine (SVM)) to merge and filter indels as well as SV¹.

Finally, variant caller integration could reduce the false-positive detections and the genotyping errors as well. For example, the Genome Strip decreased genotype errors by using different strategies to detect and genotype the CNVs^{54,62}. In this context, in the same way as variant discovery, by combining the genotypes obtained from calling, which could decrease genotype errors. For example, Manta, Lumpy, and Pindel genotypes are highly accurate in reporting the genotypes⁹, so combining their decisions should decrease genotype errors.

1.2.7. The correlation of variants across the human genome

It is known that different human traits are influenced by genetic factors, such as hair or eye colour, or our risk to develop certain diseases. SVs seems to have a similar evolutionary history as SNPs and indels, being ancestral, appearing in human history and being shared across individuals by inherited factors rather than sporadic events²². It is estimated that ~4M of SNVs¹⁰ and > 20K SV^{66,67,95} are included in a typical genome; however, the genome structure is modelled by population genetic forces, such as genetic drift, genetic flow, natural selection, among others, generating correlation patterns between genome variants¹³⁸. This idea is taken by linkage disequilibrium (LD), where the variants integrated into the same genomic interval are often correlated^{22,138,139}.

Besides, LD depends on local recombination rates. The recombination is done in meiosis and consists of exchanging the genetic material between chromosomes, breaking the chromosomes by genomic segments (haplotype block). Each haplotype block includes several alleles, which are inherited together, so one variant could tag another just by LD. Thanks to haplotype blocks, the LD patterns can be used to cover longer chromosome segments, enabling the test of one variant in each block, and recovering significant information able to associate a haplotype block to a trait or disease^{138,139}. The variants that capture the genomic segment's variation are named tag variants (tag SNPs).

However, recurrent recombinations tend to break the haplotype blocks, reducing their sizes and driving the segregation of the variants independently across generations (linkage equilibrium)²². The LD segments in African populations are shorter than in populations of European or Asian descent, due to the accumulation of recombination events. The founder events in European and Asian populations altered their genetic structure, reducing genetic variability, the population size and generational age, and increasing the length of the LD segments^{22,139}. For this reason, the LD patterns are population-specific, reflecting their demographic evolution; as such, not all casual variants are tagged by the same SNPs across all populations^{6,139} (Figure 6), demonstrating the importance to characterise the genetic background of specific geographic regions, in order to understand the genetic architecture of diseases in a single population. Besides, SNP arrays are built using tag SNPs^{139,140}, thanks to the genomic characterisation of the HapMap project, which determined that most common SNPs ($MAF \geq 5\%$) could be reduced to 550,000 SNPs in European and Asian ancestries and to 1,100,000 SNPs in African ancestries²². Tag SNPs allowed the capture of the majority of common genetic variability in humans, enabling the evaluation of large genomic segments, and ultimately, to improve GWAS.

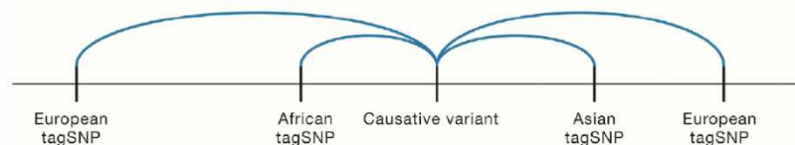


Figure 6. Linkage disequilibrium (LD) patterns are particular of populations. Casual variants can be tagged by different SNPs in different populations, due to LD patterns, highlighting the importance to characterise the genetic background of specific populations. This is one factor which explains the difficulties to replicate GWAS studies across populations. *Figure from Sirugo et al.*⁶.

Particularly, SVs are non-randomly distributed across the genome; for example, structural variants can be generated in segmental duplications by NAHR procedures, determining that SV presence depends on the local genome architecture⁶⁷. However, LD patterns between tag SNPs and SVs are hard to determine. SVs are enriched in segmental duplications, exhibiting low LD due to the paucity of the genotyped SNPs, reducing the chances to tag SVs compared to other genomic regions. Nevertheless, common SVs in segmental duplications and those in unique regions are in LD with neighbouring SNPs^{22,141}, allowing the use of tag SNPs to recover common SVs for association tests. Further SNP characterisation across segmental duplications is necessary to improve the LD between SNPs and SVs. For this reason, generating panels of genetic variability using TGS technology could improve the SNP characterisation in low complexity regions, increasing the tag SNPs chances to find LD patterns between SNPs and indirectly to recover more SVs from haplotype blocks.

1.2.8. The functional impact of variants on the human genome

The germline variants are implicated in the evolutionary adaptations and heritable diseases; thus, the characterisation of their functional impact on humans is one of the main goals in biomedicine¹⁷. Even though SNPs and indels are the most common genetic variability source, SVs due to their sizes are responsible for the majority genetic diversity between two human genomes⁸. Consequently, SVs affect more DNA stretches than SNVs and indels, with potential implications for gene function^{8,9}. For this reason, further annotations are needed in SVs, in order to increase the insights of SVs on human phenotypes.

Different annotations were applied to characterise the functional impact of SVs on gene function in humans, for instance, evaluating overlaps with intergenic or intragenic regions. Nowadays, functional interpretation is attributed to protein-coding genes. Depending on the variant location, the gene could be disrupted, losing their function and affecting human phenotype. Most SVs overlapped intronic regions, limiting their understanding and implication on traits. On the other hand, variants located in CoDing Sequencing (CDS) regions could modify the protein, affecting the gene function directly^{142,143}. However, most variants associated with traits (90%) are located outside of protein-coding regions, such as regulatory elements. Enhancers, promoters, or untranslated regions (UTRs), among others, could perturb the gene expression or indirectly modifying the function of other genes^{35,144}. These appreciations are based on predictions, helping to interpret the variant effect on human traits or diseases; however, all predictions have to be validated in the wet lab, in order to confirm these hypotheses.

Besides, some genes are less tolerant to variation than others, increasing the risk to affect their function. The predicted loss of function (pLoF) is the probability of a gene to be deleterious due to a variant, with a likelihood of clinical significance¹⁴⁵. A pLoF is ≥ 0.9 indicates low variation tolerance; these cases are named as predicted loss of function intolerance (pLI). The pLI is widely used in population genetics to predict the gene intolerance under a particular variant^{75,145,146}. The extreme case is haploinsufficiency (HI), where the heterozygous genes are insufficient to maintain the function^{145,147}. These measures do not provide information about variant dominance¹⁴⁵. For example, one disease could develop a stronger effect in homozygous patients than heterozygous.

Rare and complex diseases are two major groups with a heritable component due to germline variants. Rare diseases also are referred to as monogenic or Mendelian diseases, and

are characterised to develop the disease as a result of a single or few variants, with a negligible contribution of environmental factors⁴⁴. A great example is the Haploinsufficiency A20 (HA20) monogenic disease, caused by heterozygous variants in the *TNFAIP3* gene, producing a systemic inflammation in multiple organs¹⁴⁸. Therefore, the high penetrance of those variants in diseased individuals tends to decrease their prevalence in populations, given that these are rare and restricted in particular families⁵² (Figure 7). However, not all deleterious traits are negatively selected; for example, variants in particular populations are highly frequent, such as G6PD deficiencies and thalassemias.

On the other hand, complex (or common) diseases are influenced by environmental and polygenic factors, and it is difficult to detect causal variants. Many common variants are suspected of contributing a small fraction to disease risk, conferring low penetrance, and modulating the disease susceptibility^{44,52}. As a result, the weak variant selecting pressures increase the allele frequencies into the population (Figure 7). Besides, population migrations could interfere with modelling the genetic architecture of diseases between different geographic regions. The founder events could particularly increase the allele frequencies of weakly deleterious alleles, facilitating the introduction of multiple low-frequency variants in specific populations, resulting in different risk variants for the same disease across populations^{6,44}. Different diseases are described as complex, such as type II diabetes or schizophrenia, directly impacting to healthcare system costs¹⁴⁹. For this reason, GWAS tried to identify these risk variants in order to design new prevention and treatment strategies¹³⁹. In this direction, large catalogues of causal and risk variants were created such as Online Mendelian Inheritance in Man (OMIM)¹⁵⁰ or Genome-Wide Association Studies (GWAS) Catalogue¹⁵¹, collecting variants of rare and complex diseases⁴⁴.

Currently, to obtain a functional interpretation, different software are used to annotate genetic variability; for example, the most popular tools are SnpEff¹⁵² and Annovar¹⁵³, specialising in SNVs and indels. However, they are of limited use to characterise the wide spectrum of SVs. Recently, a new annotation tool named AnnotSV⁷⁵ improved on these limitations, using different repositories, such as OMIM¹⁵⁰, DGV⁹⁷, 1000G⁵, dbVar⁹⁶, GeneHancer³⁷, among others (further details in Table 12), helping the prediction of the pathogenic effect of SVs on humans.

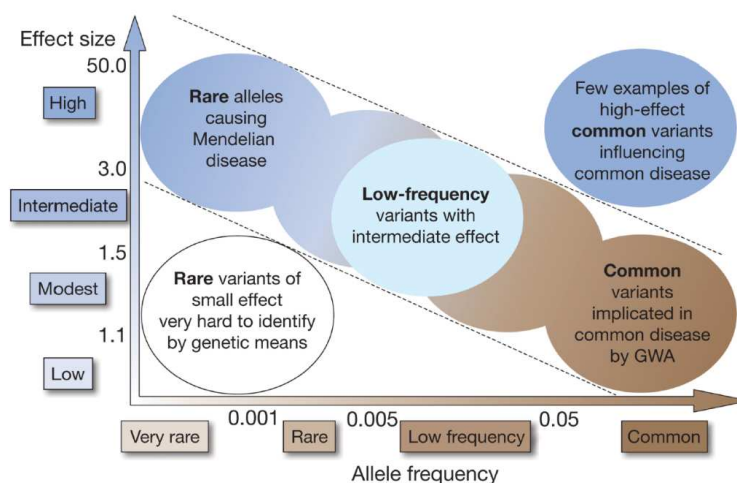


Figure 7. Genetic architecture of rare and complex diseases. In rare diseases (Mendelian), the penetrance of rare variants is high, thus, the allele frequency in a population is low (MAF < 1%). While the penetrance decreases, polygenic risk variants contribute to the complex (common) disease susceptibility, increasing the allele frequency in the population. Figure from Teri et al.⁵²

1.3 Genetic variability panels (reference panels): An invaluable resource in Genome-Wide Association Studies (GWAS)

Genome-Wide Association Studies (GWAS) have identified thousands of risk variants and their biological function, revolutionising the biomedicine field¹⁵⁴. In comparison to rare diseases, complex diseases are influenced by multiple risk variants, as well as social and environmental factors. Through the interrogation of hundreds of thousands of SNPs, GWAS allows associating thousands of risk variants to complex diseases, such as diabetes or coronary heart diseases, elucidating their genetic architecture^{52,154,155}. Thus, GWAS open new opportunities to find new therapeutic targets in order to treat heritable diseases¹³⁹.

The rationale behind GWAS is the comparison of the genetic background between cases (diseases) and controls (healthy), providing statistically significant associations at each polymorphic site between variants and disease susceptibility^{52,155} (Figure 8). GWAS rely on exploit the LD principle, where the statistical power to find associations between variants and traits depends on the allele frequency of variants, sample size, the distribution of effect size of causal genetic variants in the population, and the LD between genotyped variants and unknown causal variants¹⁴⁶.

As expected, the degree of LD between tagSNPs and rare casual variants is according to allele frequency, limiting the GWAS analysis to common variants ($MAF \geq 5\%$). In brief, the LD is linear to sample size. Thus, due to the allele frequency of rare variants ($MAF < 1\%$), the power of common variants to detect associations between rare casual variants is negligible¹⁵⁵. For this reason, increasing the sample size will improve the GWAS resolution, increasing the chances to find new risk variants in complex diseases.

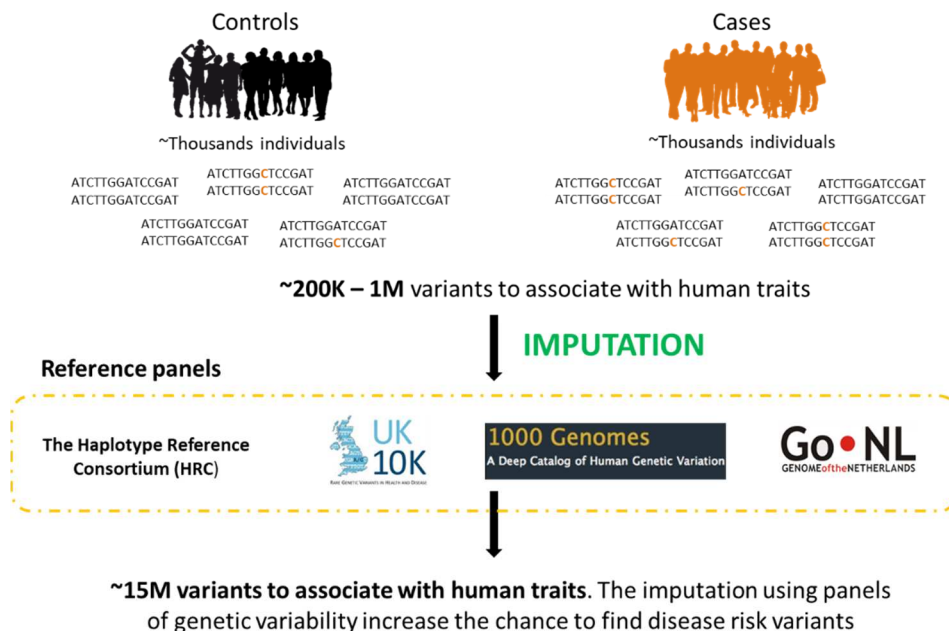


Figure 8. Workflow of GWAS analysis. The GWAS approach consists in comparing the genetic background between cases and controls, in order to find risk variants of the disease. To improve the statistical power and resolution of GWAS, the imputation analysis based on LD, allows recovering unobserved variants, by inferring genotyped tagSNPs to the haplotypes from a fully sequenced reference panels (see further details in section 1.3.2).

Despite the GWAS contribution to improving the genetic architecture insights of complex diseases, associating more than 50,000 risk variants¹⁵⁴, different constraints are involved in this type of studies. For example, the GWAS studies explain a small proportion of complex diseases heritability, mainly due to the low capacity of genotyping arrays to detect rare variants (hypothesised to be highly penetrant in diseases), limitations in evaluating structural variants, incapacity to detect gene-gene interactions, and the influence of environmental factors⁵². However, the main objective of GWAS is to explain the contribution of genetic variability to complex diseases, instead of to explain the whole heritability. To decrease the missing heritability in GWAS, increase the sample size is necessary. For example, in 2009, one locus was related to schizophrenia using 3,000 cases. Then, in 2014, the number of loci increased to 108, using 35,000 cases, explaining a substantial proportion of heritability¹⁵⁵.

On the other hand, the majority of GWAS has been conducted on individuals of European descent (52%) or Asian (21%). Notwithstanding the growing number of studies including African, and Hispanic descent, these disparities are unacceptable in GWAS^{6,155}. The main reason is that GWAS is based on LD, and this factor is closely related to the population evolution of each geographic region, each with its specific LD patterns, which influences how well a tag SNP captures a casual variant. Thus, the risk variants for complex diseases can differ between populations, complicating GWAS replication across different ethnic groups⁶. For example, in type 2 diabetes, GWAS identified different loci with high risk in East Asian (*KCNQ1*), Mexican (*SLC16A11*) and Greenlandic (*TBC1D4*) populations, evidencing differences in risk allele frequencies among populations¹⁵⁴. Nevertheless, common variants contribute in small fraction to the genetic architecture of complex diseases, and are expected to be evolutionarily old and shared across different populations, which is encouraging to find common patterns between ethnicities^{154,155}. Despite this, some common risk loci still differ in allele frequency or effect size across populations. In summary, there is a need to characterise specific populations¹⁵⁵ in order to find particular risk variants, which will enable the improvement of the diagnostic, prevention and treatment of complex diseases, contributing to the implementation of precision medicine.

Besides, the imputation of SNP arrays using panels of genetic variability (reference panels), allows to include unobserved variants in GWAS, and to lift the restriction to SNP array data, which only probes a small fraction of genome variability¹⁵⁴⁻¹⁵⁶ (see the detailed explanation of imputation in section 1.3.2). For this reason, the creation of panels of genetic variability is of paramount interest, allowing for the evaluation of more variants in GWAS, and thus, increasing the chances to find new associations between genetic variants and complex diseases.

In brief, a reference panel is a subset of human haplotypes (cohort), which are highly genetically characterised, by having detected their genetic variability by multiple means and reconstructed their haplotypes by phasing approaches. Since the HapMap project facilitated the design of first SNP arrays used for GWAS¹⁵⁶, the improvements in creating complete reference panels using NGS technologies, have allowed the analysis of the whole genome variability from different cohorts to detect rare, low-frequency and structural variants^{52,156}. 1000G was the first reference panel which sequenced 2,504 individuals from multiple ethnicities, detecting a wide spectrum of genetic variability⁵. Since then, several more reference panels have been created, focusing on a specific population (ex: GoNL¹¹⁰) or larger multi-ethnic panels (ex: Haplotype Reference Consortium (HRC)¹⁵⁷), increasing their sample size and coverage (Table 1). Therefore, by using reference panels in GWAS, the statistical power of the analyses will increase,

and they will be able to include more untyped variants, with a special emphasis on rare and low-frequency variants, which are normally not captured by SNP arrays^{154,156}.

| Reference panel | Last Release | Sample size | Coverage WGS | Number of SNPs and indels | Number of SVs | Ancestry |
|----------------------------|--------------|-------------|--------------|---------------------------|-------------------------|--------------------|
| HapMap2 | 2007 | 270 | Genotyped | 3.8M | None | Multi-ethnic |
| HapMap3 | 2008 | 1,115 | Genotyped | 1.6M (SNPs) | None | Multi-ethnic |
| 1KJPN* | 2014 | 1,070 | 32,4X | 24.6M | 82,620 (del, Ins, CNVs) | Japanese |
| 1000G phase3 | 2015 | 2,504 | 7.4X | 88M | 68,818 | Multi-ethnic |
| UK10K[▲] | 2015 | 3,781 | 7X | 45.5M | 18,739 (large del) | British |
| GoNL-SV[▲] | 2016 | 769 | 14.5X | 21.6M | 35,510 | Deutch |
| HRC | 2016 | 32,488 | Diverse | 39.2M (SNPs) | None | Multi-ethnic |
| Iceland | 2017 | 15,220 | 34X | 31.1M | None | Icelandic |
| Estonian | 2017 | 2,244 | 30X | 16.5M | None | Estonian |
| SG10K | 2019 | 4,819 | 13.7X | 98.3M | None | Multi-ethnic Asian |
| TOPMed[#] | 2019 | 53,831 | 38X | 410M | None | Multi-ethnic |

Table 1. Reference panels available. The main limitations of reference panels were the sample size and coverage. Including more samples, the performance to impute rare variants increases; thus, the new generation of reference panels increase the sample size. Besides, increase the coverage, improving the variant detection and genotyping. The majority of reference panels do not detect SVs; just two panels characterise all SV types, being 1000G the unique catalogue with these variants.

* Reference panel not available; [▲] Public panels do not include SVs; [#] Not published, Paper in bioRxiv

However, most reference panels do not include SVs in their catalogues, despite their importance in complex diseases (Table 1). For example, in neurodevelopmental diseases, such as schizophrenia or autism, specific deletions are involved in both diseases²². The main reasons are the false-positive detections using short reads, low coverage, and the high computational requirements, increasing the complexity of SV discoveries. Currently, 1000G phase3 is the only reference panel available with SVs. Nevertheless, the project used low coverage, hindering variant detection and genotyping. Thus, sequencing samples at high coverage (30X) could improve the imputation performance, and help to include more SVs with an accurate genotype, a necessary step to find more variants in LD^{116,117}. Besides, including SVs in GWAS could explain more heritability in complex diseases^{52,132,158}.

Finally, generating more reference panels is necessary to include more samples and variants in GWAS. Hence, performing variant calling in specific populations could increase the variant discoveries from particular geographic regions, and help find different risk variants across populations. Besides, including SVs in reference panels could improve the GWAS resolution, increasing the chances of finding new risk variants in complex diseases.

1.3.1. The role of phasing in reference panel creation

The haplotype is a combination of alleles from different loci, along a single chromosome, which tend to be inherited together (Figure 9A). The haplotype information is not trivial and has relevant applications, such as detecting positive variant selection by looking for long haplotypes common in the population, estimating the recombination rate, understanding gene function, allowing to decipher the compound heterozygosity cases, or performing disease association studies, among others^{159,160}. Besides, the haplotypes can be used to generate haplotype reference panels, necessary to improve the statistical power of GWAS, by including more untyped variants by the imputation analysis^{5,139,157,161,162}.

DNA is currently sequenced, ignoring the haplotypes from parents; thus, the variant callers only provide the genotype of variants, and not their relative phasing. To generate a reference panel, we have to resolve the human haplotypes as well, from the genotype information. This technique is named haplotype phasing technique, and it aims to recover the two possible haplotypes of an individual^{159,160,162}.

The phasing is only relevant for heterozygous variants, where the nucleotide content differs between homologous chromosomes. Many strategies are designed to phase the genotypes, in which can be catalogued as 1) experimental techniques, 2) population-based strategies, and 3) those who use a combination of population-based strategies with sequencing reads or parental information^{159,160}. The experimental approaches determine the haplotypes directly. However, these are more expensive and labour intensive than population-based strategies, hindering their application in populational studies. The population-based strategies are widely used; their objective is to decipher the most likely set of haplotypes given the input genotypes, using the haplotypes from a population^{159,162} (Figure 9B). Shapelt2¹⁶³, Beagle¹⁶⁴ and Eagle2¹⁶⁵ use this approach, based on a Hidden Markov Model (HMM), in which the mutation rate (emission probability) and recombination rate (transition probability) are applied to the phasing, to predict the haplotype^{159,160,162}. The transition probability is used to determine the sequence changes between different recombination points, and then the emission probability evaluates the haplotype mutability with respect to haplotypes already resolved¹⁶². In other words, the HMM has different nodes, understood as a group of haplotypes. Each node has its mutation rate, and is related to others by recombination points (transition probability). So, when a genotyping dataset is phased, each node solves a haplotype stretch using the haplotypes, the emission and transition probabilities, obtaining the most likely set of haplotypes (Figure 9B). Beagle and Shapelt2 were used to generate the 1000G and GoNL haplotype reference panels.

Besides, when sequencing coverage is high, the reads can be used to improve the population-based strategy. The reads can be thought of as mini-haplotypes, containing phase information, and so could to improve the phasing performance^{159,162,163}. The read or paired-ends which cover at least two heterozygous alleles are named Phase Informative Reads (PIRs)¹⁶³. Besides, it was found that 33% of heterozygous variants were covered by PIRs in a single sample¹⁶³, highlighting the potential of PIRs in haplotype estimation, and improving the phasing of rare and singleton variants¹⁶². The new Shapelt4¹⁶⁶ update uses the WhatsHap¹⁶⁷ tool to recover the PIRs, and then this information is used to improve the phasing of genotypes. However, WhatsHap cannot capture PIRs for SVs, limiting their use in SNVs and indels. This last

approach has not yet been applied to generate a haplotype reference panel, due to the low coverages used or the computational resources requirements.

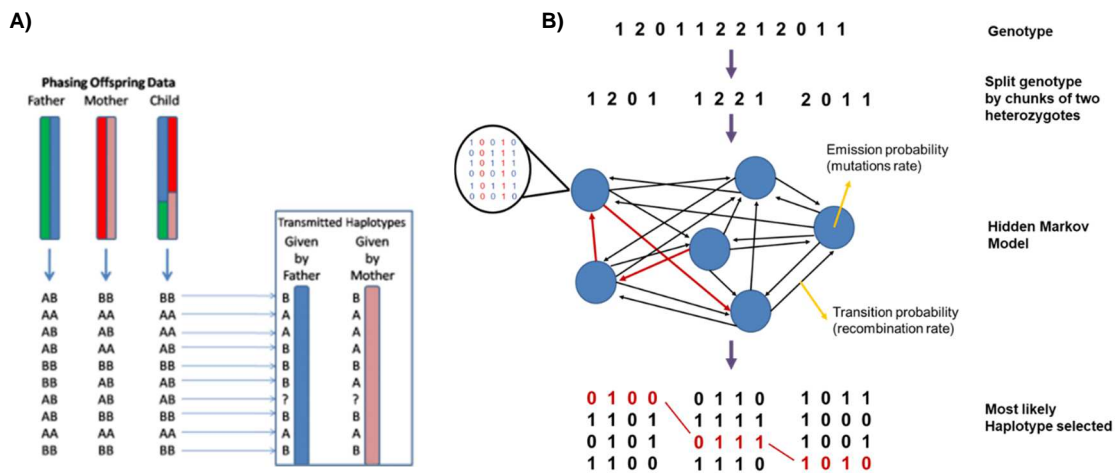


Figure 9. Phasing description. **A)** Phasing consists of resolving the haplotypes of individuals, grouping the alleles which tend to be inherited together. For example, using the genetic information from parents, the haplotype of offspring can be deciphered. The exception happens when parents and offspring are all heterozygous at a locus. In those cases, the population haplotypes can be used to resolve the ambiguous regions. **B)** The Hidden Markov Model (HMM) is used to estimate the haplotypes from an individual. Shapeit2, first divided by chunks the individual genotype data, englobing a specific number of heterozygotes. Then, a HMM based on the haplotype frequencies from a population grouped in a node, transition probabilities (solid black arrows) and emission probabilities (mutation rate of each position into the node), are used to estimate the most likely haplotype. (0= homozygote reference; 1= heterozygote; 2= Homozygote alternative). Figure 9A obtained from <https://www.plob.org/article/11643.html>. Figure 9B. Adapted from Marchini et al.¹⁶².

Although phasing can be applied for structural variants as well, a few studies are available about their accuracy. Nowadays, different studies use long reads to phase SVs¹⁶⁸, but these approaches are not cost-effective to reproduce at a large scale. Besides, the haplotype reference panels, such as 1000G or GoNL (Table 1) used the MVNcall¹⁶⁹ tool to include SVs in their catalogues^{5,161}. The MVNcall is an imputation software, which genotypes and phases the SVs obtained from low coverage (7-15X) sequencing approaches, using a haplotype scaffold, constituted by biallelic SNVs and indels^{162,169}. So, using a haplotype scaffold, MVNcall imputes the SVs obtained from sequencing calling, obtaining their genotypes, and then phased those genotypes following LD patterns¹⁶⁹. This approach tries to reduce the genotype limitations from low coverage in SVs. Thus, to ameliorate SVs' phasing performance, increasing the sequencing coverage will improve the genotype accuracy, allowing the SV inclusion of current phasing tools, instead of inferring their genotypes by imputation approaches.

1.3.2. Genotype imputation in GWAS studies

Genotype imputation is used to predict the variants not directly assayed in a sample of individuals, enabling to inexpensively approximate whole-genome sequence data from SNP array data^{156,170}. This statistical approach facilitates meta-analysis studies, combining the results of different studies, equating the set of variants obtained from different SNP array data. Moreover, the imputation also improves the resolution of a genetic region, increasing the chances to find a casual variant (fine-mapping). Finally, in combination with haplotype reference panels, imputation

increases the power of GWAS, including untyped variants such as structural variants or low-frequency and rare variants (MAF < 5%), normally not captured by SNP arrays^{156,170}.

The rationale behind imputation is that two unrelated individuals can share different DNA stretches, derived from a common ancestor. Thus, the variants detected by commercial arrays (usually genotyping SNP arrays) can be used to identify haplotype blocks (DNA segments) shared between genotyping arrays (sample) and haplotype reference panels (Figure 10). So, the haplotypes of a sample can be understood as a mosaic of short segments of related haplotypes found in the reference panels, enabling to impute unobserved variants in the study sample by LD patterns^{156,162,170}.

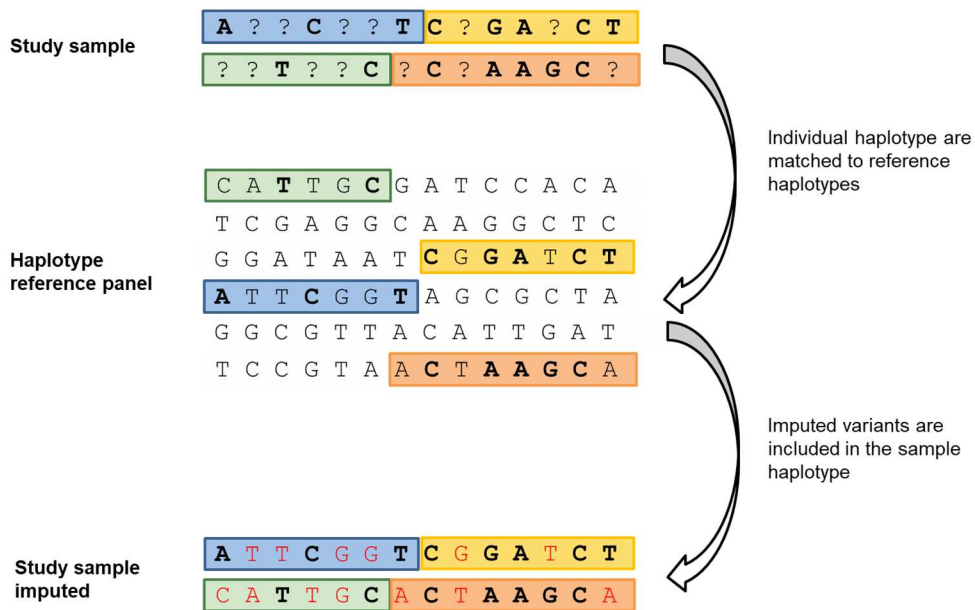


Figure 10. Genotype imputation description. The pre-phased tagSNPs from SNP array data (top) are used to match short segments from a reference haplotypes set (middle). Therefore, the haplotypes from the study sample can be represented as a mosaic of segments of a haplotype reference panel. Consequently, untyped variants can be inferred, generating the final haplotype of the sample, including the observed and unobserved variants. *Figure adapted from Marchini et al.*¹⁶².

Different imputation tools have been developed, which MINIMAC, IMPUTE2 or BEAGLE among the most popular¹⁵⁶. New updates of these tools focus on decreasing the computational requirements and increasing the execution speed. On the other hand, MVNcall infers the genotypes using the genotype likelihoods (GL), estimated from SNP array or low-coverage sequence data^{156,162}. This approach includes complex rearrangements and SVs in haplotype reference panels^{5,161}. However, imputation is computationally demanding due to the high rate of missing data. For this reason, pre-phasing the genotyping data before imputation allows a reduction of the computational burden, without compromising the accuracy. This decreases the complexity of the imputation step because two haplotypes can be contrasted directly, rather than against all pairs of haplotypes^{156,170}.

The quality of variants imputed is affected by several factors^{156,162,170}, such as the allele frequency of imputed variants, genomic context or the size of reference panels. Imputing rare variants (MAF < 1%) is harder than common ones, due to the low chances to observe those alleles in the reference panel. This complicates finding LD patterns, and the set of template

haplotypes available for matching decreases, resulting in lower imputation accuracy. For this reason, increasing the sample size of reference panels could improve the imputation of low-frequency and rare variants, which are more likely to be related to diseases¹⁵⁴. For example, HRC or TOPMed increased the imputation accuracy of rare variants as compared to 1000G, where the sample size was small. Besides, an increase in the coverage of sequencing will improve the genotypes from reference panels, which in turn correlates with the accuracy of inferred haplotypes¹⁵⁶. In addition, including more SNPs in a SNP array, could improve the chances to find shared haplotype segments between sample study and reference panels, improving the GWAS.

To evaluate the imputation accuracy, IMPUTE2¹⁷¹ uses the “info score” metric, and MINIMAC or BEAGLE the r^2 . These two measures are correlated and can be interpreted similarly. The value ranges between 0 and 1, indicating the precision of variant imputed. Usually, all variants with an info score < 0.3 are discarded for association tests¹⁵⁶.

Besides, demographic properties could affect the imputation performance^{139,156,162}. If the study sample has the same ancestry as the reference panel, the probability of matching different haplotypes increases. In this direction, the low haplotype number of each population in multi-ethnic reference panels decreased the power to impute low-frequency, rare, and that population-specific variants^{4,6,156}. For this reason, the generation of population-specific reference panels, including more haplotypes of a specific demographic region, could increase the chances to impute low-frequency and rare variants, more likely related to diseases.

Overall, combining population-specific and multi-ethnic reference panels could help for disease studies, increasing the chances to impute casual variants, since a hybrid panel is enriched with rare alleles^{156,161}. In this direction, IMPUTE2 can combine two different reference panels, improving the imputation performance (mainly for rare and low-frequency variants)¹⁶¹. However, many more reference panels are already built (Table 1), increasing the need for methods able to perform imputation using several whole reference panels, to obtain more variants for GWAS. In this context, GUIDANCE¹⁷² is able to impute the study sample against multiple reference panels independently in one run, recovering the most accurate set of variants imputed from each panel, and thus improving the GWAS resolution.

In conclusion, imputation analysis is useful to increase the number of variants in GWAS, opening new opportunities to find causal variants in complex diseases. Therefore, creating more haplotype reference panels will help increase the sample size, ameliorating the imputation analysis. In addition, generating more population-specific reference panels will increase the chances to find causal variants in particular populations, filling a crucial need in precision medicine.

1.4 The rationale of this thesis

Genome Wide Association Studies (GWAS) have revolutionized biomedicine, associating more than 216K causal and risk variants in multiple diseases or human traits. The imputation of SNP array data to haplotype reference panels has allowed the inclusion of more variants in GWAS, increasing their statistical power and chances to find more disease-associated variants¹⁵⁶. However, more samples and haplotypes from specific populations are needed in order to increase the imputation of low-frequency and rare variants^{156,162}. Besides, generating

population-specific panels could help understand the genetic bases in complex diseases, finding specific risk variants across populations⁶. This is particularly interesting in precision medicine, because the diagnostic, prevention and treatment efficacy could differ depending on the ethnic group.

Nowadays, sequencing a cohort at high coverage (30X) is possible due to the decreasing costs over the years, improving the accuracy of genome variability detection, genotyping and phasing. For example, *de novo* assembly strategies require high coverage to detect variants correctly^{18,76,80}. Besides, high coverage allows the detection of SVs with more accuracy than low coverage, providing new genomic rearrangements in SV catalogues^{9,34}. In addition, high coverage allows the genotyping of SNVs, indels and SVs correctly, thanks to more available signals (reads) to determine their allele dosage. This results in more variants in LD, essential to building a haplotype reference panel of quality^{86,117,118}. Finally, phasing can be improved using the reads as mini-haplotypes, enabling the generation of sample haplotypes more precisely, and thus, to built a much more accurate haplotype reference panel^{162,163}.

Structural Variants are the main source of genetic variability across populations, and are implicated in several human traits or complex diseases^{8,9}, such as neurodevelopmental ones, highlighting the importance to include these rearrangements in haplotype reference panels. Currently, just GoNL and 1000G include the majority of SVs types^{5,110}, where 1000G is the unique dataset available. These projects detect SVs at low and medium coverage, limiting the SV detection performance. Besides, the SV genotypes are inferred using a haplotype scaffold^{5,110,156,162}, evidencing the genotype limitations of low coverages. Since the costs of TGS technologies does not decrease, the best option to characterise SVs is to increase the coverage of NGS sequencing, and to combine different variant callers to obtain accurate catalogues of SVs^{9,11,95}. Ultimately, this will lead to the generation of new haplotype reference panels that include a more precise SVs genotypes.

In this context, this thesis has been performed in collaboration with the GCAT-Genomics for life (GCAT) project¹⁷³. The GCAT project is a prospective long-term study that tries to evaluate the effects of epidemiological, environmental, and omic factors such as genomics, metabolomics, epigenomics, and proteomics on chronic diseases^{173,174}. GCAT sequenced 808 volunteers using NGS at high coverage (30X) from the North-east region of Spain (Catalonia) at ages between 40-65 years, grouping all samples as Iberian ethnicity (IBS). This data enabled us to generate a population-specific panel of genetic variability of the Iberian population, improving the detection of SVs as well as their genotypes. This resource is of paramount interest for many reasons: 1) This represents the first population-specific Iberian reference panel. 2) This panel includes a detailed characterisation of genomic rearrangements in the Iberian population, improving the imputation of SVs and population-specific variants. 3) The combination of this novel reference panel with others previously published, will help improve the performance and resolution of GWAS studies. Finally, 4) This resource provides information on novel genomic rearrangements, which will be useful to improve the insights of genetic variability in humans.

The Iberian population has a complex demographic history, unusual across European regions, mainly influenced by Muslim rule, with estimated proportions of north-west African-like DNA in the Iberian population ranging from 2.4-10.6%¹⁷⁵. Besides, the genetic footprint of Iberians is present in the Latin American populations, due to colonisation, sharing DNA segments between these ethnic groups¹⁷⁶. These particularities could be beneficial for GWAS because many

haplotypes could be shared across different continental groups^{139,156}. Hence, generating a haplotype reference panel from the Iberian population could contribute to perform better GWAS in these communities, normally underrepresented⁶. In Europe, the Iberian population differs genetically within populations, clustering ethnic groups by geographic regions, demonstrating different genetic particularities as a result of demographic evolution^{175,177}. In addition, the genetic structure in the Iberian Peninsula is not homogeneous; for example, there are isolated populations such as Basque Country^{175,177}, increasing the complexity to apply an accurate GWAS in this population.

Besides, several studies have revealed different particularities of inherited diseases in the Iberian population, finding a high prevalence of some rare diseases¹⁷⁸, such as mild phenylketonuria phenotypes or high variant heterogeneity causing cystic fibrosis in comparison to central and northern European populations. However, the allele frequencies of risk variants of complex diseases, such as type 2 diabetes, neurodegenerative or cancer are not different from other European populations¹⁷⁸. Despite this, the characterisation of the genetic background of the Iberian population could help to find private risk variants, improving the diagnosis, prevention and treatments of complex diseases. Thus, generating a panel of genetic variability of the Iberian population could improve precision medicine in Spain, helping to maintain the current healthcare system.

2. OBJECTIVES

The main objective of this thesis has been to generate an haplotype reference panel of genetic variability for the Iberian population, mainly focused on Structural Variants, for which our knowledge is still lacking. This resource will provide new opportunities to understand the effect of SVs on diseases and improve the comprehension of genetic variability effects on humans.

Additionally, other objectives were derived from the main one:

- 1) The generation of a strategy to characterise the landscape of germline variation from whole-genome sequenced cohorts, especially for structural variants.
- 2) The generation of filtering strategies to minimise the fractions of false-positives among raw calls.
- 3) The application of this strategy to 785 genomes from the GCAT cohort to catalogue and annotate their genetic variation.
- 4) The construction of haplotype blocks using phasing strategies and generate an haplotype-based reference panel, capturing a wide spectrum of Iberian genetic variability.
- 5) The comparison and evaluation of the possibilities of this panel for variant imputation and interpretation for GWAS.

3. MATERIAL AND METHODS

This chapter can be divided into two blocks. The first block, including sections from 3.1 to 3.4, describes **the strategy designed to characterise all genome variability** of the GCAT samples. The second block, from sections 3.5 to 3.12, is focused on **creating the Iberian reference panel**, which comprises the BAM file construction, goes through the building of the haplotype reference panel, and terminates with the functional implications of Structural Variants (SVs).

3.1 Creation of *in-silico* sample

In this section, we describe in detail the *in-silico* sample (artificial sample) created in-house. The paragraph is divided into three blocks: the first one describes the *in-silico* sample and its content. In the second one, we summarise the process of creating their BAM file. Finally, we illustrate the benchmarking of the different software for variant detection (variant callers), which allowed us to make a preliminary draft of which tools were the most suitable to build a reference panel of the Iberian population, including 808 GCAT-Genomes for life (GCAT) samples¹⁷³.

The artificial sample (*in-silico*) has been generated for mainly two reasons: 1) **to understand the functionality of the variant callers used** to characterise human genetic variability, and additionally to calibrate the parameters of these tools. 2) To our knowledge, **there is currently no real sample with a well characterisation of large SVs well characterised**. Thus, we generate an *in-silico* sample containing the following known variant types: Single Nucleotide Variants (SNVs), small Insertions-Deletions (Indels, size from 1 to 30 bp), and SVs.

There are different ways to generate an artificial sample using simulations¹²⁶. We used the **ART¹²⁷ (ART-illumina) strategy**, simulating the Illumina sequencing, the same Next-Generation Sequencing (NGS) platform used to sequence the real samples from the GCAT project. Using the FASTQ files generated with ART-Illumina, we evaluated the alignment errors produced by the variants. In the following subsections, we detail the steps to generate our artificial sample.

3.1.1. *In-silico* sample description

To generate an *in-silico* similar to a real sample, we inserted in the reference genome (hg19) some real variants **known in humans**. These variants were obtained from popular projects such as the 1000 Genomes¹ (1000G) and the PanCancer project¹⁷⁹. Additionally, **we inserted artificial SVs not present in these projects**, using python scripts (version 2.7.13) developed in-house.

The *in-silico* sample includes SNVs, Point Mutations, Indels and different types of SVs such as large Deletions (DELs), Insertions (INSs), Inversions (INVs), Duplications (DUPs), Copy Number Variants (CNVs), Translocations (TRAs), Transposons (TRPs), Viruses (VIRs) and Pseudogenes (PSGs), covering the broad spectrum of genome rearrangements. Below we describe the particularities of each rearrangement. Table 2 overviews all variant label information from *in-silico*.

SNVs and small indels were selected from 37 samples of 1000G (complete sample list in Supplementary Table 1). We mixed different alleles of these samples to create one new haplotype with real variants. Table 2 shows all variants obtained from 1000G labelled with the flag "Germline." Additionally, we included Point Mutations and small Deletions and Insertions ("indels" flag name) from the PanCancer project to complete the selection.

We consider different types of **DELs**, according to **1) The size; 2) If the variant was obtained from a project; 3) if it was associated with a TRA; 4) and associated with a complex event**. The flags for the different DEL types were:

- “no_indels”: Were obtained from indel files of the PanCancer project and had a size larger than 100 bp.
- “consensus”: Were obtained from SV files of the PanCancer project. The sizes rank from 101 to 9000 bp.
- “big_del”: These were the nine largest Deletions in the *in-silico* obtained from the PanCancer Project, with a size more than 10000 bp.
- There were two random DEL types related to complex events:
 - “random_transDel_IDtranslocation” was a non-reciprocal TRA. The *IDtranslocation* was a number that connects the DEL with their related TRA. The size of these DELs and TRA is the same.
 - “random_SVtype_flank”: These were small DELs (larger than ten bp) flanking another SV. Usually, these DELs were not detected by the variant callers.

INSS were catalogued in different groups. Below, all INS type are described (Table 2):

- “random_ins_new”: **The INSS with these flags were not obtained from any project mentioned before**. These sequences were created randomly using a python script developed in-house (version 2.7.13).
- “random_SVtype_shard”: These INS were genomic shards⁴⁹, small fragments interposed between breakpoints. These events were inserted between both breakpoints or on one side of INVs, TRAs, DUPs, VIRs, and PSGs. The genomic shard events only appear in variants larger than 200 bp.

INVs were obtained from two different sources:

- “consensus_inv”: Obtained from the PanCancer project.
- “random_inv”: INV randomly generated using an in-house script.

DUPs were classified using the different labelling listed below. Besides, we included some mCNVs, mainly tandem duplications:

- “random_dup”: These DUPs were generated randomly. These events were inserted near to the original fragment, so they were considered tandem duplications.
- “random_dup_inv”: These DUPs were generated randomly and inserted in inverted mode near the original fragment.
- “Variant_in_chr12_dup”: We duplicated the whole allele1 from chromosome 12. This duplicated allele include the same variants as allele1. In addition, we inserted 57 new variants in it.
- “consensus_tran”: These were dispersed DUPs. These DUPs were inserted in a different location of the *in-silico* genome and obtained from the PanCancer project.

- “random_dup_inv_tandemnumber” / “random_dup_tandemnumber”: These events were mCNVs generated randomly. Those DUPs were repeated more than two times. The number after the word “tandem” indicated the number of tandem repeats composing a CNV (ex: random_dup_tandem3).

There were different **TRAs**. The different types of TRAs were catalogued as reciprocal and non-reciprocal:

- Using a python script, we generated the non-reciprocal TRA randomly. As we mentioned for the DEL variants, these TRAs contain an *IDtranslocation*. Here we show an example:

Translocation flag= random_trans_547091

Deletion flag= random_transDel_547091

The numerical string indicates the relationship between the DEL and TRA. This link allows the identification of all TRA and their respective DEL. Both inter- and intra-chromosomal TRAs were generated.

- There were four reciprocal TRA. The steps to construct these events were described in section 3.1.2. The flags related to these events were "translocation_chr15_chr9" / "translocation_chr9_chr15", "translocation_chr16_chr8", and "translocation_chr8_chr16, respectively.

VIRs are events related to the insertion of an exogenous viral sequence into the human genome. We selected retroviruses described in the NCBI database for humans. There were two parameters to catalogue the VIRs, 1) the genomic region where the VIRs were inserted and 2) if the VIR sequence was completely inserted or truncated. At least 10% of viruses wherein intronic regions.

- “virus_intron”: Related to VIRs inserted in the intronic regions, and the sequence was truncated.
- “virus_intron_complete”: The VIRs were inserted in the intronic region, and the sequence was completed.
- “virus_random”: The VIRs were inserted randomly in the genome.

We included in the *in-silico* **PSGs** described in the literature and included in humans from the NCBI database. All PSGs had the same flag, “random_GEN.” The *GEN* is the gene-name related to the PSG (ex: random_MYH11).

TRPs were of different types, such as ALUs, LINES, and SVAs. All TRPs were obtained from the Repbase database¹⁸⁰. The flag associated with these events is “random_retrotransposon_nametranposon.”

All variants were randomly distributed between two files, one for each chromosome haplotype (1 and 2 respectively), excluding the telomeric and centromeric regions. This procedure generated heterozygous SVs only. To obtain **homozygous ones**, we copied some variants to the homologous haplotype. The word “Homozygotic” in the label indicates that a variant was homozygous (ex: random_inv_Homozygotic). The homozygous and tandem duplications were the unique variants repeated, and they have to be considered as **one variant**.

| FLAG | VARIANT TYPE | PROJECT | TOTAL EVENTS |
|--|--------------------|--------------------|--------------|
| <i>Germline</i> | SNV | 1000G | 4,871,660 |
| <i>Germline</i> | Insertion/deletion | 1000G | 454,963 |
| <i>PointMutation</i> | SNV | PanCancer | 723 |
| <i>indel</i> | Insertion/deletion | PanCancer | 3,084 |
| <i>no_indel</i> | Deletion | PanCancer | 269 |
| <i>consensus</i> | Deletion | PanCancer | 26 |
| <i>random_transDel_IDtranslocation</i> | Deletion | Random with python | 674 |
| <i>big_del</i> | Deletion | PanCancer | 9 |
| <i>random_SVtype_flank</i> (Flanked Deletions) | Deletion | Random with python | 8 |
| <i>random_ins_new</i> | Insertion | Random with python | 440 |
| <i>random_SVtype_shard</i> (Genomic shard) | Insertion | Random with python | 9 |
| <i>consensus_inv</i> | Inversion | PanCancer | 14 |
| <i>random_inv</i> | Inversion | Random with python | 241 |
| <i>random_dup</i> | Duplication | Random with python | 458 |
| <i>random_dup_inv</i> | Duplication | Random with python | 17 |
| <i>random_dup_inv_tandemnum,</i> <i>random_dup_tandemnum</i> (mCNV) | Duplication | Random with python | 62 |
| <i>consensus_tran</i> | Duplication | PanCancer | 14 |
| <i>random_trans_num</i> | Translocation | Random with python | 604 |
| <i>translocation_chr8_chr16,</i> <i>translocation_chr16_chr8</i> | Translocation | Random with python | 2 |
| <i>virus_intron, virus_intron_complete,</i> <i>virus_random</i> | Virus | Random with python | 89 |
| <i>random_GEN</i> | Pseudogene | Random with python | 9 |
| <i>random_retrotransposon_nametransposon</i> | Transposon | Random with python | 100 |
| <i>random_SVtype_Homozygotic</i> | All SV Variants | Random with python | 1,150 |
| <i>Variant_in_chr12_dup</i> | All Variants | Random with python | 56 |

Table 2. Description of variant types inserted in the in-silico generated sample. All variants obtained from the 1000G or PanCancer projects have the flag consensus, Germline, PointMutation, indel, or no_indel. Python scripts generated the variants with the random flag. The transposons and viruses were obtained from the NCBI or Refseq databases and were inserted randomly in the genome.

3.1.2. Procedure to insert variants into the reference genome and create the *in-silico* sample

Once all *in-silico* variants were listed, a human genome reference sequence was selected to insert the variants into them. All FASTA files were downloaded of the hg19 reference genome assembly by chromosome (<http://hgdownload.cse.ucsc.edu/goldenpath/hg19/chromosomes/>). A Perl script developed by our group was used to add the variants to the FASTA files (v5.18.2).

For each chromosome, we generated two haplotypes with their respective variants. We started to insert the variants at the end of the chromosome, to avoid the alteration of positions from the reference sequences. Using this strategy, all variants were inserted in the FASTA files at base-pair resolution. Next, we generated the reciprocal translocations. These events were created using a python (version 2.7.13) script. In this way, we had an *in-silico* sample ready to use for simulating sequence.

We used ART¹²⁷ (version 2.5.8) to simulate the sequencing and create synthetic NGS sequencing reads. This approach allowed the evaluation and benchmark tools for variant discovery and read alignment. In our simulation, we used the Illumina platform to replicate the sequencing of GCAT samples (section 3.5.1). The sequencing description is detailed in Table 3. The command used to run ART-Illumina is the following:

```
art_illumina -ss HS20 -i input.fa -rs 4 -qs 1 -qs2 1 -d _idallele_ -p
-l 100 -f 15 --mflen 500 --sdev 20 -o out
```

We sequenced each haplotype separately; for this reason, the coverage of sequencing (15X) is half of the **total coverage in the *in-silico* (30X)**. On the other hand, the GCAT samples were sequenced with a read length of 150 bp, in contrast to the *in-silico*, where the simulated length was 100 bp. This feature is important to detect SVs, because when the read length increases, variant callers can more easily find SVs^{67,181}, so reducing the read length increases the complexity of detecting the SVs by variant callers.

| Features of sequencing | Parameters |
|------------------------|------------|
| Sequencing system | HiSeq 2000 |
| Read length | 100 bp |
| Insert size | 500 bp |
| Inner mate distance | 300 bp |
| Simulation | Paired-end |
| Coverage per allele | 15 X |

Table 3. Sequencing description of *in-silico*.
bp = Base Pair

The ART-Illumina produced two FASTQ files per chromosome, one for each allele. Finally, with the FASTQ files, we constructed the BAM files using the following pipeline:

1- Converting the FASTQ to SAM

We used the BWA tool (version 0.7.15-r1140) with the new algorithm called “mem” to get these files. We aligned the *in-silico* to the Reference Genome (RG) “decoy” (hs37d5).

In this step, for each allele, we merged the two paired-end reads generated by ART-Illumina.

```
bwa mem -M -t 4 hs37d5.fa input_first_paired.fq.gz
input_second_paired.fq.gz | gzip > allele_output.sam.gz 2>
allele_output.sam.err
```

Converting SAM to BAM and sorting the BAM files

To convert the SAM files to BAM files, we used Samtools (version 1.5):

```
samtools view -uS allele_input.sam.gz | samtools sort -m
4000000000 -o allele_output.bam 2> allele_output.bam.err
```

2- Merging the BAM files and filtering it following the Best Practices of GATK

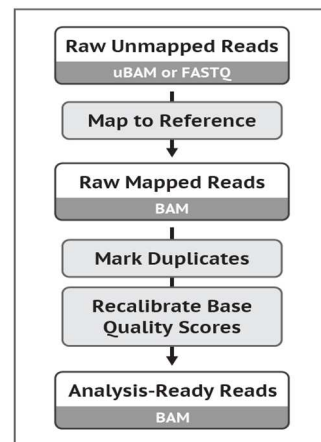
We used Samtools (version 1.5) to combine all alleles and chromosomes in a single BAM file:

```
samtools merge -rh header_bamfile output_insilico.bam
input_allele_chromosome.bam1 input_allele_chromosome.bam2 ...
```

We applied some data clean-up operations in order to correct some technical biases and make the data suitable for downstream analyses. Otherwise, we would have increased the risk of false-positives in the variant calling step. In Scheme 1, we illustrate the steps to carry out the Best Practices of GATK⁷².

We marked the duplicated reads as recommended by **the Best Practices of GATK⁷²**, using PICARD (version 1.108).

After marking the duplicated reads, we indexed the BAM file using Samtools (version 1.5). Finally, we recalibrated the Base Quality Scores (BQSR) of the BAM file using two modules (`VariantRecalibrator` and `ApplyVQSR`) of the GATK4 package (version 4.0.11). We recalibrated the bases considering 1) the base cycle, 2) the original quality score from the NGS platform, and 3) the dinucleotide context.



Scheme 1. Pipeline followed to create the BAM files. Figure taken from of GATK consortium.

3.2 Benchmarking of different variant callers

To select/optimize the variant callers to analyse the genome variability of the GCAT samples, we tested different tools on the *in-silico*, classified into tools for SNPs/Indels and tools for SVs. We used different software and combined its results in order to improve the accuracy of variant

detection and reduce false discoveries. An overview of variant callers described below can be found in Table 13.

All software explained below was installed and executed in the Marenostrum4 supercomputer from the Barcelona Supercomputing Center (BSC). This machine is composed of 48 racks housing 3,456 nodes, with a total of 165,888 processor cores and 390 Terabytes of main memory. Compute nodes are equipped with:

- Each node includes 48 cores.
- L1d 32K; L1i cache 32K; L2cache 1024K; L3 cache 3,3792.
- 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 were supported by vectorisation instructions such as SSE, AVX up to AVX-512.

3.2.1. SNV and Indel calling

3.2.1.1. Haplotype Caller (GATK4 version 4.0.2.0)

We ran Haplotype caller⁸⁹ (HC) to detect SNVs, indels (1 to 30 bp), mid-size deletions (31 to 150 bp), and INSs. The strategies used to identify the variants were Split-read (SR) and *de novo* Assembly (AS). We used java version 8u131 to run HC following three steps: 1) Run the Haplotype caller in `gvcf` mode; 2) Run the `GenotypeGVCF` module, and 3) apply a variant filtering with `VariantFiltration` module. For all the steps, we used the default parameters and 16 CPUs.

3.2.1.2. Deepvariant (version 0.6.1)

We ran Deepvariant⁸¹ to detect SNVs, indels, and mid-size deletions. This program uses a Machine Learning (ML) algorithm based on a deep neural network, which accurately improves genome variability detection. This tool was executed following three steps by default parameters: 1) Run `make_examples.zip`, 2) Run `call_variants.zip`, and 3) Run the `postprocess_variants.zip`. The first step was parallelised in order to improve their performance. We used 48 cores for all the executions.

3.2.1.3. Strelka2 (version 2.9.2)

We used Strelka2⁹⁴ to discover SNVs, Indels, mid-size deletions, and INS. This caller uses the *de novo* assembly (AS) strategy. This tool requires two main steps: 1) Run the `configureStrelkaGermlineWorkflow.py` using default parameters, and 2) Run the `runWorkflow.py` with `-m` local flag. To execute Strelka2, we required 48 cores to reserve an entire node memory for each execution.

3.2.1.4. Platypus (version 0.8.1)

Platypus⁹¹ was used to detect SNVs, Indels, MNPs, and DELs up to 300 bp. This tool uses the strategy of local *de novo* assembly (AS). Before running Platypus, we imported Bcftools (version 1.6), python (version 2.7.13) and htlib (version 1.5). We ran Platypus using default parameters, including the following flags: `--assemble=1 --assembleBrokenPairs=1 --mergeClusteredVariants=1 -nCPU 16`. These flags allowed the detection of DELs with

lengths between 50bp to 2kb and INS between 50-500 bp. To run Platypus, we used 16 CPUs, for two main reasons: 1) to parallelise the execution to improve the performance and 2) for memory issues.

Platypus was discarded in the calling of the GCAT samples for two reasons: 1) the flag `--assemble=1`, produced different computational errors. 2) The values of recall in the *in-silico* and GIAB sample (section 3.3) were the lowest of all callers, that detect SNPs and Indels (Table 14 and Table 15). Also, the inclusion of this caller in the Logistic Regression Model (LRM) produced no improvements in the precision and recall values of the model (Table 14 and Table 15).

3.2.1.5. VarScan2 (version 2.4.3)

VarScan2⁹³ was used to detect SNVs and indels. This tool uses the SR strategy plus the information of map quality, coverage, and base quality. We ran Varscan2 in germline mode, as follows: 1) we executed the mpileup module from Samtools (version 1.5) by default, including the following parameters `--no-BAQ --min-MQ 1`. Then, 2) we ran VarScan2 using the modes `mpileup2snp` and `mpileup2indel` to detect SNVs and Indels, respectively. Besides, we used the following flags (`--min-coverage 10 --min-var-freq 0.20 --p-value 0.05`) to remove the low-quality variants detected by this variant caller. For memory requirements, we used 16 CPUs.

VarScan2 was discarded in the calling step on real samples for the following reasons: 1) As this caller can be run in a multiple sample mode to improve calling and genotyping, we did a pilot study to run four samples altogether, generating in the first step a mpileup file of 753G in eight hours. However, the time consuming to process this file in the second step was more than 48 hours. For space and time reasons; thus, the analysis of the 808 GCAT samples with this caller was not computationally feasible. 2) Besides, we included this caller in the LRM without improving the accuracy results of SNVs and Indels (Table 14 and Table 15).

3.2.2. Large Structural Variant (SV) calling

Currently, there is no available tool able to find all types of SVs and lengths accurately. It is mainly due to the strategies used by the variant callers, which have strengths and weaknesses. Furthermore, the sequencing read length and coverage are also fundamental factors for the proper detection of SVs. For these reasons, we used different tools to improve the detection of SVs and to filter out false-positives. This section will explain which callers we run/evaluated in the *in-silico* sample to select the optimal SV detection in the GCAT samples.

3.2.2.1. Delly2 (version 0.7.7)

Delly2¹⁰¹ was used to detect DELs, DUPs, INVs, INs, and TRAs. The recent version combines different strategies to improve variant detection, like split-read (SR), Discordant-Reads (DR), and Read-depth (RD). We ran this tool with default parameters. We modified the `-t` parameter to `DEL, DUP, INV, IN, and TRA`, depending on the variant type to analyse. We excluded the telomere and centromere regions with `-x` flag, because they are known to be prone to false-positives. For memory requirements, we needed 24 CPUs to execute Delly2 on Marenostrom4.

3.2.2.2. Manta (version 1.2)

Manta¹⁰⁴ was used to detect DELs, DUPs, INVs, INs, and TRAs. This tool combines SR, DR, and *de novo* assembly (AS) strategies. To run Manta, we followed two steps: 1) we ran the `configManta.py`, which scans the genome to find SV associated regions. 2) We ran the `runWorkflow.py`, obtained from the previous step to provide a score and the VCF output of genome variability in the sample. We executed the steps using default parameters and 24 CPUs for memory issues.

3.2.2.3. Pindel (version 0.2.5b9)

Pindel¹⁰³ was used to detect DELs, DUPs, INVs, INs, and TRAs. Initially, this caller used the SR strategy, but recent updates used DR too. Some Pindel parameters were modified to improve SV detection, allowing the detection of interchromosomal events and long insertions. Pindel was executed by chromosome using the following flags: (`-a 3 -C -k -l -I -M 8 -T 6 -x 5 -v 10 -c 1 -R hs37d5 -d Feb2009`). In the required config file, we included the insert size of 300 bp Table 3, needed for variant identification. We converted the BCF to VCF, using the `pindel2vcf` module using default parameters. To run Pindel on MareNostrum, we used 8 CPUs, 6 were used to parallelise the execution, and the other 2 CPUs were needed for memory issues.

The output of Long Insertions does not provide the genotype, that is crucial for developing a reference panel of genetic variation. For this reason, we generated a custom script to obtain the genotypes from the variants reported by Pindel, using the read from the BAM. To obtain accurate genotypes from a variant calls, we needed the total coverage of the position where Pindel detected an INS together with the altered reads. All reads were selected in a window of 10 bp from the breakpoint. We filtered all reads with mapping quality ≤ 20 . The final genotype was determined with the following formula:

$$\frac{\text{Total altered reads}}{\text{Total coverage}} * 100$$

If the proportion of altered reads was $\leq 20\%$, the genotype was 0/0, if the fraction was between 0.20 and 0.80, assigned a 0/1 genotype, and if the fraction was ≥ 0.80 , we assigned 1/1 genotype.

3.2.2.4. Lumpy (version 0.2.13)

Lumpy¹⁰² was used to detect DELs, DUPs, INVs, and BNDs (break-end orientation). This variant caller uses SR, DR, and a generic module (like RD). Before running Lumpy, we pre-processed the *in-silico* BAM file, following the recommendations of Lumpy developers. We extracted from the BAM file the split reads with `extractSplitReads_BwaMem` module provided by Lumpy and discordant reads with Samtools (version 1.5). Next, the BAMs were sorted using Samtools. We used `Lumpyexpress` with default parameters after the pre-processing stage, including the `-P` label. Finally, we used `SVtyper`¹¹⁹ to genotype. We ran it according to the developer recommendations. After the genotyping step, we filtered out those variants whose quality was ≤ 20 . Also, we discarded variants with SVTYPE = BND and if the two chromosomes reported by the variant caller were the same. We used 12 CPUs to extract discordant and split reads and 24 CPUs and one CPU for `Lumpyexpress` and `SVtyper` tools, respectively.

3.2.2.5. Whamg (version v1.7.0-311-g4e8c)

Wham⁷⁸ is a variant caller used to detect DELs, DUPs, INSSs, and INVs. The detection strategy is SR and DR, plus an additional Machine Learning algorithm (ML) to classify the SVs by type. We used the Whamg version as recommended by the developers, and with default parameters. Finally, we used SVTyper to obtain genotypes, as Whamg cannot produce them. To run Whamg, we parallelised the execution by 48.

3.2.2.6. SvABA (version 7.0.2)

SvABA⁷⁹ classifies all SV types by breakpoint orientation using the *de novo* Assembly (AS), SR, and DR strategy. We ran SvABA in germline mode, as recommended by the developers. We applied an internal validation of the SV type reported by SvABA in the *in-silico*, and it was discordant. Thus, we did not use the SV type information to catalogue the SVs. To improve their performance, we parallelised the execution by 16 CPUs.

3.2.2.7. CNVnator (version v0.3.3)

CNVnator¹⁰⁶ was used to discover Copy Number Variations (CNV), such as DEL and DUP. This caller applies the RD strategy, and detects CNVs larger than 200 bp. We ran the CNVnator following all the steps recommended by the developers, using a bin size of 100, which corresponds to the read length of the sequencing. Finally, we converted the .root file to VCF using a Perl script, "cnvnator2VCF.pl", from the CNVnator toolkit. We used 12 CPUs to execute this tool for memory reasons.

3.2.2.8. Popins (version damp v1-151-g4010f61)

Popins⁸⁰ was used to discover *de novo* insertions (INSSs). This tool uses the AS strategy, improving the detection of exogenous sequences, and is able to detect INSSs ≥ 100 bp. To run this variant caller, we followed different steps: 1) assembly, 2) merge (skipped due to we used just *in-silico* sample), 3) contigmap (generate the supercontigs file), 4) place-refalign, 5) place-splitalign, 6) place-finish, and 7) genotyping. All steps were done using the default parameters.

The NONANCHOR variants were discarded as recommended by the Popins developers. In all steps, we used 48 CPUs to reserve all the memory of a node for each execution and to parallelise steps 1 and 3 by 48.

3.2.2.9. MELT (version 2.1.4)

MELT¹⁰⁷ was used to detect Mobile Element Insertions (MEIs) such as Transposons (TRPs). The strategy applied to identify the TRPs were SR and DR. This variant caller was executed in a SINGLE mode. The following flags were used to run MELT (-bamfile -c -e -t -h -r -k -w). We adapted the MELT execution to each type of TRP, divided by ALUs, LINE1s, and SVAs. MELT requires high computational resources, so we used 24 particular CPUs with high memory.

3.2.2.10. ViFi (no version reported)

ViFi¹⁰⁸ was designed to detect viruses (VIRs). This tool uses SR and DR to detect the viruses inserted in the human genome. Before running ViFi, we applied two pre-processing steps.

1) We build a new database that incorporated all viral sequences from NCBI “ftp://ftp.ncbi.nih.gov/refseq/release/viral/”. First, we downloaded (at date 07/01/19) and merged all the files (*.genomic.fna). Second, we adapted the “merge.fna” file to the .fas file format, using a script developed in-house. Finally, we included the reference genome (hg19) provided by ViFi to obtain the final database. We indexed this file using the BWA tool. 2) We converted the *in-silico* BAM file to a FASTQ file using Biobambam2 (version 2-20.65) to spread the FASTQ files by pair-ends 1 and 2, respectively. Before executing ViFi, we imported python (version 2.7.13), BWA (version 0.7.15), and Samtools (version 1.5). Then, we ran the ViFi in a basic mode, as recommended by developers. To run ViFi, we required 24 particular CPUs of high memory and more than ten days to process the *in-silico*. This tool requires a high amount of computational resources; for this reason, we discarded it for variant detection in the GCAT samples.

3.2.2.11. VERSE (VirusFinder2) (version 2.0)

VERSE¹⁰⁹ was used to detect VIRs in the human genome. The strategy to identify the viral integrations is a combination of SR, AS and DR. We imported the following programs to run VERSE properly: CREST (version 1.0), SVDetect (version r0.8_threads), BLAST (version 2.6.0), BOWTIE2 (version 2.3.2) and BWA (version 0.7.15). Then, we generated a viral database to run VERSE; for this reason, we downloaded the same files as ViFi (section 3.2.2.10), and we combined and produced the FASTA file using a python script developed in-house. Finally, we generated all remaining data of our database using the following command of BLAST:

```
makeblastdb -in ncbi_virus.fa -dbtype nucl -out ncbi_virus
```

To run VERSE, we modified the template-config.txt file. We included the paths of FASTQ files of *in-silico* (described in section 3.2.2.10), as well as all other remaining routes needed to run the variant caller properly. We used 24 CPUs to run VERSE in MareNostrum4.

Due to problems in the execution of VERSE in real samples, we discarded it.

3.2.2.12. Genome Strip (Version 2.0)

Genome Strip⁹⁸ was designed to detect DELs, DUPs, and multiple Copy Number Variants (mCNVs) in cohorts. This tool uses various strategies such as SR, DR, and RD. This variant caller requires different tools and a specific computational environment: a) Java version 1.7, b) R tool (version 3 or newer), c) Samtools and Htslib, d), and the LSF environment. We followed three steps to run the Genome Strip: 1) SVPreprocess, 2) SVDDiscovery, and 3) SVGenotyper. All steps were run using default parameters, as developers recommended. We used 48 CPUs to run each step of the Genome Strip against the *in-silico*.

We could not run the package designed to detect DUPs and mCNVs in the *in-silico* sample due to incompatibilities between the Supercomputer Nord3 and this package. In addition, when we ran the Genome Strip with real samples, we were not able to finish the executions for the detection of DELs, DUPs, and mCNVs. For this reason, we discarded it from the project.

3.2.2.13. Pamir (version 1.2.2)

Pamir¹⁸² was designed to detect INSSs. This tool combines the SP, DR, AS, and One End Anchored (OEA) strategies. We ran Pamir in default mode, including the flag `-p`, to report the name of the sample, that we process. Finally, the output was genotyped using a python script provided by Pamir developers:

```
python genotyping.py input.vcf genome.fa.masked fastq_pair_1.fq.gz
fastq_pair_2.fq.gz name_file 0 1 path_all_files_generated_by_pamir 1000 1
```

It took us 48 CPUs for more than ten days to analyse the *in-silico* sample, also a test with a real sample, generating a size file of 881Gb after 22 days. Given these high computational requirements, we discarded this tool from future analyses.

3.2.2.14. AsmVar (version 2.0)

AsmVar¹⁸³ is able to detect DELs, DUPs, INSSs, INVs, and TRAs. This tool uses the AS strategy to detect all SVs. Before variant detection, we realigned the *in-silico* using the LAST aligner (*de novo* aligner), as developers recommended. This software required the following steps: 1) Variant detection, 2) Altalignemt, 3) Genotyping, 4) RecalibrationVg, and 5) a Variant Quality Score Recalibration (VQSR). AsmVar was further discarded from future steps due to its high memory requirements.

3.2.3. Recall, Precision, and F-score

Recall, precision, and F-score are the metrics used to determine the accuracy of detecting variants for each caller in our *in-silico* sample. These metrics were calculated using True-Positive (TP), False-Negative (FN), and False-Positive (FP) variant calls. These parameters were evaluated for each SV type, SNVs, and indels independently.

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

For indels and SVs, we used these metrics to evaluate the breakpoint resolution reported by the callers. Besides, for indels and SVs, we used recall and precision to analyse the accuracy to report the breakpoint correctly.

We applied a variant filtering to the whole VCFs obtained by each software. **We selected the variants which PASS all variant caller filters.** We also discarded all variants detected in 1) Decoy sequences, 2) Y chromosome, 3) MT chromosome, and 2) variants with genotypes reported as "0/0" or "./."

3.2.3.1. Evaluation of breakpoint-error for Indels

Indel detection usually occurs at base-pair resolution. However, the larger the length of the variant, the worst is the resolution of breakpoint reported by variant callers. For this reason, we studied the breakpoint-error using as a gold standard the *in-silico* sample. First, we normalised the outputs obtained by each caller (section 3.2.1) following the recommendations of the Global Alliance for Genomics and Health (GA4GH) (detailed documentation in section 3.6.1.4). We compared the outputs with the *in-silico* at a base-pair resolution to obtain TP and FP rates.

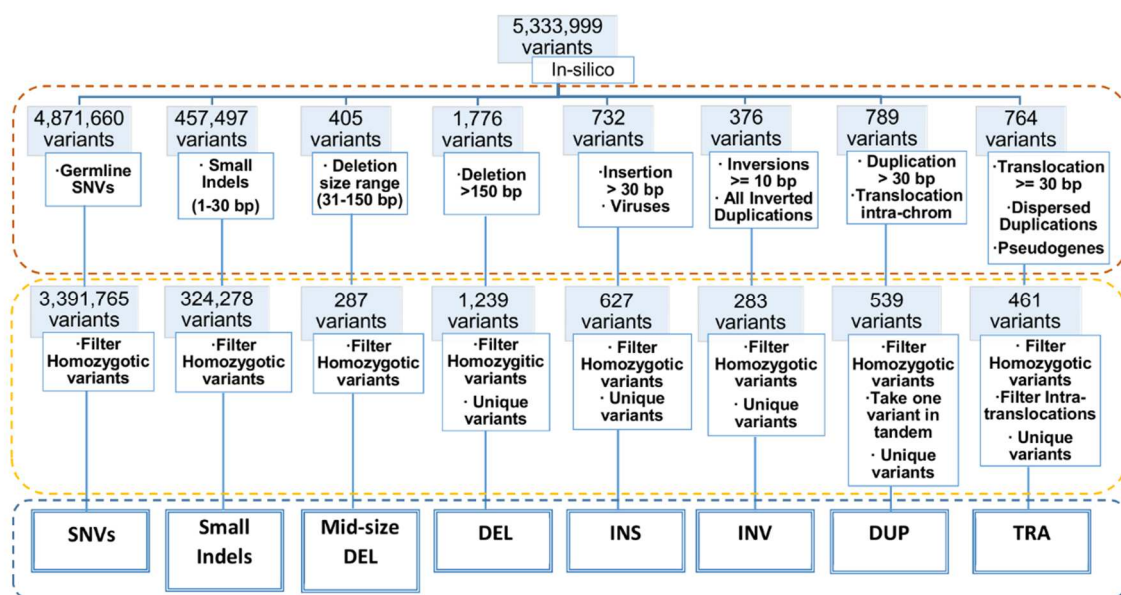
Second, we calculated recall and precision based on the indel size. When indels larger than 30 bp, we grouped them by batches of 10 bp. Finally, using R (version 3.3.1), we plotted the distribution of recall and precision based on the Indel size (Figure 15).

Variants between 1 and 30 bp were classified as small Indels, with no breakpoint-error. When the size was between 31-150 bp, we catalogued variants as mid-size deletions, for which we reported an associated a breakpoint-error of ± 10 bp. Insertions were not catalogued due to the size was not reported by variant callers.

3.2.3.2. The categorisation of *in-silico* variants

Before performing the analysis of recall, precision, and F-score of all variant callers, we used the *in-silico* sample list to compare the outputs of each algorithm. We divided all the genome variability of the *in-silico* by variant type, and we used this classification to calculate the metrics.

To avoid duplicates from each variant type, we filtered out from the *in-silico* list the variants with the ‘‘Homozygotic’’ flag and taken the duplications with the ‘‘tandem’’ flag once. Besides, some callers could fail to report the type of SV, but not their breakpoint, so we grouped these inconsistencies into the SV type group reported by variant callers. In Scheme 2, we illustrated how we arrange and filter different variants by type:



Scheme 2. Classification and filtering of the *in-silico* variants by type of variant. Each dashed line colour (- - - ; - - - ; - - -) indicates the grouping and filtering steps, and different variant type respectively. The **white boxes** of grouping and filtering, indicate the criteria used to obtain the unique variants of each variant type. The **blue boxes** shows the number of variants after grouping and filtering step. In the grouping step, we also included those variants where the callers detect the breakpoint, but fail the variant type.

Of note in the *de novo* insertion (INS) group, small variants from other variant types were further included; these additional variants were 32 Duplications, 2 Inversions, 1 Pseudogene, 47 Translocations and 37 small insertions (indels) (all numbers are obtained after filtering homozygous and duplicate variants), which were misclassified by variant callers, and were therefore added to the INS group, obtaining a total of 727 variants for INS benchmarking.

3.2.3.3. Determination of SV breakpoint-error

Reporting the position of a structural variant with high definition in the genome is one of the challenges in the calling step. As the length of variants increased, discrepancies between callers reporting of the same variant in the same position also increased. For this reason, we studied the error in breakpoint definition associated with each caller, in order to know which algorithms were able to report the concordant variants.

We evaluated which breakpoint-error was the most accurate to use, based on the F-score metric and type of SV. If we used large ranges, we could combine different SVs wrongly. The breakpoint-error allowed us to establish a range of positions where the variant can be considered the same among variant callers; even more, it gave information about how accurate a caller is reporting a breakpoint position correctly. The thresholds were calculated as follows:

$$Range_{lower_point} = Position\ reported\ by\ caller - window\ breakpoint_error$$

$$Range_{upper_point} = Position\ reported\ by\ caller + window\ breakpoint_error$$

We used as a reference the positions of variants in the *in-silico*. We evaluated different breakpoint-errors for each algorithm and SV type (10, 20, 50, 100, 200, and 300 bp) to know which produced the best F-scores. The breakpoint-error has been selected according to sequencing particularities, coinciding the largest with the insert-size of the *in-silico* (300 bp).

The outputs obtained from “section 3.2.2” (except MELT, ViFi, VERSE, Genome Strip, and AsmVar) were used to generate the range intervals for each variant. We considered a variant as a TP if the position in the *in-silico* overlapped with the breakpoint-error of the caller. Otherwise, it was classified as an FP. We repeated the same analysis for each SV type, and all breakpoint-error considered. Table 16 shows which window breakpoint-error was selected for each variant caller and SV type.

3.2.3.4. Evaluation of variant caller metrics

We estimated and evaluated recall, precision, F-score for each variant caller, and variant type (except for MELT, ViFi, VERSE, Genome Strip, Pamir, and AsmVar (see section 3.2.2)). This section describes how we calculated the metrics for SNVs, small indels, and SVs.

3.2.3.4.1 SNV and Indel metrics

For each variant caller, we used as a gold standard the list of *in-silico* variants (Scheme 2). We classified the variants as TP according to the following criteria: (i) If the chromosome reported by algorithms were the same as the *in-silico*; (ii) If the positions overlapped at a base-pair resolution between the algorithms and gold standard; (iii) The Reference and the alternative alleles were the same. All three criteria had to be met to obtain a TP classification. In all the other cases variants were classified as FP. All the variants not detected by variant callers but present in the *in-silico* were classified as FN. Finally, we calculated the metrics, as mentioned in section 3.2.3.

We generated four datasets (including TP and FP to calculate recall, precision and F-score), following the same criteria exposed before. Two datasets included SNV variants and the other two the indel variants. The difference between both SNV and indel datasets were the number of variant callers included, were two used the outputs from Haplotype caller, Deepvariant,

and Strelka2, and the other two included all callers mentioned in section 3.2.1. Besides, all datasets contained the *in-silico* information. These datasets will be used to validate the Logistic Regression Model (section 3.4.1.1).

3.2.3.4.2 SV metrics

To evaluate the metrics of each algorithm and SV type, we used the *in-silico* sample as the gold standard. The criteria to classify the variants as TP were: 1) The chromosome reported by algorithms coincided with the *in-silico*; 2) The SV type was the same as the *in-silico*; 3) The *in-silico* position overlapped in the breakpoint-error of variant callers evaluated in section 3.2.3.3; 4) The length of the variant reported by algorithm was 80% Reciprocal Overlap (RO) with the variant length of the *in-silico*. If one or more of those criteria was not met, the variant was classified as FP. The variants in the *in-silico* sample that were not detected by variant callers were classified as FN. Then, recall, precision, and F-score were calculated as previously described in section 3.2.3.

Following the criteria mentioned previously, we generated a dataset for each type of SV (just TP and FP to calculate recall, precision and F-score), combining all outputs from callers selected. Besides, the variant list of the *in-silico* has been included in each dataset. These datasets were used to create/validate the SVs Logistic Regression Model (LRM) (section 3.4.2.1).

3.2.3.4.3 Recall and Precision to detect SVs by size

Size is a determinant factor for detecting SVs. The larger the SV size, the lower the mapping quality of the reads in the region; this leads to misinterpretations and increases false-positive detections. Therefore, for each SV type, we calculated the F-score by intervals of length ((30-50], (50-75], (75-100], (100-125], (125-150], (150-300], (300-500], (500-1000], (1000-2000], (2000-3000], >3000) and SV type. This analysis allowed to include the variant length as a covariate in the Logistic regression model (section 3.4) to filter out the potential false-positive detections (Figure 16).

3.2.3.5. Evaluation of genotype errors

The callers reported a genotype for each variant. This parameter is important to generate a reference panel because a good genotype will improve imputation. For this reason, we evaluated the genotype error of variant callers for each SV type, using the *in-silico* sample as a gold standard.

For the *in-silico* files, we reported as “1/1” all the variants with the “Homozygous” flag (Table 2) and all the other variants as and “0/1”. Finally, we calculated the genotype error of each variant caller for heterozygous, homozygous, and the combination between them, using the following formula:

$$\text{Genotype error} = \frac{100 - \text{Correct Genotypes}}{\text{All Genotypes}} * 100$$

All variants with a called genotype matching the *in-silico* sample were defined as “Correct Genotype”. “All Genotypes” refers to all the genotypes of variants introduced and reported by the variant caller. The same formula was applied for heterozygous (0/1) and homozygous alternative (1/1) genotypes individually.

3.2.3.6. Selection of a strategy to construct all the BAM files of GCAT samples

To improve variant detection, we evaluated and selected the best strategy to obtain the BAM files for the GCAT samples. Table 4 illustrates the three analysed strategies analysed:

| | Human genome | GATK Best Practices |
|-------------------|--------------|---------------------|
| Strategy 1 | Hs37d5 | Not applied |
| Strategy 2 | Hs37d5 | Applied |
| Strategy 3 | Hg19 | Not applied |

Table 4. Strategies that were explored to obtain the BAM files of the GCAT project. Each colour is related to Figure 19 and Figure 20 strategies.

The FASTQ files of the *in-silico* sample were used to construct a BAM file for each strategy. First, these files were created following the steps explained in section 3.1.2, but considering Table 4 differences. Second, we ran all the variant callers explained above (except for VarScan2, Platypus, Genome Strip, AsmVar, Melt, VIFI, Verse, Pamir, and Popins) using each generated BAM file. Finally, we calculated recall, precision, and F-score (section 3.2.3), and we applied the Wilcoxon test to analyse if there were significant differences between these strategies (Figure 19, Figure 20).

3.3 Genome in a Bottle sample

The benchmarking analyses explained in the previous section were performed on an artificial sample built *in-silico*. To further explore the behaviour of algorithms in a real sample, we used the NA12878¹²³ from the Genome in a Bottle Consortium (GIAB) to validate the calling of SNVs and indels. We could not do a similar evaluation for the calling of SVs, because there is currently no real sample to validate whole SV types.

Firstly, we downloaded the down-sampling 30X BAM file of NA12878 (RMNISTHS_30xdown sample.bam) from ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NIST_NA12878_HG001_HiSeq_300x/. Secondly, we converted the BAM file to a FASTQ file using Biobambam2 version 2-2.0-65, and we applied the strategy2 (section 3.2.3.6) to reconstruct it. Finally, we ran all callers from section 3.2.1 on the NA12878 sample.

To evaluate recall, precision, and F-score, we used as a gold standard the “HG001_GRCh37_GIAB_highconf_CG-IIIIFB-IIIIGATKHC-Ion-10X-SOLID_CHROM1X_v.3.3.2_highconf_PGandRTGphasetransfer.vcf.gz” file of the NA12878 sample. We downloaded it from ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/latest/GRCh37/, which contained all the variants validated by the GIAB consortium. We normalised this file using the pre.py tool from the Global Alliance for Genomics and Health (GA4GH), obtaining the same representation in the vcf for SNVs as well as indels. The accuracy metrics were evaluated following the steps of 3.2.3 section.

Finally, we generated four databases (divided by SNVs and indels), two with Haplotype Caller, Deepvariant and Strelka2, and others two combining all algorithms from section 3.2.1. To create them, we merged all the outputs from callers and gold standard file following the criteria of

section 3.2.3.4.1. Those datasets were used to create a Logistic Regression Model for SNVs and indels (section 3.4.1.1).

3.4 Increasing accuracy detection using a machine learning algorithm

One of the variant calling challenges is increasing precision and recall, mainly for SVs. For this reason, we developed a Logistic Regression Model (LRM) for each variant type to improve the performance of the calling through different patterns and discriminative variables. The LRM was suitable because of the small number of features and the large number of variants considered. Furthermore, the LRM allowed to estimate parameters indicating the accuracy of each caller and to make predictions based on the sum of the estimates in the logistic regression equation. This section explains the strategy followed to create the LRM for each variant type.

3.4.1. Logistic Regression Model for SNVs and small Indels

3.4.1.1. Training and Testing of the model

The LRM for SNVs and INDELS was trained using the GIAB sample (section 3.3) and tested using the *in-silico* sample (3.2.3.4.1). The input of the LRM was a merged dataset of the VCF outputs from the callers following the same criteria to calculate the TP, FP, and FN section 3.2.3.4.1. We developed a specific LRM for SNVs and Indels independently using the R software (version 3.3.1) and the ISLR package. The function to fit the LRM was:

```
(PASS ~ Deepvariant + Haplotype caller + Strelka2, data = database, family = "binomial")
```

The outcome of the LRM was a binary variable (PASS), indicating if the variant was predicted to be present in the GIAB sample. The independent predictor variables were the genotypes reported by the variant callers indicating their detection pattern.

3.4.1.2. Genotype reported by LRM for SNVs and small Indels

The consensus genotype between Haplotype caller, Deepvariant, and Strelka2 was considered as the genotype of the LRM.

3.4.2. Logistic Regression Model for SVs

3.4.2.1. Training and Testing of the LRM for SVs

For SVs, the LRM was trained using 10-fold cross-validations for a random subset of variants (70%) from the *in-silico* and was tested using the remaining subset of variants (30%) of the *in-silico*. The input of the LRM is a merged dataset of the VCF outputs from the callers following the same criteria to calculate the TP, FP, and FN (section 3.2.3.4.2). We developed a specific LRM for each SV type independently using the caret (version 6.0-85) and e1071 (version 1.7-3) R packages. The function to train the LRM was:

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

The outcome of the LRM was a binary variable (PASS), indicating if the variant was predicted to be present in the *in-silico*. Next, for each SV type, we generated a specific LRM. We studied which

independent variables were discriminative versus non-discriminative or redundant. Table 5 shows the variables included for each SV type to fit the best LRM.

| Mid-size Deletion | DEL | INS |
|---|---|--|
| Deepvariant, Haplotype Caller, Strelka2, Manta, Whamg, Delly2, Lumpy, Pindel, SvABA, Num of callers detected, Reciprocal overlap ≥ 0.8 | Manta, Whamg, Delly2, Lumpy, Pindel, SvABA, CNVnator, Variant size, Strategy | Manta, Whamg, Delly2, Pindel, SvABA, Popins, Haplotype caller, Strelka2, Num of callers detected |
| INV | DUP | TRA |
| Manta, Whamg, Delly2, Pindel, SvABA, Lumpy, Variant size | Manta, Whamg, Delly2, Lumpy, Pindel, SvABA, CNVnator, Variant size, Reciprocal overlap ≥ 0.8 | Manta, Delly2, Pindel, SvABA, Lumpy |

Table 5. Independent predictor variables used to train the Logistic Regression Model for each SV type. Variant caller variables are the genotypes indicating the presence or the absence of the variant. The **variant size** is divided into different ranges, example (500-1000, 1000-2000 bp, 2000-3000, > 3000), each range is a predicted variable. The **number of detection callers**, indicates the number of different algorithms detecting each variant, (for example 2, 3, 4-5), each number is a predicted variable. The **reciprocal overlap ≥ 0.8** , is a numerical variable indicating the size of the overlapping between two different algorithms. The **strategy**, is the number of different strategies used by callers to detect a variant.

3.4.2.2. Strategy to report the position and length of variants by LRM

The software methodology and variant type are factors associated with breakpoint and length accuracy of called variants. We used the *in-silico* sample as a gold standard to evaluate the accuracy of the algorithms to report breakpoints and lengths by SV type. Then we merged the different variant caller outputs in a multi-sample VCF consensus file.

The breakpoint definition of the consensus regions was determined by 1) the precision of each caller to report the position within a breakpoint-error of ± 10 base pairs, and 2) the number of variants detected by each algorithm (Supplementary Figure 1). Table 6 describes the priority to report the breakpoint by SV type in the consensus multi-sample VCF. On the other hand, the variant length was consistent between tools; thus, we considered the median length reported by the callers, excluding CNVnator, which got the worst predictions (Supplementary Figure 2).

| Structural Variant Type | Order |
|-------------------------|--|
| Mid-Deletion/Deletion | Pindel > Whamg > Delly2 > Manta > Median of remaining callers |
| Insertion | Pindel > Delly2 > Strelka2 > SvABA > Manta > Median of remaining callers |
| Duplication | SvABA > Pindel > Delly2 > Whamg > Median of remaining callers |
| Inversion | Lumpy > Pindel > Delly2 > Median of remaining callers |
| Translocation | Manta > Median of remaining callers |

Table 6. Order to report the breakpoint. The most precise algorithm was used to report the breakpoint by the merge algorithm. When the breakpoint-error was large, we reported the median.

3.4.2.3. Genotype reported by LRM for SVs

The LRM predicted the variant as a true-positive or false-positive. However, it still needed to define the genotype. The genotype for each SV type was defined following specific strategies. Below we show the strategy applied.

1. Mid-size Deletion and DEL strategy

As the genotype error between callers was relatively low (Supplementary Figure 4A), we reported the most frequent genotype between them. When the consensus genotype was not possible to report, we included a missing “./.”.

2. INS strategy

We reported the most frequent genotype between algorithms. For variant sizes between 30 and 50 bp, we did not use Manta and Whamg detections, because they do not report any genotype. When the consensus genotype was not possible to report, we included a missing “./.”.

3. DUP strategy

The genotyping error of callers is high for this type of SVs (Supplementary Figure 4E); for this reason, we developed a genotyping method using the BAM information of the *in-silico* sample. To obtain the **total coverage** of each DUP, we reported the median from all the reads that covered twice the length of the variant, in both upstream and downstream directions of the breakpoint reported by the LRM. The **altered reads** were obtained from the breakpoint reported by LRM as follows. We counted the split reads in a window of ± 10 bp, discarding the Hard-clipped reads and those containing INSs or DELs in the CIGAR. Finally, we calculated the proportion of altered reads over the total coverage.

$$\frac{\text{Altered reads at breakpoint}}{\text{Total Coverage region}} * 100$$

If the proportion of altered reads was ≤ 20 %, we genotyped as homozygous reference (0/0) (no variant present); if the proportion was between 0.20 and 0.80, we genotyped as heterozygous (0/1), and if the proportion was ≥ 0.80 , we genotyped as homozygous alternative (1/1).

4. INV strategy

The genotype of INV reported by LRM was selected based on the order of best callers to genotype, excluding SvABA, due to the large error rates (Supplementary Figure 4D). The order is shown below:

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

5. TRA strategy

The genotype error displayed by the callers in TRA detection was high (Supplementary Figure 4F). Then, we re-genotyped the variants using the BAM information of the *in-silico*. **The total coverage** was obtained by counting all the reads covering the breakpoint reported by the LRM in a window of 4bp. **The altered reads** were obtained by discarding (i) Hard-clipped reads, (ii) counting all reads with map quality ≥ 20 , and (iii) reads with a label different from 151M in the CIGAR. Finally, to report the genotype, we applied the same formula described for DUP.

3.4.2.4. Filtering out the SVs following the GoNL strategy

The GoNL project built a panel of genetic variability including SVs¹⁶¹. As a filtering strategy, GoNL retained the SVs that were detected by at least two variant callers. We extended this strategy to three and four different algorithms and calculated recall and precision, considering all these criteria (Figure 17, Supplementary Figure 3).

3.5 The GCAT project

The previous sections describe how we succeed in understanding, 1) the best way to construct the BAM files, 2) selecting and improving the performance of callers use it to detect SNVs, Indels, and SVs, and 3) building for each variant type a Logistic Regression Model to increase recall and precision of calling. We then applied this knowledge to analyse samples from the GCAT project. This section will explain the characteristics of the GCAT samples, the BAM files construction, the quality control, and coverage analysis.

3.5.1. **Genomic data features**

The GCAT project is designed to associate genetic and environmental factors to complex diseases, such as diabetes or respiratory diseases. In this context, 19,267 GCAT volunteers were recruited from the general population of the Northeast region of Spain (Catalonia) in different areas such as coastal, mountain, rural, or urban areas. The **unrelated participants** had an age between **40-65 years**¹⁷³ with 16% **non-Caucasian**, mostly of American-Hispanic origin. The complete protocol for the settings of the GCAT cohort is detailed in Obon-Santacana et al.¹⁷³. Table 7 describes how the sampling has been done:

| Study purpose | Fraction sample | Vacutainer tube | Volume mL | Transport T°C | Time to PMPPC | Aliquots n (T°C) | Control assay |
|--------------------|--------------------------------|-----------------|-----------|---------------|-----------------|------------------|---------------------------|
| Genomic/epigenomic | Buffy coat | EDTA | 10 | 4 | max 24 hours | 2 (-80) | SNP array, qPCR, PCR, STR |
| | Highly concentrated buffy coat | Blood bag | 480 | 18 | max 48 hours | 2 (-80) | SNP array, qPCR, PCR, STR |

Table 7. Description of sampling for Genomic analysis in the GCAT project. This table is adapted from Obón-Santacana et al.¹⁷³.

For this thesis, we used SNP array (of 5,459 volunteers) and Whole Genome Sequencing (WGS) data (808 volunteers) from the genomic analyses. Below we detailed the features of each dataset.

| Study purpose | Number of participants | Fraction sample | Platform | Machine | Analysed |
|-------------------------------|------------------------|-----------------|--|------------------------------------|---|
| Genotype | 5,459 | Buffy coat | Infinium Multi-Ethnic Global (MEGAEX2) array | HiScan confocal scanner (Illumina) | 2×10 ⁶ SNPs, InDels |
| Whole-genome Sequencing (WGS) | 808 | Buffy coat | Illumina TruSeq PCR free/Illumina paired-end SBS | HiSeq 4000 sequencer (Illumina) | 30X coverage Read length 150 bp Insert size 600 |

Table 8. Summary of Genomic data for this thesis. This table is adapted from Obón-Santacana et al. paper¹⁷³.

3.5.2. BAM file generation

Before analysing the samples obtained from the WGS platform, we pre-processed the FASTQ files of 808 volunteers to obtain the BAM files, which we used for downstream analyses. The BAM files were created following the strategy2 as described in section 3.2.3.6, but with some differences. In this section, we will explain these particularities.

3.5.2.1. Selection of the Reference Genome, based on the sample gender

The majority of sequencing projects align all reads from FASTQ files against a Reference Genome (RG) to obtain the BAM files. The RG includes chromosome Y (chrY), so in our opinion, the variant detection in chromosome X for female samples was not well-curated because they do not contain in their karyotype the chrY. For this reason, we wanted to know the effect of chrY in female and male samples in the read alignment and variant detection.

The GCAT samples included 409 Female and 399 Male samples. We selected two male (JID259, JID439) and two female samples (JID250, JID297) to perform this study. For these samples, we constructed the BAM files following the strategy2 described in section 3.2.3.6. We repeated the process twice for each sample, using the RG with and without chromosome Y. We removed the chromosome Y from the hs37d5 decoy RG, using an in-house script. Finally, we counted all reads mapped on chromosome X, and we compared the results of these two RG.

We ran Haplotype caller to detect SNVs and Indels, as mentioned in section 3.2.1.1, for the different BAM files created. Finally, we evaluated the differences between them, using variants with genotype 0/1 and 1/1.

3.5.2.2. Data structure of the WGS samples of the GCAT project

The FASTQ files of the 808 GCAT samples were generated in batches. Four batches included 192 samples, one batch included 24 samples, and the remaining batch included 16 samples. Each sample contained multiple LANES grouped by the Multiplex index. Furthermore, the paired-end was generated in separate files:

FASTQ paired-end 1

H52MCDSXX_1_60idt_1.fastq.gz

H52MCDSXX_2_60idt_1.fastq.gz

FASTQ paired-end 2

H52MCDSXX_1_60idt_2.fastq.gz

H52MCDSXX_2_60idt_2.fastq.gz

3.5.2.3. The construction of BAM files in the GCAT project

The BAM files were constructed using the hs37d5 reference genome (without chrY in female samples) and following the GATK Best Practices. Firstly, we aligned each pair-end FASTQ file. Then, we merged the BAM file for each sample according to the LANE. A scheme of the procedure is illustrated in Figure 11:

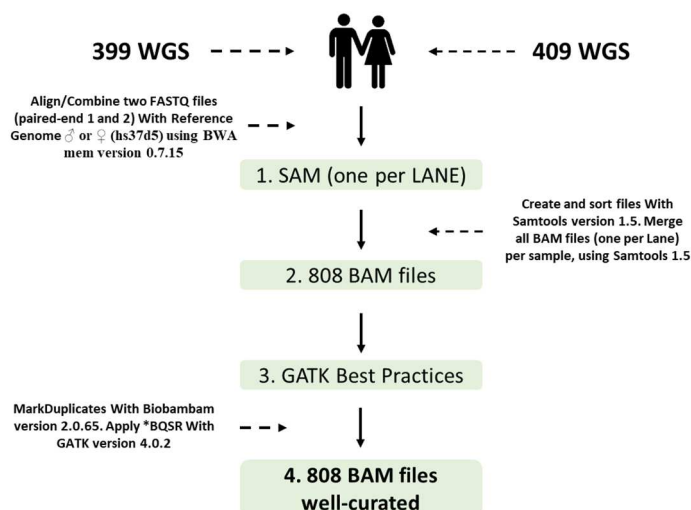


Figure 11. Steps followed to construct the BAM files of the GCAT project. The reference genome used to create the SAM files depended on the sample gender. We cleaned the BAM files following the Best Practices of GATK. Detailed documentation in section 3.1.2.

* BQSR: Base Quality Score Recalibration

3.5.3. Quality Control (QC) of the BAM files and sample ancestry analyses

Checking the quality of the 808 BAM files allowed us to determine if the samples were well constructed or contained inter-sample contamination produced from a bad manipulation of DNA material. Obtaining a good quality of alignment was necessary to improve the variant calling, **decreasing the false-positive detections**. In addition, as we mentioned in section 3.5.1, some GCAT samples had an American-Hispanic origin, so to create an Iberian reference panel, we filtered out those samples genetically non-Iberian related. Finally, we evaluated the level of relatedness between samples.

3.5.3.1. Alignment quality

To evaluate the alignment quality, we ran Picard (version 2.18.11), Biobambam (version 2-2.0.65) and Alfred⁷¹ (version 0.1.16). The statistics evaluated are described in Table 9:

QC metrics and Inclusion criteria

Fraction purified reads > 0.90

Fraction reads aligned in pairs > 0.95

0.495 < Strand Balance < 0.505

250 bp < Mean insert size < 350 bp

Standard deviation of insert size < 50 bp

*Fraction of duplicated reads < 0.1

27X < Mean Coverage < 37X

+ Fraction of paired reads mapped in the same chromosome > 0.88

Table 9. Alignment Quality control metrics.

* Biobambam2 tool; + Alfred tool

In total, 75.4% of samples fulfilled the QC metrics described in Table 9. 24.6% of samples which no pass the thresholds were further inspected. Besides, the variant callers were executed correctly across samples, determining that they passed the QC. However, the **JID748 sample** was discarded from downstream analyses due to irregularities observed when executing the variant callers. Figure 12 shows some descriptive measures of the samples based on QC metrics:

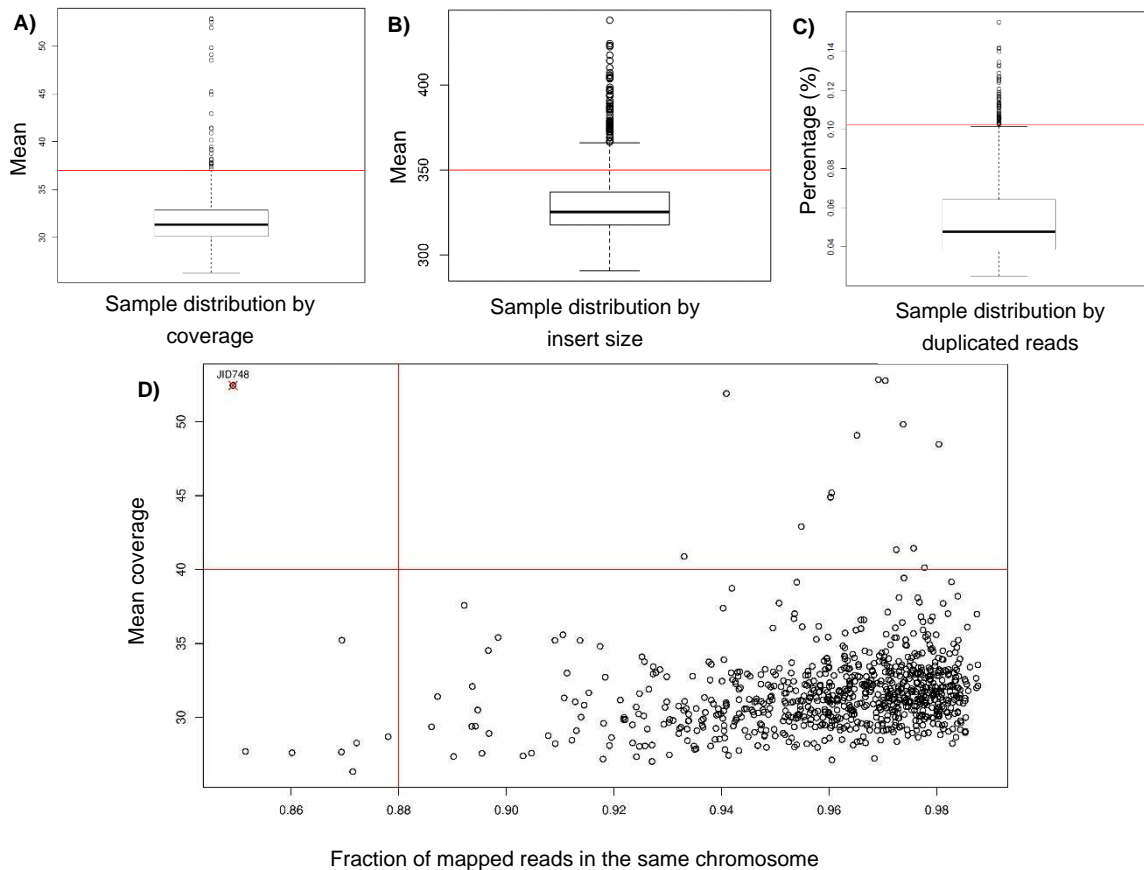


Figure 12. Sample distribution based on QC metrics. **A)** Mean coverage. 29 samples exceeded the threshold. **B)** Insert size distribution. 112 samples exceeded the threshold. **C)** Fraction of duplicated reads. 58 samples exceeded the threshold. **D)** Fraction of mapped reads in the same chromosome. 7 samples exceeded the threshold. One sample exceeded the threshold of coverage mean > 40 and the fraction of mapped reads in the same chromosome. This sample was discarded for variant calling irregularities.

The (—) indicates the Table 9 thresholds.

3.5.3.2. Contamination analysis

We used VerifyBamID¹⁸⁴ to determine inter-sample contamination or swapped ID samples. We used this tool with genotyping data (array-based) and array-free approaches. The thresholds to determine if the sample was contaminated are shown in Table 10. For this purpose, we selected 570 GCAT samples from the array data that were also sequenced. Using this array data and all 807 WGS BAM files, we ran VerifyBamID with `--best --ignoreRG --maxDepth 30 --precise` arguments. All the metrics obtained from VerifyBamID showed that none of the GCAT samples were contaminated (Figure 13).

| Threshold contamination array-based | Threshold contamination array-free |
|---|---|
| [CHIPMIX]>> 0,02 and/or [FREEMIX] >> 0,02 | [FREEMIX] ≥ 0.03 and [FREELK1]-[FREELK0] is large |

Table 10. Thresholds to determine the contamination/ID swap from GCAT samples. When the array data was provided, if $\geq 2\%$ of non-reference bases were present in reference sites, the sample was contaminated. The array-free strategy used the allelic frequency estimations to determine if a sample was contaminated or not.

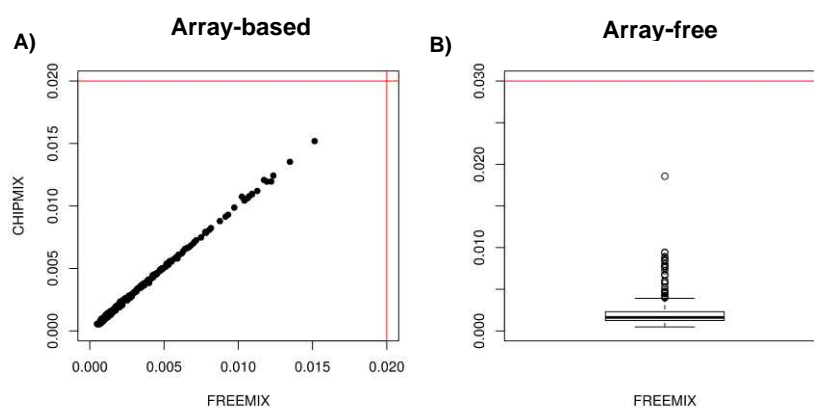


Figure 13. Contamination analysis of GCAT samples. **A)** Contamination distribution of 570 samples with array and Whole Genome Sequencing analysis. **B)** Contamination Distribution of 237 samples with Whole Genome Sequencing analysis.

The (—) indicates the Table 10 thresholds.

3.5.3.3. Population structure using reference ancestries

We evaluated the population structure of GCAT cohort, based on Principal Component Analysis (PCA). As we mentioned in section 3.5.1, 16% of the GCAT samples were of American-Hispanic origin, so we discarded from the 807 WGS samples those who were non-Iberian representative.

Firstly, we ran the Haplotype Caller tool and selected the PASS filter variants from the VCF file. Secondly, we used PLINK (version 1.90b6.7 64-bit) to keep ~ 1 million of SNVs with a Minor Allele Frequency (MAF) $> 1\%$ (discarding rare and monomorphic variants) and Linkage Disequilibrium (LD) $r^2 < 0.2$ (obtaining independent variants). Finally, we applied two PCAs based on a different population of known ancestries: 1) the first PCA discarded the non-European samples from GCAT, using the SNVs from 1000G and GCAT. We filtered out **16 GCAT samples** (Figure 21A, Figure 21B) according to the high euclidean distance of PC1, PC2 and PC3 from the

center of the distribution of GCAT samples. 2) The second PCA was performed to get only the Iberian samples. We used the webserver from the LASER¹⁸⁵ project (<https://laser.sph.umich.edu/>) with variants from European samples of the 1000G and the Population Reference Sample (POPRES¹⁷⁷) projects, and GCAT samples. **2 GCAT samples** were discarded, according to the mean \pm 4sd criteria (Figure 21C). After performing these two PCAs, we discarded **18 samples** for genetic discrepancies of Iberian ethnicity.

3.5.3.4. Identity by Descent analysis (IBD)

Discarding samples with a high level of relatedness allowed us to remove population frequency biases for variants in the reference panel. We used PLINK (version v2.00a2LM) to estimate Identity by Descent (IBD probabilities) in 789 GCAT samples. We identified one full-sibling pair and one first-cousin relationship. These pairs of individuals showed probabilities of sharing 0, 1 and 2 IBD alleles equal to (0.3,0.48,0.22) and (0.78,0.22,0), which are close to the theoretical values (0.25,0.5,0.25) and (0.75,0.25,0) for full-siblings and first-cousins, respectively¹⁸⁶. For each of the related pairs, we discarded the sample with the **highest proportion of missing genotypes** (Figure 21D).

3.5.3.5. Population structure without reference ancestries

We applied an additional PCA without reference ancestries to obtain an homogenous sample within the GCAT cohort, avoiding the stratification between volunteers in the same population. We discarded from this analysis **two additional GCAT samples** (Figure 21E). After all the QC steps applied, we used **785 samples to construct the panel of genetic variability of the Iberian population**. Supplementary Table 2 shows the 23 GCAT samples discarded.

3.5.4. The impact coverage on SV calling

Whole-Genome Sequencing (WGS) of the 808 GCAT cohort samples was performed at high coverage (30X). We evaluated the coverage effect on SV discovery as follows. Using Picard (version 2.18.11), we downsampled ten randomly selected samples at different coverages: 5X, 10X, 15X, 20X, and 25X. Then, we performed the variant calling for each sample and coverage (section 3.2.2). Finally, we filtered the calls, selecting from each VCF the variants that passed (PASS flag in VCF file) all the software filters (Figure 22).

3.6 Variant calling in the GCAT samples

In section 3.2, we described how we ran all variant callers for a single sample. Many algorithms are known to improve the detection and genotyping of variants when running multiple samples simultaneously. This section reviews how to execute the algorithms in a multi-sample mode.

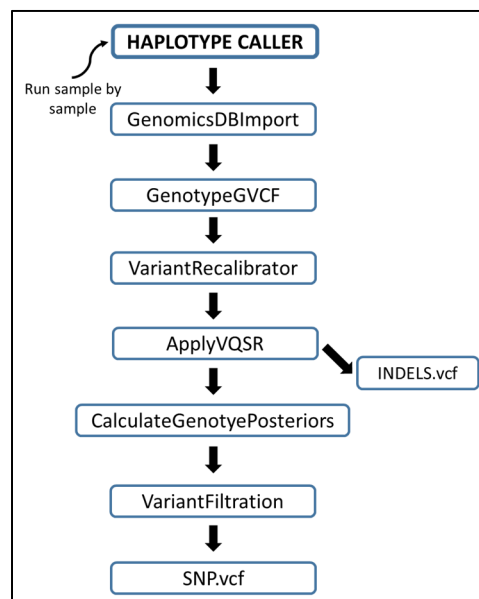
3.6.1. SNV and Indel calling

SNVs and indels were called with Haplotype caller, Deepvariant, and Strelka2. We normalised the indel and SNV VCF files, to obtain the same variant alternative representation among these tools, which is needed to apply an efficient merge between variant callers (section 3.7).

3.6.1.1. Haplotype Caller

Haplotype Caller was run by sample and chromosome as mentioned in section 3.2.1.1, including the flags `--ERC GVCF`, `--dbsnp dbsnp_138.b37.vcf`, `--L chrM` `--G Standard Annotation`, and `--G StandardHCAnotation`. We used the ♂ or ♀ Reference Genome (section 3.5.2.1) based on the sample gender.

We combined all the 785 samples using the `GenomicsDBImport` module and default parameters. Next, we parallelised the execution by chromosome and batches of 1 MB. Also, we included the label `--batch-size 50` to decrease the computational time of each execution. Then, we re-genotyped all sample variants with the `GenotypeGVCF` module, including the flags `-v gendb://merge`, `--G StandardAnnotation`, `--new-qual`. Finally, we combined together all batches and chromosomes again, in-house developed script. To detect chromosome Y variants, we followed the same steps using the 388 male samples only.



Scheme 3. Steps to run Haplotype caller in multi-sample mode.

We applied the Variant Quality Score Recalibration (VQSR) as recommended by GATK developers, to reduce false-positive calls. We ran the `VariantRecalibrator` (using the label `-AS`) and `ApplyVQSR` modules in default mode. These modules were executed separately for SNPs and Indels, using the `-module` flag.

These steps were repeated for chromosome Y independently. We also applied the VQSR for SNVs in the mitochondrial (MT) chromosome.

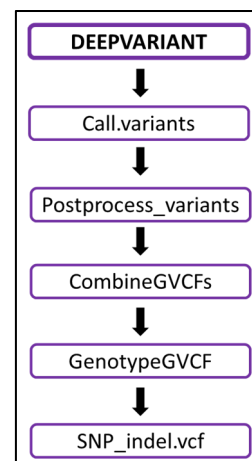
To improve the quality SNV genotypes, we executed a genotype refinement using the `CalculateGenotypePosteriors` and `VariantFiltration` modules with default mode. Variants with a genome quality below 20 were filtered out. In Scheme 3, we illustrate all the steps to execute the Haplotype Caller properly.

3.6.1.2. Strelka2

Strelka2 was run sample by sample, as described in section 3.2.1.3, as this tool is not able to run in a multi-sample mode if the chromosome number is different between samples. Therefore, we could not run all the samples at the same time because male and female samples differ in the presence of the Y chromosome in the BAM file.

3.6.1.3. Deepvariant

Deepvariant was run as described in section 3.2.1.2. After applying the three Deepvariant modules, we genotyped all variants together using two GATK modules, as recommended by the developers. First, we combined the 785 samples by chromosome and batches of 1 MB, with the CombineGVCFs module. Next, we genotyped all the detected variants using the GenotypeGVCF module and default parameters. The genotyping step was performed for chromosome Y separately. In Scheme 4, we illustrate the steps followed to run Deepvariant in the 785 GCAT samples. Finally, we merged all chromosomes and batches in a single VCF file.



Scheme 4. Steps to run Deepvariant in multi-sample mode.

3.6.1.4. SNV/Indel VCF Normalisation

Callers use different rules to report calls, especially in the case of indels and Multiple Nucleotide Variants (MNVs). Standardising the outputs of the variant callers is therefore required to merge of Deepvariant, Haplotype Caller, and Strelka2 correctly.

We used the GATK module SelectVariants to split each multi-sample VCF obtained from Haplotype Caller and Deepvariant by sample. Default parameters were used to execute the SelectVariants module, including the following labels: `--exclude-filtered`, `--remove-unused-alternates`. Then, using an in-house script, we first separated the MNPs from biallelic variants, and then we split biallelic variants into SNVs and indels. Next, the VCF normalisation was accomplished with the `pre.py` tool, designed by the Global Alliance for Genomics and Health (GA4GH). We ran `pre.py` as recommended by developers, including the flags `--L`, `--decompose`, `--pass-only`, and `--threads`. Finally, we created a final VCF per sample and chromosome, discarding MNVs and variants for which the ALT allele was not specified.

3.6.2. Mid-size and Structural Variant calling

Structural Variant (SV) detection was carried out with Delly2, Manta, Pindel, Lumpy, SvABA, Whamg, CNVnator, Popins, and MELT. These tools allowed us to cover a variety of SVs types, except for mCNVs. The obtained outputs were used in section 3.7.

3.6.2.1. Delly2

Delly2 CALL (Scheme 5) was run as described in section 3.2.2.1, using the ♂ or ♀ reference genome based on sample gender. We could not apply the entire Delly2 pipeline for samples JID673, JID748, and JID727 to find TRA events due to time and computational resources issues. For this reason, we applied the first step for TRA detection. For the remaining SV types, we ran the MERGE and CALL bcf modules of Delly to combine and re-genotype all variants, using the default parameters as recommended by developers. Next, we used Bcftools MERGE to

combine all the bcfs obtained from the re-genotyping step. We executed it following the recommendations of Delly developers. Finally, we used the FILTER module of Delly to delete redundant variants and find confident germline SVs.

We repeated the same pipeline for each SV type. Also, we ran the Delly2 pipeline to detect chromosome Y variants independently, using the male samples only.

The Delly FILTER module provides a multi-sample vcf. In order to apply our custom merge, we split the VCF by sample. We used the GATK module SelectVariants, as described in section 3.6.1.4, obtaining 5 VCF files per sample, one for each of the SV type analysed.

3.6.2.2. Manta

Manta was run as described in section 3.2.2.2. For BAM construction issues (section 3.6.1.2), we could not execute all samples in multi-sample mode.

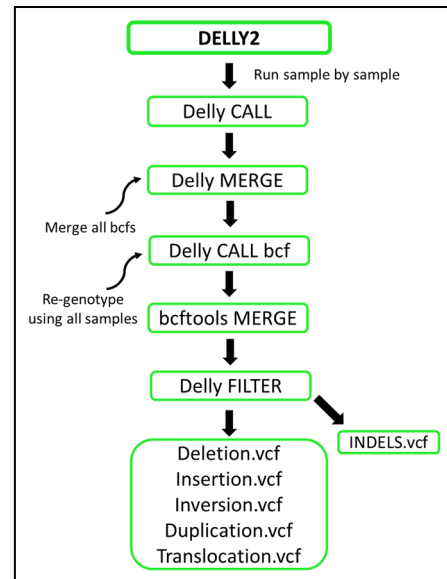
3.6.2.3. Pindel

Pindel was run by chromosome and sample, as described in section 3.2.2.3. In the config file, we included the insert size information, obtained from the Picard tool section 3.5.3.1. We could not execute the pipeline in chromosome 2 for samples JID272, JID278, JID286, JID309, JID541, JID673, JID727, and JID748 due to computational requirements. Then, we converted the Pindel format files to VCF with the pindel2vcf module, in exception for translocations, due to format incompatibilities. Finally, to genotype large insertions, we applied our custom genotyper, described in section 3.2.2.3.

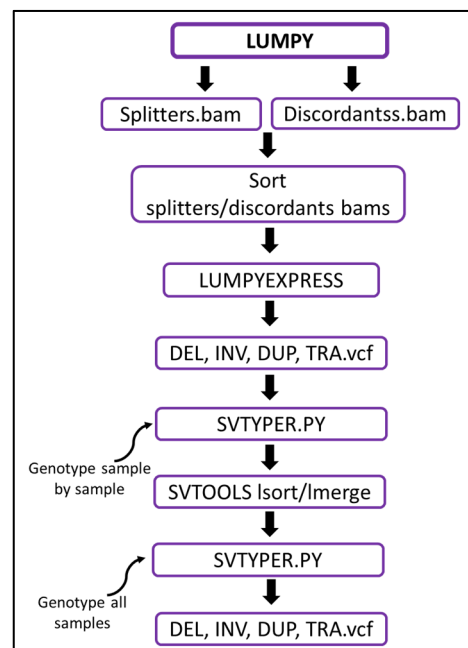
3.6.2.4. Lumpy

Lumpy and SVTYPER were run as described in section 3.2.2.4. We re-genotyped all the samples together to improve the genotype quality and recall of Lumpy. Firstly, we used the lsort and lmerge modules of SVTOOLS to combine the VCFs and merge the redundant variants. In module lmerge, we used a breakpoint-error of 50 bp ($-f\ 50$) to consider a variant as the same between different samples. Next, to improve the re-genotyping performance of SVTYPER, we divided the VCF obtained from lmerge into batches of 15K variants.

We repeated the same pipeline with chromosome Y, using the male samples.



Scheme 5. Steps to run Delly2 in Germline and multi-sample mode.



Scheme 6. Steps to run Lumpy in multi-sample mode.

3.6.2.5. SvABA

SvABA was run sample by sample, as described in section 3.2.2.6. The SV type classification of SvABA was not used in SV detection, so we used the variant detection to reinforce the variant prediction as true-positive or false-positive with the Logistic Regression Model.

3.6.2.6. Whamg

Whamg was run sample by sample, as described in section 3.2.2.5. This tool is able only to call SVs and not SV genotypes. The reference genomes used to detect the variants differed depending on the sample gender. Then, we used the SVTYPER tool described in section 3.2.2.4 to genotype all variants detected by Whamg.

3.6.2.7. CNVnator

CNVnator was run sample by sample, as described in section 3.2.2.7. We used the read-length (150) in the bin size parameter.

3.6.2.8. Popins

Popins was run as described in section 3.6.2.8, with some additional steps, allowing the execution in a multi-sample mode. All steps were executed as recommended by developers. We executed the following steps for each sample: 1) assembly, 3) contigmap, and 7) genotype. Then, steps 2) merge, 4) place-refalign, 5) place-splitalign, and 6) place-finish, were executed with all samples together. To genotype the samples, we used the `-m RANDOM` label.

3.6.2.9. Melt

Melt was run in multi-sample mode using the MELT-SPLIT pipeline. This pipeline is composed of (i) Pre-processing BAM files, (ii) MEI discovery by sample; (iii) Merge all samples to determine the breakpoint accurately; (iv) Genotype the variants together; (v) MakeVCF file.

3.7 Variant Calling integration

The integration of multiple variant callers allowed us to improve the performance of variant calling and increase precision and recall. This section describes the pipeline followed to integrate the variant caller outputs in a consensus VCF file and filter it.

3.7.1. VCF pre-processing

Before merging the outputs from variant callers for each sample, we applied some filters to clean up the detected variants. We normalised the VCFs and discarded the MNPs and those SNPs and indels which the alternative allele was not specified. Next, for the VCFs of SV set, we filtered out the low quality/NO_PASS variants, the 0/0 and ./ variants and those not mapped in autosomal and sexual chromosomes. In addition, we removed the variants catalogued as NOANCHOR in the VCFs from Popins and the BNDs variants that had both the breakpoints in the same chromosome in the VCFs from Lumpy. Finally, the variants were grouped by SV type in independent files.

In the files obtained for SNPs and indels after the pre-processing step the chromosome, we included position, reference allele, alternative allele, and genotype information. The information of SVs files (with the exception of INSS and TRAs) included chromosome, position initial, position final, SV length, and genotype. The TRA variants included the second chromosome, while the length was not provided. The INS variants included chromosome, position, and genotype.

3.7.2. Merging all VCFs per sample and per variant type

After the pre-processing step, we merged the VCF outputs from the variant calling for each sample as follows. For SNVs and Indels, we merged variants by (i) chromosome, (ii) position at base-pair resolution, and (iii) same REF/ALT alleles. For SVs, we merged variants by (i) variant type, (ii) chromosome, (iii) overlapping breakpoint-error of the variant (section 3.2.3.3), and (iv) reciprocal overlap $\geq 80\%$ between callers (this filter was not applied for INS and TRA SVs). Using these merged databases, we applied the LRM (section 3.4.2) by variant type and sample. Based on the LRM results, we considered as a TP if the LRM prediction was ≥ 0.5 . Otherwise, we considered the variant as a FP. For SNVs, we considered a variant as TP if at least two callers had detected it. On the other hand, for SVs, we reported the maximum breakpoint-error, the number of callers and number of strategies that had detected the variant, the breakpoint position, genotype, and length of the variant, following the strategies described in section 3.4.1.2 and sections 0, 3.4.2.3 for each SV type.

3.7.3. Combining all samples in a single VCF

For each variant type, we combined the 785 GCAT samples as follows. For SNPs and indels, we merged individuals by (i) chromosome, (ii) position at base-pair resolution, and (iii) REF/ALT alleles. For SVs, we merged individuals by (i) variant type, (ii) chromosome, (iii) maximum breakpoint-error of the merged variant (Table 16), and (iv) reciprocal overlap $\geq 80\%$ (with the exception of INS and TRA) between individuals. Using these merged databases, we calculated the TP proportion for each variant as determined by LRM in the 785 GCAT samples. **This proportion is referred to as the quality score of the merged variant.** Then, we considered a variant as PASS if the quality score was ≥ 0.5 . Besides, we reported the length and position of each SV as the median length and median position of all the samples that had the SV. Additionally, we used the sample alleles to provide populational information for each variant, such as the allelic count (AC), the Minor Allelic count (MAC), the Allele Frequency (AF), the Minor Allele Frequency (MAF), or the population variation (POPVAR) which indicates if the variant was common, low-frequency, rare, a doubleton or a singleton in the cohort. Then, we also reported variant-specific features, such as breakpoint-error (ERRBKP), SVTYPE, among others. All information was organised in a VCF file.

3.7.4. Variant Quality Control

The final set of variants was obtained as follows. We considered PASS, the variants of quality score ≥ 0.5 and discarded the monomorphic ones. Next, we used PLINK to remove variants in Hardy-Weinberg Disequilibrium (Bonferroni correction p-value $< 5 \times 10^{-8}$) and those with $\geq 10\%$ of missings.

3.8 New discoveries and validation of the GCAT variants

The genetic characterisation of a specific population genetic allowed us to discover new variant rearrangements and their effect on diseases. However, variant calling is known to produce false discoveries, and if variant filtering is not well applied, it could introduce noise in the dataset, driving misinterpretations, and results with questionable quality. In this section, we describe how we evaluated the new discoveries obtained from the GCAT cohort and the quality of the dataset.

3.8.1. Comparative studies with different datasets

3.8.1.1. SNVs and indels

We compared the GCAT final set of SNVs and indels (section 3.7.4) with the NCBI dbSNP Build (version 153) dataset downloaded from <https://ftp.ncbi.nlm.nih.gov/>. We considered dbSNP as the gold standard, and we merged it with the GCAT set in two different ways. First, merging by (i) chromosome, (ii) position at base-pair resolution, and (iii) REF/ALT alleles. Second, merging by (i) chromosome and (ii) position. Finally, we determined the number of variants shared in both databases and the number of unique variants in the GCAT set.

3.8.1.2. Structural Variants

We compared the final set of SVs with (i) The Genome Aggregation Database (gnomAD.v.2) (<https://gnomad.broadinstitute.org/downloads>), (ii) the Database of Genomic Variants (DGV) (<http://dgv.tcag.ca/dgv/app/downloads?ref=GRCh37/hg19>), (iii) the Human Genome Structural Variation Consortium set (HGSVC) (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/) (iv) the Ira M Hall dataset (https://github.com/hall-lab/sv_paper_042020), (v) the 1000G project (Phase3) (<ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/>) and (vi) GoNL (release 6.2) as reference datasets. We matched each of the reference datasets and the GCAT variant set by (i) variant type, (ii) chromosome, (iii) 1000 bp breakpoint-error and (iv) reciprocal overlap $\geq 80\%$ between datasets. We determined the number of variants shared in at least one dataset, and the number of variants unique to the GCAT variant set (Figure 25).

The same procedure was further applied considering the 1000G (Phase3) and GoNL datasets to estimate the number of new imputable SVs (Figure 25). For translocations, *de novo* insertions and transposons, the reciprocal overlap filter was not applied, due to the SV length being unavailable.

3.8.2. Experimental validations

3.8.2.1. Validation of SNVs and indels using the GCAT SNP-array

We compared the SNV/indel final set (section 3.7.4) with the GCAT SNP array (Table 8). First, we selected 570 GCAT samples that had both WGS and SNP-array data. Then, we merged both datasets by (i) chromosome, (ii) position at base-pair resolution, and (iii) REF/ALT alleles. Finally, we calculated recall and genotype concordance for each sample based on 732,978 SNPs and 1,168 indels in the SNP array, considering 1-23 chromosomes (Figure 34).

3.8.2.2. Copy Number Variation (CNV) validation using Comparative Genomic Hybridization (CGH)

We hybridised five GCAT samples in a CGH array, using as a reference the GIAB sample NA12878. We used the silver standard CNV set¹⁸⁷ (downloaded from <https://bmcbgenomics.biomedcentral.com/articles/10.1186/s12864-017-3658-x#Sec21>) of NA12878 to validate large deletions and duplications (>20 Kb), corresponding to a minimum of 5 consecutive CGH probes for variant detection. The CGH technique does not validate small sizes due to their resolution and the distance between probes (3 Kb). The probe intensity changes between GCAT and reference samples indicated duplications (intensity gain) or deletions (intensity loss). When no intensity change was appreciated, both samples included the same variant or was a true negative. We validated large deletions and duplications (>20 Kb) detected by variant callers in two different ways: (i) variants present in silver CNV list and GCAT samples, overlapping in a window of 1 Kb, with no differences in probe intensities and; (ii) variants in GCAT samples with probe intensity changes.

3.8.2.3. Inversions validation using a verified dataset

To check the GCAT dataset's accuracy, we used a set of inversions experimentally validated from the InvFEST project (Lerga-Jaso et al., in preparation). Each validated variant included the GRCh37 coordinates, the allelic frequency among different continents, and the length. We considered 64 Non-Homologous (NH) inversions and inverted duplications because variants generated by Non-Allelic Homologous Recombination (NAHR) procedures are located in repetitive regions, making them harder to detect by short reads, as performed in the GCAT samples. Then, we matched this InvFEST project subset with the GCAT final inversion set (section 3.7.4) considering an overlapping a window of ± 1000 bp. Finally, we calculated recall, allele frequency (using European frequencies) and length concordances of inversions from the GCAT and InvFEST datasets (Figure 35).

In parallel, as described in (Lerga-Jaso et al., in preparation), the accuracy of GCAT inversion calling and genotyping was evaluated using 51 NH InvFEST inversions. We selected the SNVs discovered from variant calling in the GCAT samples (section 3.7.4) in a window of ± 100 Kb from 51 NH inversion breakpoints. Then, following the protocol of Lerga-Jaso et al., we imputed inversions using the reference panel of experimentally-validated inversions provided by the InvFEST Project (Lerga-Jaso et al., in preparation), and the SNPs previously recovered. Inversion genotypes were called with a posterior probability higher than 0.8 and were classified as missing otherwise. Besides, if the inversion had perfect tag SNPs ($r^2 = 1$) present in the GCAT calling, those missing genotypes were recovered based on the tag SNP genotypes. Then, genotype concordance and variant calling accuracy were estimated by comparing the genotypes obtained from GCAT calling and InvFEST imputation for all the 785 GCAT samples (Figure 35).

3.9 Creation and integration of haplotypes sets

We processed the consensus vcf files (section 3.7) to create a haplotype-resolved panel to perform imputation. The sequencing of the GCAT samples at high coverage (30X) decreased the level of uncertainty of genotyping^{86,118}, allowing us to avoid the use of genotype likelihoods (GL) or Phred-scaled likelihoods (PL) in the phasing step. This section describes the benchmarking of different phasing strategies and the pipeline to construct the GCAT haplotype panel.

3.9.1. Benchmarking of different phasing strategies

Since the initial SV-resolved panels^{5,110}, phasing technology has evolved, producing new program updates and strategies, making it more efficient, and improving haplotype estimation. Unfortunately, little is known about the efficacy to phase SVs by these tools. This section analyses the benchmarking of different phasing strategies to select the most accurate to generate the Iberian-GCAT panel. To benchmark different phasing strategies, we evaluated the imputation accuracy of SVs in a subsample of 95 individuals characterised by having both WGS and SNP-array data available. We considered the quality of imputed variants as a metric to determine the best phasing strategy.

3.9.1.1. Sample pre-processing

We used the GCAT SNP-array data (Table 8) to perform imputation after applying a Quality Control (QC)¹⁷⁴ to filter the data. The `gcat_core` array includes 756,773 SNVs for 4,988 samples before applying the AT-CG and Minor Allele Frequency (MAF) > 0.1% filters. To benchmark the phasing strategy, we used 7,146 SNVs on chromosome 22 for a subset of 95 samples from GCAT (Supplementary Table 3), having both SNP-array data and WGS data, in order to evaluate the amount and the genotype concordance between variants imputed from SNP-arrays and called from WGS.

To create a pilot reference panel, we used the variants from chromosome 22, obtained in the variant calling integration step (see section 3.7). This combined subset included 195,106 SNVs, 24,321 indels, and 128 large Deletions (>150 bp). In building the panel, we used 690 individuals, after discarding the 95 samples previously selected for the benchmarking. Finally, with PLINK2, we re-tested for Hardy-Weinberg Equilibrium (HWE) (`--hwe 1.639629e-07` midp (Bonferroni Correction)), we filtered out all variants with $\geq 10\%$ of missings (`--geno 0.1`), and removed all singletons and doubletons (`--maf 0.0026`) from the `vcf`.

3.9.1.2. Phasing strategies

There are different possible strategies to create a panel of genetic variability. We analysed the accuracy to generate the phased set using Shapelt2¹⁸⁸ (version v2.r904), Shapelt2 plus MVNcall¹⁶⁹ (version 1.0), Shapelt4¹⁶⁶ (version 4.1.3), Shapelt4 plus MVNcall, and Shapeit4 plus WhatsHap¹⁶⁷ (version 0.18).

3.9.1.2.1 Shapelt2 (with SVs)

Shapelt2 is a tool mainly used to phase SNVs and indels, and it has been used in projects similar to GCAT, such as 1000G and GoNL, to create haplotype sets. For this reason, we evaluated the capacity of Shapeit2 to phase all variant types, including SVs. We executed Shapelt2 as follows:

```
Shapeit --input-vcf snps_indels_del_to_phase_HWE_no_double_single
QC01.vcf --input-map genetic_map_chr22_combined_b37.txt --output-max
snps_indels_del_to_phase_HWE_no_double_singleQC01.haps snps_indels_
del_to_phase_HWE_no_double_singleQC01.legend --thread 48
```

Finally, using an R script developed in-house, we adapted the `.haps` and created the `.legend` files needed to run IMPUTE2 for the subsequent imputation.

3.9.1.2.2 Shapelt2 and MVNcall

This strategy has been considered to create an SV-resolved haplotype set in the 1000G and GoNL projects^{5,110}. We reproduced their pipeline to estimate its accuracy to phase SVs. MVNcall is used to phase MNPs (Multiple Nucleotide Polymorphism), complex indels and SVs, and it requires a pre-built haplotype scaffold. We constructed this haplotype scaffold by selecting SNPs and indels from chromosome 22 filtered previously described and by analysing it with Shapelt2 as described in section 3.9.1.2.1.

MVNcall requires a Genotype Likelihood (GL) or a Phred-scaled Genotype Likelihoods (PL) for each variant. The LRM did not report GL and PL, so taking advantage of high coverage, we introduced for each variant the PL as 0 in the genotype reported by variant calling integration step (section 3.7), and 225 for the remaining genotype probabilities. Variants with “.” genotypes were discarded. We executed MNVcall as follows:

```
mvncall --int 1 51304566 --sample-file SNP_indels_GCAT1.sample
--glfs Del_final_to_phase_nodots_no_singleton_doubleton_hwe.vcf --sca
ffold-file SNPs_indels_final_to_phase_HWE_no_double_singleQC01.haps
--lambda= 0.1 --o Del_final_to_phase_mvncall_1_51304566_all.vcf
```

Finally, we combined the Shapelt2 and MVNcall outputs and generated the final .hap, .legend, and .sample with an in-house script.

3.9.1.2.3 Shapelt2, MVNcall with PIRs

Shapelt2 can use sequencing reads to improve the phasing of rare and singleton variants for samples with high coverage (30X). Reads that spanned two heterozygous sites were labelled as Phase Informative Reads (PIRs) and considered as mini-haplotypes¹⁶³.

We downloaded the PIR module (extractPIRs.v1.r68.x86_64.tgz) from the official Shapelt2 web page. We executed this module only for SNPs as recommended by the developers. Then, we ran Shapelt2 as described in section 3.9.1.2.1, including the `-input-pir` flag. Finally, we executed MVNcall, as described in section 3.9.1.2.2.

3.9.1.2.4 Shapelt4

The new Shapelt4 (version 4.1.3) uses the Positional Burrows-Wheeler Transform (pBWT) algorithm. The method stores the haplotypes at each iteration in a pBWT data structure, facilitating locally matching haplotypes to be identified in a given window. This Shapelt version is highly accurate and computationally efficient compared to other phasing algorithms. Before running Shapelt4, we compressed with `bgzip` and indexed with `tabix` the pre-processing file (section 3.9.1.1) as developers recommended. Then, we ran Shapelt4 as follows:

```
shapeit4 --input snps_indels_del_to_phase_HWE_no_double_singleQC01.vcf.gz
--map genetic_map_chr22_combined_b37.txt --output SNP_indels_SV_GCAT_all
.vcf --pbwt-depth 8 --seed 123456 --region 22 --thread 48 --sequencing
--log SNP_indels_SV_GCAT_all_ok.log
```

Finally, we executed Bcftools (`convert --haplegendsample`) to convert the vcf into `.hap`, `.legend`, and `.sample` files, required to run IMPUTE2.

3.9.1.2.5 Shapelt4 and MVNcall

To improve SV phasing, we applied the same strategy developed in section 3.9.1.2.2 but changing the Shapelt version.

3.9.1.2.6 Shapelt4 and WhatsHap

Shapelt4 uses the WhatsHap tool to extract PIRs, by grouping heterozygous genotypes into phased sets when overlapped by the same sequencing reads. However, this tool is able to recover this type of reads from SNPs and indels only.

With the current WhatsHap version, it's not possible to parallelise multiple executions, so to increase the performance, we divided the multi-sample VCF file (pre-processed) into individual VCF files, one for each sample. We ran WhatsHap as follows:

```
whatshap phase -o sample_snps_indels_SV_filtered_final_to_phase_QC01_
shapeti4_whatshap.vcf.gz --tag PS --reference genome.fa --indels
--chromosome 22 snps_indels_SV_filtered_final_to_phase_QC01_shapeit4_
sample_ok.vcf.gz
```

Once we got all outputs, we combined all the individual VCFs in a new multi-vcf file with an in-house script, including in the column FORMAT the “GT:PS” string, only for SNVs and indels, because is the information obtained from WhatsHap. Finally, we executed Shapeit4, including the `--use-PS 0.0001` flag, and Bcftools as described in section 3.9.1.2.4.

3.9.1.3. Imputation using different phasing strategies

We imputed SNP array data for the 95 GCAT individuals using the different reference panels obtained with the strategies described in the previous sections. Before imputation, we applied a “pre-phasing” step to the array data to reduce the computational costs without compromising accuracy¹⁸⁹. We used the `--input-ref` and `-H` flags for Shapeit2 and Shapeit4, respectively, to pre-phase the array with the reference panels created in section 3.9.1.2.

IMPUTE2¹⁷¹ (version 2.3.2) was used to impute the phased array data. We applied the same command for all the different reference panels, dividing by batches of 5 Mb. We executed IMPUTE2 as follows:

```
impute2 -use_prephased_g -m genetic_map_chr22_combined_b37.txt
-h reference_panel.hap.gz -l all_snps_reference_panel.legend.gz
-known_haps_g gcat_test_imputation.hap.gz -int 5 Mb batch -o
output_file.gen -o_gz
```

Finally, we merged all the batches and converted the `.gen` files to VCF using PLINK.

3.9.1.4. Selecting the strategy to phase the GCAT samples

The phasing strategy is one of the key factors determining imputation quality. We assessed the best phasing strategy by counting the number of variants of high quality imputed variants (info score ≥ 0.7) and calculating the genotype concordance between imputed and WGS data. In our hands, the best results were obtained using Shapelt4, with slight differences between the Shapelt4 and Shapelt4 + WhatsHap strategies, in which the latter one recovered 85 SVs compared to 81 with Shapelt4 alone (Figure 36A). The differences came from rare variant imputation, where Shapelt4+WhatsHap showed an improved imputation quality (Supplementary Figure 8A). For this reason, we used **Shapelt4+WhatsHap** to phase the GCAT samples.

3.9.2. Pipeline to construct the Iberian-GCAT haplotype panel

We followed the strategy described in section 3.9.1.2.6 to create a haplotype-resolved panel of GCAT using the 785 samples. We generated a haplotype panel for each autosomal chromosome and the X chromosome. The pipeline is illustrated in Figure 14:

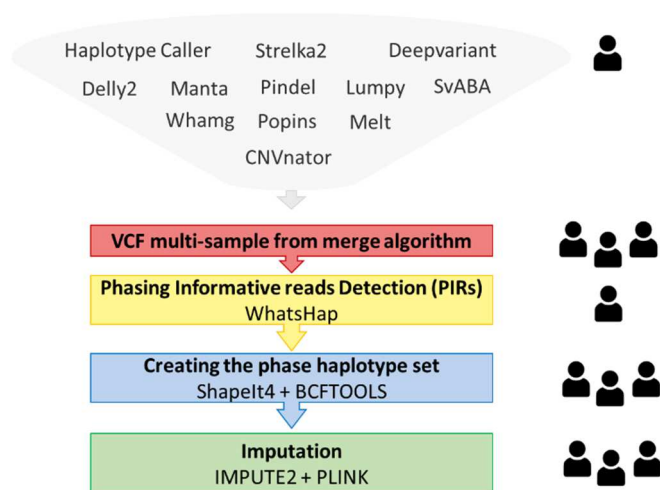




Figure 14. Pipeline followed to construct the Iberian-GCAT phase haplotype set.

-  Step executed sample by sample.
-  Step executed with all the 785 GCAT samples together.

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, we coded the heterozygous genotypes in NOPAR as “./.” for male samples. On the other hand, to improve the imputation of chromosome X, we included two specific flags, `-chrX` and `-sample_g`, and we grouped by gender the samples from the array. Finally, we executed the pipeline described in Figure 14.

3.10 Imputation using the Iberian-GCAT reference panel

Once the Iberian-GCAT reference panel was created, we validated its performance in imputation with different SNP-genotyping arrays. We considered the same 95 samples of the GCAT array (section 3.9.1.1), as well as, the array data from 1000G, as an alternative, non-

population-specific dataset. We evaluated different aspects of imputation, such as 1) the genotype concordance between imputation and calling, 2) the imputation differences using a reference panel the PIRs information, 3) the relevance of including the A/T-C/G variants in the array for imputation resolution, and 4) the SV effect in the imputation quality.

3.10.1. Imputation analyses using the GCAT SNP-genotyping array

Similar to section 3.9, we created a whole-genome pilot reference panel of 690 GCAT individuals using the VCF files obtained in the merge step (section 3.7.4). The variants included in this panel were obtained after filtering out variants with $\geq 10\%$ of missing and removing monomorphic variants with PLINK. Using this pilot reference panel, we evaluated the imputation accuracy on 95 independent GCAT individuals using 754,593 whole-genome SNVs from the SNP array (Table 8) that passed a strict Quality Control¹⁷⁴. Before performing imputation for the GCAT array of 95 samples (section 3.9.1.1), we converted the GCAT SNP-array data to VCF format using PLINK, fixing the ALT and REF alleles as A1 and A2, respectively, with the `--a2-allele` flag. We filtered out all indels and further 36 SNPs with a REF allele discordant from the reference genome.

The following analysis was done to improve imputation performance. 1) All variants in the array were in the forward strand¹⁷⁴, thus overcoming the typical strand issues arising from real array data for variants with A/T or C/G alleles. For this reason, we evaluated the impact of those type of variants in imputation, compared to the exclusion that is commonly applied. 2) To evaluate the effect of PIRs in the panel, we constructed two reference panels, as described in sections 3.9.1.2.4 and 3.9.1.2.6, respectively. Imputation was executed as described in section 3.9.1.3 (Supplementary Figure 8).

To evaluate the genotype concordance, we selected the variants with an info score ≥ 0.7 from the imputation results. Then, we compared the imputed genotypes with those from WGS calling. The genotype concordance was evaluated by sample. When both genotypes were identical, we considered the variant as well imputed (Figure 37). Additionally, using PLINK, we evaluated the number of variants with linkage disequilibrium (LD) $r^2 > 0.2$ in a window of 1 MB for each common SV type (MAF $> 5\%$) to find a correlation between genotype concordance and low LD (Figure 38A). Finally, we assessed the effect of SVs for SNP and indel imputation by 1) exploring the info score in a window of 100 bp from the reported SV breakpoint (Supplementary Figure 9), and 2) the number of SNPs and Indels recovered without SVs (Figure 38B).

3.10.2. Imputation quality using the array data of 1000G

To evaluate the efficiency of imputing variants in different populations, we download the 1000G 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. The SNP array data was available for 2,318 samples (48.1% males) from 19 populations and 5 continental groups with 2,458,861 SNPs¹¹¹.

3.10.2.1. Sample filtering and Quality control of the 1000G array

The array data were filtered for samples or markers of low quality. To evaluate the sample gender, we downloaded the `igsr_samples.tsv` file from <https://www.internationalgenome.org/data-portal/sample>. From this, we filtered out 41 samples with unknown gender. We further

discarded 395 related samples ($\geq 2^{\text{nd}}$ degree relatedness), using the 2013606_g1k.ped file downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130606_sample_info/. Finally, we discarded two Masai samples due to low sample size, obtaining an array of 1,880 samples. The array samples were divided into the 19 populations included in the 1000G study.

We then used PLINK to perform a QC in the 19 population subsets individually, to filter out low quality and non-informative variants, and re-analyse sample relatedness. Below we list the filters applied:

- Variant filtering:
 - Discard variants in LD `--indep-pairwise 50 5 0.2`
 - Discard variants with $\text{MAF} \leq 0.4$ `--maf 0.4`
 - HWE equilibrium with Bonferroni correction `--hwe Bonferroni correction`
 - Missingness call rate $\geq 10\%$
 - Discard all A/T and C/G variants

In contrast to the GCAT array, we discarded the A/T and C/G variants because the strand direction was unknown in the 1000G array.

- Sample filtering:
 - Discard related samples `--rel-cutoff 0.05`
 - Excess of heterozygosity ± 2 sd `--het`
 - Missingness call rate ≥ 0.1

We split males from females using the gender information from igsr_samples.tsv file for chromosome X. Finally, chromosome X was divided into PAR and NOPAR regions.

3.10.2.2. Imputation of non-Iberian samples with the Iberian-GCAT panel

Using the Iberian-GCAT reference panel (section 3.9.2), we imputed the 19 populations from 1000G separately (section 3.9.1.3), and we evaluated imputation accuracy using as a reference the SV characterisation performed by Audano et al.⁶⁷. We download the EEE_SV-Pop_1.ALL.sites.20181204.vcf file from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/working/20181025_EEE_SV-Pop_1/VariantCalls_EEE_SV_Pop_1/. This file contained 15 samples sequenced with PacBio, aligned against the GRCh38 genome and SV-characterised with the SMRT-SV tool. We applied a liftover to convert the VCF GRCh38 genomic coordinates to GRCh37 with the liftOverPlink.py tool. Nine samples included in 1000G SNP-array (HG00514, HG00733, NA19240, NA19434, HG01352, HG02059, NA12878, HG02106 and HG00268) were used to evaluate the precision and recall of imputation. For SVs, we compared imputed variants with an info score ≥ 0.7 with the reference SVs⁶⁷, considering a window of ± 50 bp to determine overlap between the two datasets. The Audano dataset was used in this case in terms of variant presence/absence only. We evaluated the concordance of SVTYPE and variant length between Audano characterisation and GCAT dataset. Additionally, we discarded the variants with genotypes 0/0 and “./.”. (Figure 39).

To calculate SV genotype concordance, we used instead data reported by Hickey et al.¹²¹ as reference. This project genotyped three samples from Audano et al.⁶⁷ (HG00514, HG00733, and NA19240), using short reads and applying a variation graph implemented in the vg toolkit. We download the svpop-vg-HG00514.vcf.gz, svpop-vg-HG00733.vcf.gz, svpop-vg-NA19240.vcf.gz files from <https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=vgs2019/vcfs/>. We converted the VCF GRCH38 genomic coordinates to GRCh37 coordinates with the liftOverPLink.py tool. All variants with “.” in the reference were discarded. Also, the genotypes of imputed TRAs, TRPs, INs, and DUPs were compared against the INs reported by Hickey et al. (Figure 40).

3.11 Benchmarking the reference panels

To demonstrate that the GCAT panel was an invaluable resource for population studies or Genome-Wide Association Studies (GWASs), we imputed the GCAT SNP- genotyping array data with different reference panels. The GCAT array data contain 756,773 SNPs for 4,988 samples (detailed description in section 3.9.1.1). We checked the percentage of missings genotypes by sample and chromosome using PLINK, and we discarded three samples (JIDraw1, JIDraw3, JIDraw3), showing $\geq 10\%$ of missings. Finally, we filtered out the 537 samples present in the array and the GCAT panel, obtaining an SNP-genotyping array of 4,448 samples.

To reduce the time to perform the imputation with different reference panels, we used GUIDANCE¹⁷². This tool can pre-phase haplotypes and impute genotypes using multiple reference panels in a single execution, making the analysis faster and lower computational resources demanding than running all panels individually. We used an update of GUIDANCE, which executes Shapelt4 to pre-phase the array and IMPUTE2 to perform the imputation. This algorithm was executed as developers recommended.

The benchmarking imputation was performed with the Iberian-GCAT and the five most popular reference panels. Table 11 shows all panels used:

| Reference Panel | Release | Variant Information |
|---------------------------------------|---|---|
| GoNL | Release 5.4 | SNVs and Indels |
| GoNL SV* | Release 1_20161013 | SNVs, Indels, and SVs (exception INS) |
| 1000G | phase 3, v5a.20130502 | SNVs, Indels, and SVs (exception INS and TRA) |
| Haplotype Reference Consortium (HRC)* | Release 1.1 | SNVs |
| UK10K | Release 2012-06-02, updated on 15 Feb 2016 | SNVs and Indels |
| GCAT | Release 1 | SNVs, Indels, and SVs |

Table 11. Panels used to benchmark the imputation analysis. The benchmarking was performed using different reference panels. GoNL SV, 1000G and GCAT were used to evaluate the imputation accuracy of SVs.

* Database not available for open access

To be consistent with all projects, we considered as SVs variants with lengths ≥ 50 bp. The remaining variants were SNVs and indels. After imputation, we retained imputed variants with an info score >0.7 and MAF > 0.001 , and evaluated which panel gave better imputation values. Results were divided by variant type. To highlight the importance of using different panels in imputation for downstream analyses, we evaluated the number of unique variants recovered by different panels as follows. For SNVs and indels, we considered a variant being the same between panels if the position and ALT allele coincided. For SVs, variants had to be of the same SV type and overlapping in a window $\pm 1,000$ bp to be considered the same across different panels (Figure 44A, B).

To evaluate the quality of imputation by variant frequency, we calculated the average the info score (r^2) for rare, low-frequency, and common variants (Figure 44C, D).

3.12 Biological impact of Structural variants

SVs are an essential source of genetic variability in the human genome, but little is known about their effect on human phenotypes. Recently, several initiatives have appeared, trying to elucidate it, and showing the importance of SVs in diseases^{10,11,65}. This section describes the SV distribution in different populations and how we annotate the SVs to characterise their potential effects on human phenotypes.

3.12.1. Structural Variant distribution in the worldwide populations

We evaluate allele frequency, distribution of variant type, and quality of imputed SVs in different human populations. We imputed each 1000 Genomes population separately with the Iberian-GCAT reference panel (section 3.10.2). We used Shapelt4 to pre-phase the array data and IMPUTE2 to impute the variants.

Imputing each population independently, allowed us to evaluate if some population was genetically too distant from Iberians, discouraging the use of the Iberian-GCAT panel for imputation studies. IMPUTE2 evaluates the imputation quality applying different cross-validations between array and imputed genotypes. To evaluate the imputation quality by population, we considered the most confident imputed genotypes (max prob ≥ 0.9), which shows the percentage of imputed genotypes that match with genotypes from the array. This information was used to calculate the median score of all max prob per each population (Figure 41).

The SV distribution among different populations was determined using the `.gen_info` file provided by IMPUTE2. This file contains information on frequency (`exp_freq_a1`) and certainty (`info`) of imputed variants. We evaluated SVs with info score ≥ 0.7 and the length ≥ 50 bp. The singletons and doubletons were not evaluated because each population dataset contained around 100 samples, so there were no variants with allele frequency below 0.01. All variant types and allele frequencies (we used the `exp_freq_a1` values) distributions were plotted with an R script developed in-house, using the following libraries (`dplyr`, `data.table`, `ggplot2`, `viridis`, `ggrepel`, `forcats`, `RColorBrewer`, `ggmap`, `maps`, `ggforce`, `scatterpie`) (Figure 42).

3.12.2. Functional impact of Structural Variants

The functional impact of SVs was performed with AnnotSV⁷⁵ (version 2.3.3). In addition, we used the GWAS catalog to evaluate how many SVs were tagged by SNVs ($LD \geq 0.8$) and their effect on diseases/traits.

3.12.2.1. Structural variant annotation using AnnotSV

We used AnnotSV to obtain functional, regulatory, and clinical information of SVs. This tool uses multiple datasets to evaluate the effect of SVs in humans. Table 12 shows the datasets used to annotate the GCAT SVs.

| Dataset | Release |
|---|--|
| Refseq | GRCh37 Feb. 2009 assembly |
| Deciphering Developmental Disorders (DDD) | DDG2P.csv.gz version 12_7_2020 |
| OMIM (Online Mendelian Inheritance in Man) | genemap2.txt and morbidmap.txt |
| American College of Medical Genetics and Genomics (ACMG) | ACMG SF v2.0 |
| Gene intolerance annotations from the ExAC | fordist_cleaned_nonpsych_z_pli_rec_null_data.txt and exac-final-cnvc.gene.scores071316 |
| Haploinsufficiency annotations (DDD) | HI_Predictions_Version3.bed.gz |
| ClinGen Consortium Rating System Database of Genomic Variants (DGV) | ClinGen_gene_curation_list_GRCh37.tsv The version of date 20190322 |
| DDD frequency annotations | 20191219_DDD_population_cnvc.sorted.bed |
| 1000 Genomes Project | ALL.wgs.mergedSV.v8.20130502.svs.genotypes.vcf.gz Genome build GRCh38: |
| gnomAD-SV | gnomad_v2_sv_sites.bed.gz |
| Ira M. Hall's lab | Supplementary_File_1.zip and Supplementary_File_2.zip |
| dbVar non-redundant SV (dbVar NR) | GRCh37.nr_deletions.tsv.gz |
| Topologically Associating Domains (TAD) | 20171024_boundariesTAD.sorted.bed |
| GeneHancer* | GeneHancer_V4_14_for_annotsv.zip |
| GC content | GRCh37 FASTA genome |
| Repeated sequences annotations | Last version |

Table 12. Detailed list of databases used to annotate the SVs detected in the GCAT project.

A detailed description of AnnotSV can be found in <https://lbgi.fr/AnnotSV/>

* No open acces

AnnotSV was executed using default parameters. To be consistent with our merge algorithm (section 3.7), we changed the reciprocal overlap parameter to 80% (`-overlap 80`). SnpEff¹⁵² was used in parallel to AnnotSV. SnpEff was executed using default parameters, limiting the annotation to DELs, DUPs, INVs. Then, to characterise the SVs, we divided the variants detected in our calling by frequency (common (MAF \geq 5%), low-frequency ($1\% \leq$ MAF $<$ 5%), rare (MAF $<$ 1%), and doubleton or singleton and SV type, in order to predict the functional impact of SVs in human phenotypes. Finally, we evaluated: 1) the estimated pathogenicity of variants, 2) the SV distribution in human genes, 3) the impact of SVs in Topologically Associating Domains (TAD) regions, 4) The Haploinsufficiency (HI) and predicted Loss of Function Intolerance (pLI) effect of genes in human and 5) The Top 10 diseases related to detected SVs.

We followed the American College of Medical Genetics and Genomics (ACMG) criteria¹⁹⁰ to predict the level of pathogenicity of SVs: 1) If an SV overlaps with morbid/candidates genes, 2) the pLI scores, and 3) if SV was already described as pathogenic. Based on these criteria, AnnotSV ranked the variants as:

- 1) Benign SV
- 2) Likely Benign SV
- 3) Variant of unknown significance
- 4) Likely pathogenic
- 5) Pathogenic

A detailed description of the pathogenic ranking can be found in <https://lbgi.fr/AnnotSV/ranking>.

We compared the number of SVs affecting genes in the gnomAD-SV, in Audano et al., and in the GCAT project. We ran AnnotSV for the VCFs obtained from gnomAD-SV¹¹ (gnomad_v2.1_sv.sites.vcf.gz file) and Audano et al.⁶⁷ (EEE_SV-Pop_1.ALL.sites.20181204.vcf file).

3.12.2.2. Evaluation of SVs using the GWAS catalog

Genome Wide Association Studies (GWASs) usually are performed to investigate the link between variants and complex diseases/traits. Nowadays, this technique is applied in numerous studies, finding relations between variants and phenotypes. Significant associations identified by GWASs are deposited in the GWAS catalog.

This invaluable repository allowed us to determine whether the variants detected in our study had been previously associated with a phenotype. We downloaded the GWAS catalog version 1.0 (e98 r2020-03-08) from <https://www.ebi.ac.uk/gwas/docs/file-downloads>, and we filter the data as follows. First, from the total 179,364 variant-phenotype associations, we selected **106,906** variants of **72,849 unique autosomal** entries identified in European ancestry. Second, we intersected 68,323 SNVs by chromosome and breakpoint with minor allele frequency $>$ 1% in our GCAT dataset (~30M SNPs). Then, using PLINK, we identified **4,733 associations (2,669**

unique SNVs) having at least one structural variant in strong linkage disequilibrium $r^2 > 0.80$ (Figure 33). From these 4,733 SNVs, we evaluated the following distributions: 1) the type of SVs tagged by SNVs and 2) the gene function impact using GWAS catalog information.

Finally, 72 SNVs (51 unique SNVs) of 4,733 associations from the GWAS catalog were linked with SVs, which in turn affected the extreme loss of function intolerant genes, according to 1) Haploinsufficiency, 2) pLI ≥ 0.9 , and SVs with 4) the pathogenic level ≥ 4 values provided by AnnotSV.

4. RESULTS

The results have been divided into three blocks. According to the major findings of the thesis, these are; 2.1) Strategy to identify and classify different software (variant calling), to improve the variant filtering and the merging of their outputs, 2.2) Characterisation the variants/genotypes of the GCAT cohort, and 2.3) Building, validation and annotation of the haplotype-resolved panel of the Iberian cohort for GWAS.

4.1 Identification and classification of Variant Callers using the Genome In A Bottle (GIAB) and *in-silico* samples

It is known that genome variability is related to human evolution and phenotype, such as anthropometric particularities or human diseases. There are different ways to discover/analyse this variability, such as Polymerase chain reaction (PCR), array analysis, mainly focused on SNPs, and Whole Genome Sequencing (WGS). This last technique requires the use of a specific piece of software known as a variant caller, to detect these genomic changes, by using different strategies such as Split-Read (SR), Discordant-reads (DR), *de novo* assembly (AS), Read Depth (RD), sophisticated Machine Learning algorithms (ML), or a combination of them.

Unfortunately, the variant callers designed to discover genome rearrangements produce False-Positives (FP) detections, especially in the case of Structural Variants (SVs). Their accuracy depends on the genomic region, variant type, length of variant, and read depth³⁴. This bias towards small variants is also seen in the newly generated HRC panel, which was constructed using only SNVs, since indels were found to be very inconsistent across projects¹⁵⁷, highlighting the difficulties of detecting SVs correctly. It has been recently documented that the combination of outputs from different variant callers allows for an improved calling accuracy^{8,9}, highlighting the importance of understanding each tool's strengths and weaknesses to merge their outputs efficiently.

Currently, several initiatives are trying to characterise variants from real samples¹²⁴ to use as a golden set, to perform benchmarking analyses of these tools. The Genome In a Bottle Consortium (GIAB) is dedicated to this aim; currently, they provide a golden set to validate SNVs and indels, and are working on a new sample to characterise SVs, focused on deletions and insertions ≥ 50 bp¹⁹¹. These sets have different limitations: 1) the validated SNVs and indels are only located in conservative regions of the genome, and 2) they do not provide all the broad spectrum of SVs, limiting the benchmarking to deletions and insertions. An alternative to compensate for this lack of information is to create an *in-silico* sample, which is a simulated genome where the user can introduce known variants and choose their genotype length and properties.

All the results presented in this section have been obtained from two samples: 1) An *in-silico* sample generated by our group (detailed description in section 3.1), and 2) the GIAB sample (NA12878) with validated SNVs and indels (detailed description in section 3.3). We used the hs37d5 reference genome to create the BAM files, and processed them following the GATK Best Practices (further details in section 3.1.2, and 3.3 for *in-silico* and GIAB sample, respectively).

4.1.1. The software selected to perform the variant detection in GCAT samples

Currently, no variant caller can detect all genome variability, due to different deterministic factors, such as the size, variant type, coverage, and the strategy of detection, among others. So, in order to have a complete calling, a thorough characterisation of tools has to be performed. In this project, different variant callers were selected based on the strategy applied to detect genome variability (Split-read (SR), Read-depth (RD), Discordant-Reads (DR), *de novo* assembly (AS), Machine Learning (ML) or a combination), and the variant type. A description of the 19 tools can be found in Table 13.

| Variant Caller | Strategy | Variant type |
|------------------|-------------|----------------------------------|
| Haplotype Caller | SR, AS | SNVs, Indels, Mid DEL |
| Deepvariant | ML | SNVs, Indels, Mid DEL |
| Strelka2 | AS | SNVs, Indels, Mid DEL |
| Delly2 | SR, DR, RD | Mid DEL, DEL, DUP, INS, INV, TRA |
| Manta | AS | Mid DEL, DEL, DUP, INS, INV, TRA |
| Pindel | SR, DR | Mid DEL, DEL, DUP, INS, INV, TRA |
| Lumpy | SR, DR, RD | Mid DEL, DEL, DUP, INV, TRA |
| Whamg | SR, DR, ML | Mid DEL, DEL, DUP, INS, INV |
| SvABA | AS, SP, DR | Mid DEL, NO SV TYPE |
| CNVnator | RD | DEL, DUP |
| Popins | AS | INS |
| MELT | DR | TRP |
| ViFi | SR, DR | VIR |
| VERSE | SR, AS | VIR |
| Platypus | AS | SNVs, Indels, Mid DEL |
| VarScan2 | SR | SNVs, Indels |
| Genome Strip | SR, DR, RD | DEL, DUP, mCNV |
| Pamir | SR, DR, OEA | INS |
| AsmVar | AS | DEL, DUP, INS, INV, TRA |

Table 13. Variant callers evaluated to analyse the genome variability. To perform an accurate variant detection, we evaluated different variant callers based on their strategy and variant detection, covering all genome variability. SR= Split read; DR= Discordant Read; RD= Read Depth; AS= *de novo* Assembly; ML= Machine Learning; OEA; One End Anchored. The variant callers coloured in red were discarded for variant detection in GCAT samples.

All tools mentioned in Table 13 were executed with the *in-silico* or GIAB samples. In section 3.2, we detail how these tools were executed. We designed a methodology to improve

the variant detection accuracy, combining the outputs of multiple variant callers outputs (section 3.2.3.4). We discarded eight tools for three reasons: 1) The number of variants detected (Figure 15), 2) Inability to improve the results of the Logistic Regression Model (LRM) (Table 14 and Table 15), and 3) Computational limitations (further details can be found in section 3.2). Finally, we used **12 algorithms to detect genome variability**.

4.1.2. Variant classification according to size

The variant and library properties can affect the variant breakpoint resolution. SNV and indels detections can be combined at base-pair resolution from different samples in a multi-sample VCF file. However, as a general remark, the larger the size of the variant detected, more difficult it is to determine with high-resolution its breakpoint, which increases the difficulty to properly identify redundant variants among samples. For example, GoNL determined that indels < 20 bp could be combined at base-pair resolution across different samples. Meanwhile, larger sizes need to include a breakpoint-error.

Following the GoNL criteria, we evaluated the accuracy of indel detection of all variant callers (section 3.2.1), considering the variant positions at base-pair resolution and matching alternative alleles (detailed methodology in section 3.2.3.1). The library properties of the *in-silico* sample (Table 3) and GCAT samples were the same, except for read length (Table 8). Figure 15 shows the recall and precision distribution of variant calling tools divided by indel size.

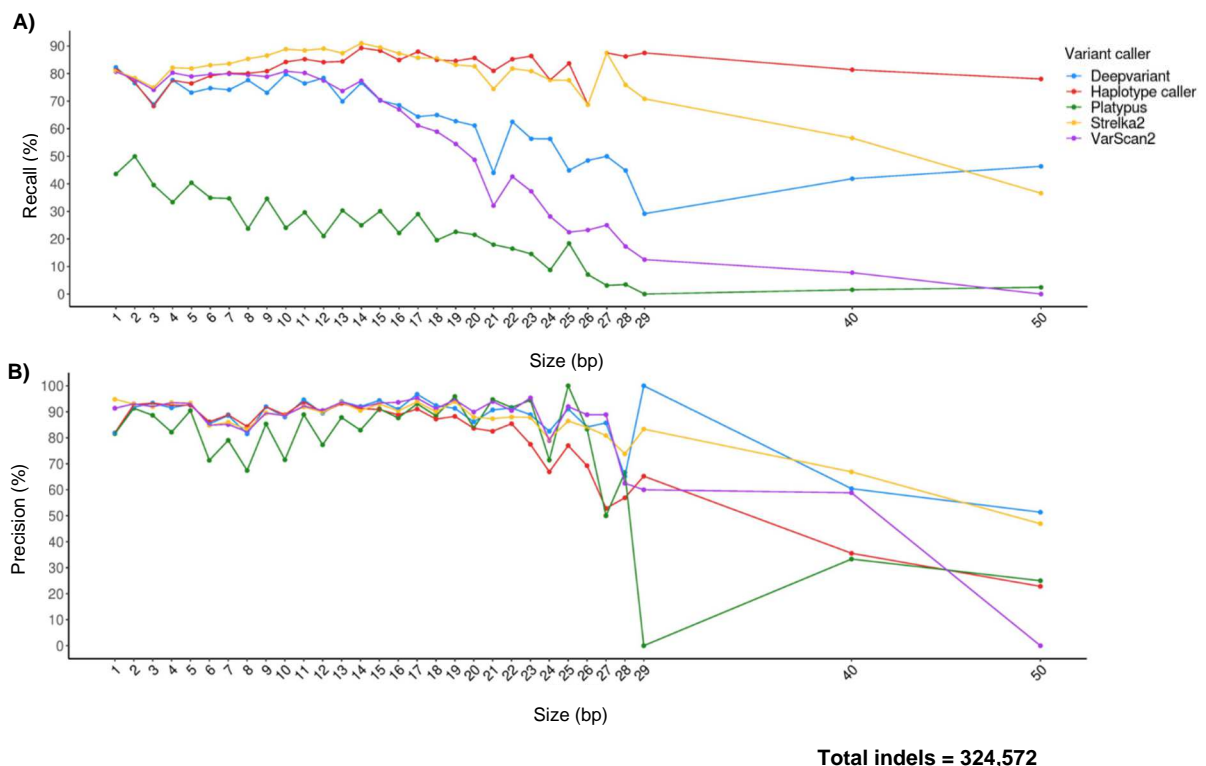


Figure 15. Variant caller indel detection benchmark: Accuracy to report at base-pair resolution with the same alternative allele. A) Recall distributed by indel size. While indel size increased, the recall decreased. Haplotype Caller detect around 80% of all indels independently of size. **B)** Precision of variant callers distributed by indel size. The precision was around 90% up to 20 bp, then this parameter decreased.

The recall decreased as the indel size increased (Figure 15A), showing the difficulties of variant callers to report the position at base-pair resolution, except for Haplotype Caller. In addition, Platypus detected a lower fraction of indels within all indel sizes as compared to other software, showing its inconsistency to report indels. The precision was around 90% until 20 bp for all variant callers (except Platypus) (Figure 15B), then it decreased, showing that the tools could work accurately to report variants with a **size up to 20% of the read length**. When the size surpassed this threshold, the variant callers included more false-positives (FP). Following these results, since the read length of GCAT samples was 150 bp: 1) Indels were considered **until 30 bp**. 2) Deletions **smaller than read length** (31 and 150 bp) were considered as mid-deletions (mid DEL). Mid insertions were not catalogued, because the variant callers did not report the length. 4) The remaining variant types > 30 bp were classified as Structural Variants (SVs), with the exception of large deletions, which were > 150 bp.

4.1.3. Benchmarking of variant callers and the Logistic Regression Model (LRM)

Variant callers that focus on the detection of SNVs and indels are highly accurate, due to small variant sizes, which facilitate a correct the read mapping compared to larger events^{8,85}. The challenge resides on the SVs, for which the accuracy varies between different detection methods, mainly due to the variant and library properties^{8,9,34,46}, showing inconsistencies across SV detections.

There are different studies that have performed benchmarking analyses of different variant callers^{9,46,128,130,192}, evaluating their detection properties, such as variant type and size, performance using different sequencing platforms, genomic context, and effect of NGS or TGS (Third Generation Sequencing) technologies. Usually, these studies also provide a ranking of variant callers showing the strengths and weaknesses of each caller in variant detection. These studies determined that there is not a single tool that can detect all structural variant types and sizes accurately.

Merging SV results from individual variant callers is a good strategy to increase the precision and recall of variant detection; in addition, the improvement is accentuated if each software applies independent detection methodologies (SR, DP, AS, RD, ML)^{9,46,76}. Currently, there are different tools to combine redundant SVs from different variant callers. SVmerge¹³³, SURVIVOR¹³⁵, Parliament2¹³⁶, or MetaSV¹³⁴ are tools that combine the outputs of variant callers by using unrelated decision logical rules, which usually are not optimal to obtain the best merge results, since the combination of imprecise callers could include noise in SV catalogues. For example, GoNL applied a logical rule-set to merge and filter the outputs¹⁶¹.

Alternatively, using machine learning approaches in the merging and filtering steps can improve the performance, since these methods analyse which variables are sufficiently “discriminative” to classify the variant as a true or false-positive, With better performance than simple logical rules. In this direction, FusorSV¹³² can merge and filter variants from independent tools, but unfortunately, it only uses two pre-set variables (SV type and size) for this purpose. Among the currently published panels, only the 1000 Genomes project used a machine-learning algorithm (Support Vector Machine model) to merge and filter indels and SVs¹.

In this context, we have performed a benchmark analysis by variant type, to evaluate the strengths and weaknesses of the variant detection software, and their accuracy in reporting SV

length and genotype. Finally, we created a machine learning model (Logistic Regression Model) using different discriminative features (detailed documentation in section 3.4.1.1 for SNVs/Indels and section 3.4.2.1 for SVs) to filter out the potential false-positive detections derived from inconsistent variant calling.

4.1.3.1. Benchmark analyses of SNVs and small Indels

Currently, variant calling tools can accurately detect SNVs and indels with a higher precision in SNVs^{128,129}. Although one of the most used variant callers is the Haplotype caller from the GATK consortium, new studies recommend other variant callers with better recall and precision results^{128,129}.

We performed a benchmarking of variant callers (section 3.2.1) for SNV and indel detections using an *in-silico* (artificial sample) and a GIAB (real sample) as golden samples. We classified/merged the results of the tools based on; 1) Variant type (SNV or Indel), 2) Position, and 3) Alternative allele (further details in section 3.2.3.4.1). Finally, we created with the GIAB data set, two versions of the Logistic Regression Model (LRM1, LRM2) for each variant type, including different variant callers. Finally, the models were validated using the *in-silico* dataset (details in section 3.4.1.1). For SNVs, Table 14 shows the recall and precision scores of each variant caller and LRM.

| A) | Variant caller metrics from GIAB sample (SNVs) | | | B) | Variant caller metrics from <i>in-silico</i> sample (SNVs) | | |
|----|--|------------|---------------|----|--|------------|---------------|
| | Variant caller | Recall (%) | Precision (%) | | Variant caller | Recall (%) | Precision (%) |
| | Deepvariant | 95.50 | 96.93 | | Deepvariant | 99.08 | 99.86 |
| | Haplotype caller | 81.90 | 96.92 | | Haplotype caller | 97.38 | 99.85 |
| | Strelka2 | 95.30 | 96.88 | | Strelka2 | 99.07 | 99.55 |
| | VarScan2 | 95.49 | 96.76 | | VarScan2 | 99.15 | 99.20 |
| | Platypus | 84.86 | 97.10 | | Platypus | 87.40 | 99.59 |
| | LRM1 | 95.68 | 96.93 | | LRM1 | 99.22 | 99.86 |
| | LRM2 | 95.50 | 96.93 | | LRM2 | 99.08 | 99.86 |

Table 14. SNV benchmark using different software and LRM combinations. To fit the model, we used the GIAB sample as the training dataset, then to validate the model, we used the *in-silico* sample (further details in section 3.4.1) **A)** Benchmark results using the GIAB sample as a gold standard. The models were trained with the GIAB sample. Haplotype Caller and Platypus had the worst recall values. **B)** Benchmark results using the *in-silico* sample as a gold standard. The models are tested with the *in-silico* sample. The validation results from the LRM1 and LRM2 models showed no differences in recall and precision.

LRM1: Deepvariant + Haplotype Caller + Strelka + Varscan2 + Platypus

LRM2: Deepvariant + Haplotype caller + Strelka2

As reported in previous studies¹²⁹, SNV detection was highly accurate, with **>96% precision**. The recall was slightly lower in the GIAB sample than the *in-silico* sample, but the trend was similar between the two samples. The recall and precision results from LRM1 and 2 (Table 14) indicated no correlation between the number of variant callers and accuracy improvement, mainly due to the high precision reported by variant callers individually. Due to irregularities in LRM decisions, the LRM2 only accepted the variants detected by Deepvariant, reflected by the same recall and precision results (Table 14). For this reason, these models were

not used to filter the SNVs from GCAT samples. However, to validate more variants in the SNV calling and improve the accuracy, we selected those variants supported by two or more callers in the GCAT calling (detailed description in section 3.7.2).

Indel detection accuracy was similar to SNVs. Precision in the GIAB sample was higher (>95%) than the *in-silico* sample (>88%) (Table 15); this could be due to the location where the variants were inserted. As mentioned in section 3.3, the variants evaluated in the GIAB sample were located in conservative regions, facilitating their detection. In contrast, in the *in-silico* sample, where indels were randomly distributed across the genome.

| A) | Variant caller metrics from GIAB sample (Indels 1-30 bp) | | | B) | Variant caller metrics from <i>in-silico</i> sample (Indels 1-30 bp) | | |
|----|--|------------|---------------|----|--|------------|---------------|
| | Variant caller | Recall (%) | Precision (%) | | Variant caller | Recall (%) | Precision (%) |
| | Deepvariant | 89.16 | 96.02 | | Deepvariant | 82.43 | 88.76 |
| | Haplotype caller | 88.15 | 95.82 | | Haplotype caller | 83.65 | 89.01 |
| | Strelka2 | 88.00 | 95.93 | | Strelka2 | 84.46 | 92.78 |
| | VarScan2 | 83.27 | 58.22 | | VarScan2 | 82.76 | 91.99 |
| | Platypus | 66.91 | 70.05 | | Platypus | 46.11 | 86.46 |
| | LRM1 | 89.17 | 96.00 | | LRM1 | 83.79 | 89.11 |
| | LRM2 | 89.25 | 95.94 | | LRM2 | 85.31 | 88.89 |

Table 15. Indel benchmark using different software and Logistic Regression Model (LRM) combinations. A) Benchmark of indels using the GIAB sample. The models were trained with the GIAB sample. The precision of Haplotype caller, Deepvariant and Strelka2 was >95%, showing a high accuracy in variant detection. **B)** Benchmark results of indels using the *in-silico* sample. The models were tested with the *in-silico* sample. The recall was lower than in the GIAB results, decreasing to >83%. The precision decreased until >88%, evidencing the effect of size in variant detection.

LRM1: Deepvariant + Haplotype Caller + Strelka + Varscan2 + Platypus

LRM2: Deepvariant + Haplotype caller + Strelka2

The LRM1 and LRM2 slightly improved the values obtained from the individual callers, with the exception of Strelka2 and VarScan2, which showed a higher precision by themselves (Table 15B). They had better precision than LRMs because most of the complex indels (ex: two contiguous indels) detected by Deepvariant and Haplotype caller were badly called. Also, the differences in recall and precision between LRM1 and LRM2 were not significant enough, which means that the inclusion of VarScan2 and Platypus in the LRM did not significantly improve indel detection (Table 15B). Thus we used LRM2 to deplete the indel detection in the GCAT variant calling.

Chromosome X was analysed separately from the autosomes. For example, in X chromosome, Deepvariant resulted in noisy indel identification, producing downstream errors in the LRM2 decisions. For this reason, in X chromosome, we accepted those indels detected by at least two software.

After evaluating the benchmarking results, only **Haplotype caller, Deepvariant and Strelka2, were selected to detect the SNVs and indels in the GCAT samples** for two major reasons: 1) They were the best callers in SNV detection, and 2) in indels, the LRM2 generated with those callers improved the **precision and recall detection in autosomal chromosomes** slightly, in comparison to the callers individually.

4.1.3.2. Measuring the breakpoint-error of each variant caller in SV discovery

The detection strategies applied by variant callers determined the breakpoint resolution of SVs. Those strategies can be categorized into four principal groups: 1) Split-Read (SR), 2) *de novo* Assembly (AS), 3) Discordant-Reads (DR) and 4) Read-depth (RD). Thus, determining a breakpoint-error of variant callers is paramount of interest, because it helps to combine redundant SVs of different variant callers.

We evaluated the resolution to report the breakpoint of several SV types, by using the *in-silico* sample, since the exact position of the SVs was known. A breakpoint-error was estimated for each variant caller, using the F-score metrics, an harmonic mean of precision and recall values, which provides the variant detection quality. The highest F-scores determined the breakpoint-error used for each variant caller (detailed description in section 3.2.3.3). Table 16 shows the breakpoint-errors selected for each variant caller.

| Callers | Strategy | Breakpoint Resolution (31-150 bp) | Breakpoint Resolution (150 bp ≥) | Breakpoint Resolution (50 bp ≥) | Breakpoint Resolution (50 bp ≥) | Breakpoint Resolution (50 bp ≥) | Breakpoint Resolution (50 bp ≥) |
|----------|----------|-----------------------------------|----------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| | | Mid DEL | DEL | DUP | INV | INS | TRA |
| Delly2 | SR+DR+RD | ±10 | ±100 | ±10 | ±10 | ±10 | ±300 |
| SvABA | AS+SR+DR | ±10 | ±100 | ±10 | ±10 | ±10 | ±10 |
| Manta | AS | ±10 | ±50 | ±20 | ±10 | ±10 | ±200 |
| CNVnator | RD | None | ±300 | ±100 | None | None | None |
| Whamg | SR+DR+ML | ±10 | ±10 | ±10 | ±10 | ±10 | None |
| Lumpy | SR+DR+RD | ±10 | ±100 | ±50 | ±10 | None | ±200 |
| Popins | AS | None | None | None | None | ±10 | None |
| Pindel | SR+DR | ±10 | ±10 | ±10 | ±10 | ±10 | ±10 |

Table 16. Selected Breakpoint-error of each variant caller by SV type. SV detections did not have base-pair resolution. To combine different variant callers, we first determined the accuracy to correctly report the SV position. The breakpoint-error allowed us to determine if a SV detected by different tools as the same since the breakpoint-errors overlapped.

None: The software does not detect this variant type.

Selecting the most accurate breakpoint-error by each variant caller allows combining redundant SVs efficiently. Table 16 demonstrated that CNVnator was the variant caller with the worst breakpoint resolution, consistent with previous studies^{77,193}. On the other hand, all variant callers that used **SR, AS, or both, showed high breakpoint resolution**, with translocations as the SV type which was reported with less resolution. **The tool which reported breakpoints more accurately in every SV type was Pindel**, proving the superiority of combining different detection strategies (Table 16).

4.1.3.3. Accuracy of detection of SVs by size

The SV size is a factor that plays a key role in its ability to detect the SVs. As the SV size increases, the mapping quality of the reads decreases, leading to misinterpretations and false-positive detections⁸. For example, the *de novo* Assembly (AS) and Discordant-Reads (DR)

strategies can detect a broader range of SV sizes than others⁹. Therefore, each variant caller is better suited to detect specific size ranges of SVs, depending on the strategy used.

To obtain a representation of variant caller performance for each SV size, we evaluated the F-score of each caller, grouped by SV type and size. We divided the SVs from the *in-silico* sample by sizes ((30-50], (50-75], (75-100], (100-125], (125-150], (150-300], (300-500], (500-1000], (1000-2000], (2000-3000], >3000), and evaluated their F-score by each size (further details in sections 3.2.3.4.2 and 3.2.3.4.3) (Figure 16).

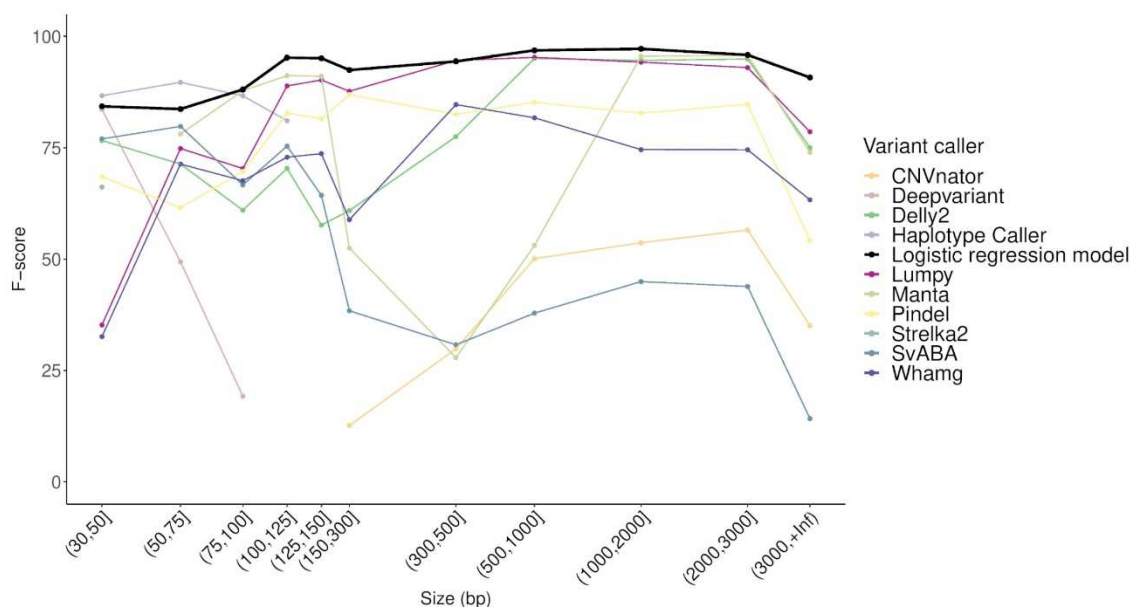


Figure 16. Overview of Structural Variant (SV) discovery distributed by size. The F-score is calculated using the recall and precision of each algorithm grouped by sizes. The F-score of variant callers fluctuated across SV sizes, improving their detection at size ranges between 100-150 bp.

Figure 16 reveals that **variant callers exhibited their best performance in detecting variants with sizes between 100-150 bp**. Besides, Delly2 and Manta showed the F-scores at sizes ≥ 500 bp, highlighting their efficacy to detect large SVs. Lumpy demonstrated high accuracy to detect SVs at sizes between 75 to > 3000 bp, covering efficiently all size ranges (Figure 16). CNVnator detected deletions and duplications of at least 125 bp, mainly due to the Read Depth strategy. Overall, **considering the size as a variable, the Logistic Regression Model outperformed each of the individual tools in SV detection across all SV sizes, obtaining F-scores > 90%** (Figure 16). These results showed the relevance of size in SV discovery performance.

Particularly, the variant callers showed different performances according to the SV type (Supplementary Figure 11). Regarding Deletions, the variant callers detected all size ranges efficiently. For Deletions > 150 bp, Pindel, SvABA and Whamg, had decreasing F-scores as size increased. Lumpy and Delly2 maintained their F-scores above 90%, except for largest sizes (Supplementary Figure 11). For Duplications and Inversions, depending on the size, the variant callers experimented different performances. For example, for duplications, Whamg had an F-score of 76% between 500-1000 bp, and of just 4% for size ranges of 150-300 bp (Supplementary Figure 11). However, **the LRM outperformed all variant callers individually in SV discovery across all SV types and size ranges**, highlighting the importance of building the LRM by including different size ranges as a variable (Supplementary Figure 11).

4.1.3.4. Benchmarking analyses of SVs between variant callers, GoNL strategy and Logistic Regression Model

The properties of variants (i.e. size, type) or sequencing libraries (i.e. read length, coverage, insert size)^{9,46}, affect the recall and precision in SV detection algorithms, reaching up to 89% of false-positive in some SV types and sizes⁹. Combining the variant caller outputs could thus improve SV detection accuracy, obtaining better SV catalogues. For example, SVmerge or FusorSV demonstrated an improvement in recall and precision by combining the output of different callers. Nevertheless, these results could be further improved in performance by including more variant and algorithm properties.

To improve the variant detection accuracy of our calling, we designed a custom merge of different variant callers for each SV type based on; 1) The variant detected was located in the same chromosome between variant callers, 2) The overlapping breakpoint-error positions, and 3) An SV size in reciprocal overlap (RO) of at least 80% between algorithms (further details in section 3.2.3.4.2). Finally, to filter potential false-positive detections, a specific Logistic Regression Model (LRM) was developed for each SV type (section 3.4.2), including discriminative variables such as the size, the number of tools that detected the same variant, among others (complete description of discriminative variables in Table 5). Figure 17 illustrates the recall, precision and F-score distribution of each variant caller, the Logistic Regression Model, and logical rules.

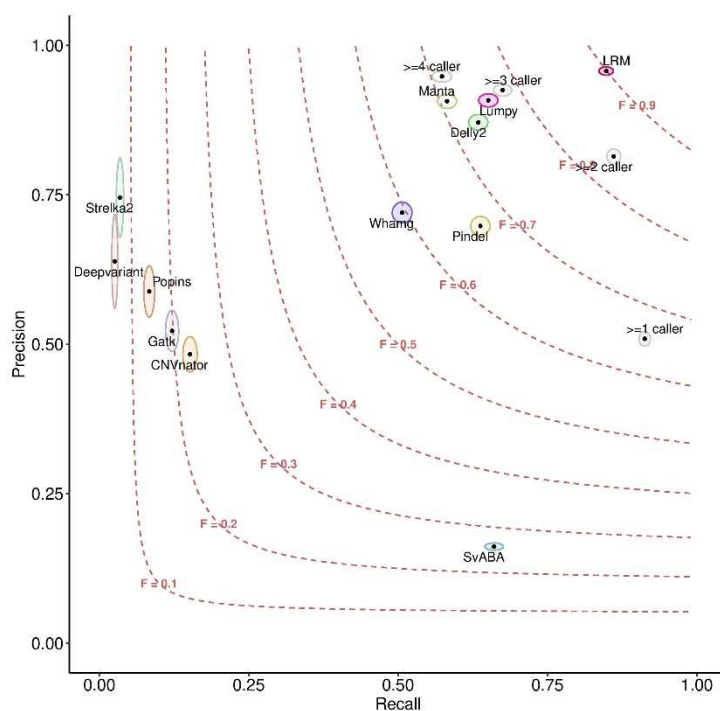



Figure 17. Overview of the Variant callers and Logistic Regression Model (LRM) benchmarking. General description of all variant discovery, evaluating the precision and recall of the algorithms and strategies considered in the project. The LRM outperformed all variant callers and logical rules strategies, obtaining an F-score of 0.9 and precision up to 0.95, without losing recall.

>=: Logical rules (ex: >= 2 callers, at least two callers and methods detect the same variant).; *LRM*: Logistic Regression Model.;  Shaded Area : Confidence interval (CI) area of each algorithm.; *F* = F-score.

These results demonstrated the LRM improvement in contrast to variant callers individually or using simple logical rules (Figure 17). Considering all variants together, the LRM obtained an **F-score of 0.9** and **precision of 0.95, without losing recall**, showing that LRM increased the strengths of all variant callers individually (Figure 17). The combination strategies by logical rules (\geq), such as ≥ 1 caller, improved the recall to 0.91 but at the cost of a heavy decrement of precision, to near ~ 0.5 . Besides, the ≥ 2 caller strategy followed by the GoNL project increased just 2% in recall, but obtained $\sim 14\%$ less precision than the LRM model, (Figure 17). This result suggested that GoNL included a large number of false-positives in their catalogue. However, in this analysis, the recalls for SV type-specific software, such as CNVnator or Popins, and those that detect small SVs (average length between 30-150 bp) were markedly lower, due to their inability to detect all variant types.

Particularly, **the LRM accuracy varied across SV types** (Supplementary Figure 3). For Deletions and Insertions, the LRM outperformed all strategies and callers individually, highlighting the relevance of building LRMs to improve variant discovery (Supplementary Figure 3A, B, C). For example, for insertions, **the F-score of LRM was 18% larger than the ≥ 3 callers rule**, the second largest F-score obtained. However, for inversions, duplications and translocations, the LRM performance was similar to the individual variant callers and logical rules (Supplementary Figure 3D, E, F). A particular instance of this is that to duplications, the LRM and the ≥ 2 caller strategy obtained similar accuracy values, with the LRM only 1.3% more precise. For inversions, the LRM performed better than the most accurate variant caller, highlighting that logical rule combinations decreased the recall and precision compared to variant callers individually (Supplementary Figure 3D). Finally, for translocations, no difference was appreciated between strategies, showing that while variant callers improved the variant discovery, the model filtered fewer SVs.

4.1.3.5. Benchmarking of genotyping between variant callers and Logistic Regression Model

Reporting an accurate genotype is necessary for population studies such as GWAS, phasing, and linkage disequilibrium analysis^{116–118,194}. For this reason, we analysed the genotyping precision of each caller and Logistic Regression Model (LRM), using as a gold standard the genotype reported by the *in-silico* sample (further details in section 3.2.3.5 and section 3.4.2.3). Figure 18 illustrates the genotype error of the tools and LRM.

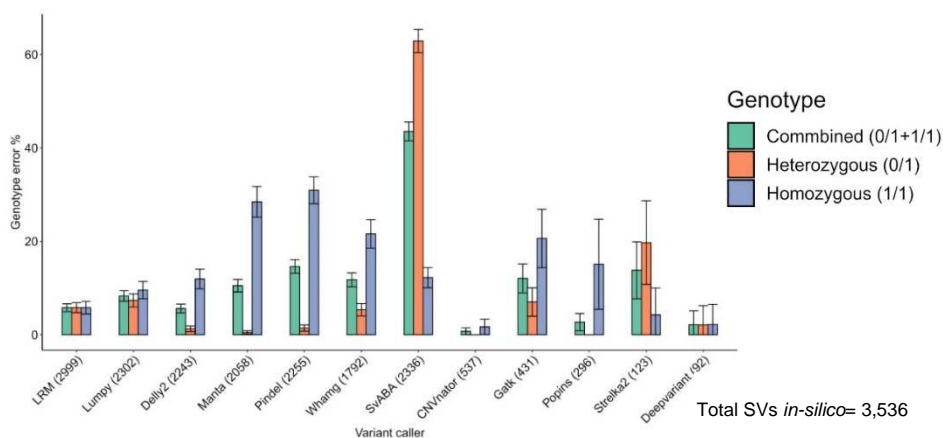


Figure 18. Genotype benchmarking between Logistic regression model (LRM) and variant callers. The genotype strategy of LRM proportionally, outperformed the accuracy compared to all variant callers, considering the number of variants genotyped, including 2,999 variants with only 5.6% of genotype errors.

The LRM genotyped ~85% of all *in-silico* sample SVs, with a genotype error of only 5.6% (Figure 18). Proportionally, the LRM genotype strategy was the most precise in comparison to variant callers individually. Besides, evaluating the genotype accuracy across SV type, the genotype concordance of LRM was higher than other variant callers (Supplementary Figure 4). Although compared with the ≥ 2 caller strategy, the LRM did not improve the variant discovery in duplications (Supplementary Figure 3E), the LRM was able to reduce the genotype error to under 20%, suggesting that our custom genotype strategy was well-designed. Moreover, for duplications, the genotype error of Manta and Pindel was 100% for heterozygous variants, which indicates a bias in their genotyping procedure. In the same way, Whamg did not report heterozygous insertions or Pindel in translocations. SvABA generated high genotype errors in all SV types (Figure 18), suggesting that their genotyping strategy had room to improve. These results demonstrated that the combination of variant callers' genotypes was the best strategy to reduce genotype errors.

4.1.3.6. Evaluation of the strategy used to generate the BAM files of the GCAT samples

Variant discovery using NGS can produce false discoveries^{62,113,195}, which drive to misinterpretation of the results obtained from variant callers. In this direction, BAM file generation plays an important role in decreasing the FDR. There are different error-prone steps in the generation of the data and BAM file, such as the production of duplicated reads in the PCR amplification step, the presence of not well-calibrated base quality scores due to systematic errors of sequencing machines, read misalignments, or artefact reads mapping to the reference genome⁷². Thus, the GATK Best Practices recommendations and the hs37d5 reference genome could decrease the false-positive detections without affecting the recall of variant callers.

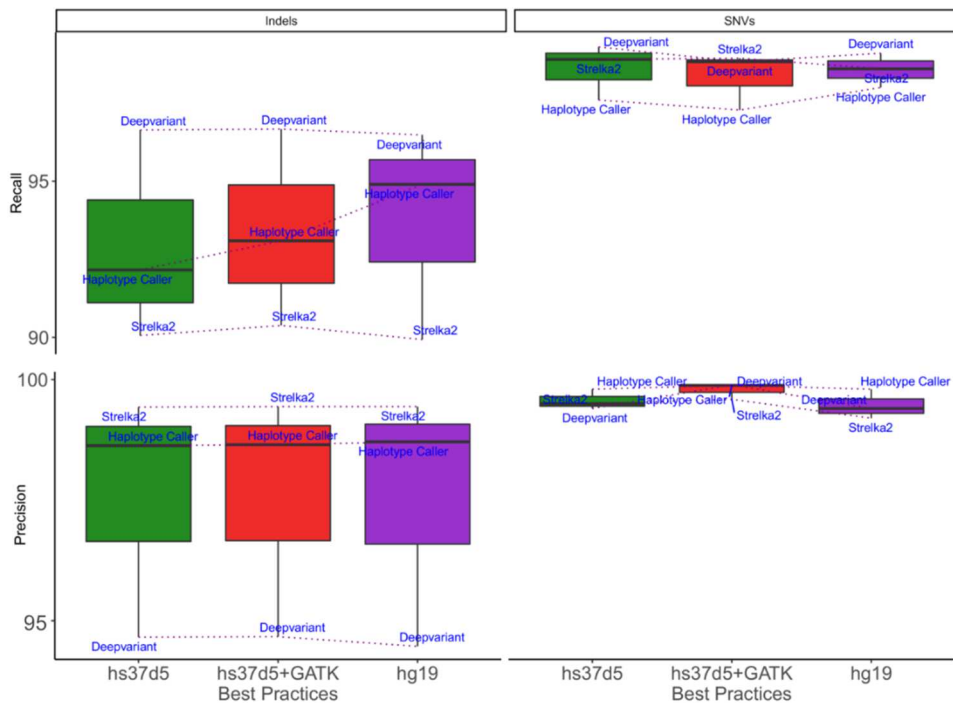


Figure 19. Accuracy in detecting SNVs and indels using different strategies to construct the BAM files. In the x axis there are the strategies followed to construct the BAM files. Then, we evaluated the precision and recall of indel and SNV for Haplotype caller, Deepvariant and Strelka2.

- - - Dotted line: Difference of Recall and Precision of a Variant caller between independent BAM file constructions.

To evaluate the best strategy to construct the BAM files in GCAT samples, we constructed the *in-silico* BAM file as follows: 1) Following the Best Practices of GATK using the hs37d5 genome, 2) Aligning the reads against the hs37d5 or 3) Aligning the sequences against the hg19 reference genome. Finally, we calculated the precision and recall for each variant caller and condition analysed (further details in section 3.2.3.6). Figure 19 and Figure 20 show the recall and precision categorized by variant type.

Figure 19 shows the effect of different BAM file constructions on SNV and indel detections. **Using the hg19 reference genome improved indel recall**, mainly for Haplotype caller, producing a better recall without losing precision. Nevertheless, the differences were not considerable between the different BAM file construction strategies. On the other hand, the recall of SNV discovery was near 100%, with **an improvement of precision when using the hs37d5+GATK Best Practices strategy**.

Overall, **the accuracy of SVs discovery did not vary when using different strategies in the construction of BAM files** (Figure 20). Mainly, the recall and precision results were constant in all SV types and variant callers, except for Delly2, for which the recall for inversions was lower when the BAM file using the hs37d5 reference genome. Besides, the strategy of hs37ds+GATK Best Practices improved the overall precision of translocation detection slightly.

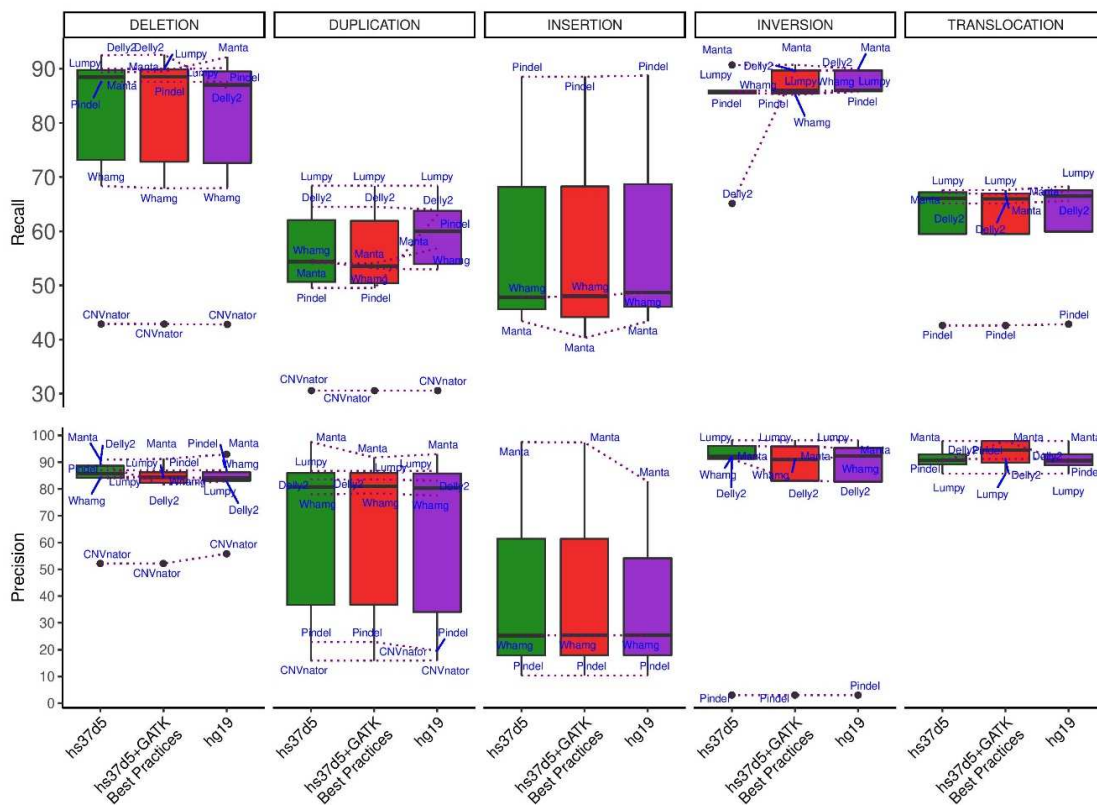


Figure 20. Structural Variant accuracy using different strategies to construct the *in-silico* BAM files.

- - - Dashed line: Difference of recall and precision of a Variant caller between independent BAM file constructions. ; ● Black dots: Outlayers.

These results (Figure 19, Figure 20) showed no difference in recall and precision when using different BAM file constructions in the *in-silico* sample. However, **the most conservative results were obtained with the BAM file generated with hs37d5+GATK Best Practices strategy**. Thus, to obtain accurate results in variant calling of real samples, we decided to create the GCAT BAM files applying the hs37d5+GATK Best Practices strategy.

4.2 Characterisation of the GCAT samples

Genetic variability is related to evolution and diseases/traits on humans^{8,9,113}, with a higher impact on the phenotypes from the Structural Variants (SV) than SNVs^{8,9}. The patterns of genetic variability among population groups are associated with geographic ancestry and can affect the disease risk or treatment efficacy differently^{6,7}. For this reason, more population-specific studies are necessary to characterise better this genetic diversity.

In this direction, this thesis was elaborated in collaboration with the GCAT-Genomics for life (GCAT) project, which provided the samples necessary to characterise at a genetic level the Iberian population with the aim of understanding the effect of genetic variability on Iberian-phenotypes. The data was collected from 19,267 participants at ages between 40-64 years in Northeast region of Spain (Catalonia). Of these, 16 % were self-reported as non-Caucasians, mainly from American-Hispanic origins (further description of the GCAT project in section 3.53.5.1 and in Obón-Santacana et al.¹⁷³).

From the variety of data collected, genomic data was utilized to understand and characterise the effect of genome variability on human-Iberian phenotypes. The GCAT Project produced two types of genomic data: 1) In 5,489 participants, the genomic profile was characterised using a genotyping array (Multi-Ethnic Global (MEGAEX2), and 2) In 808 participants, it was obtained genome-wide at **high coverage (30X)** (with HiSeq4000 machine) (detailed genomic data description in Table 8 and Obón-Santacana et al.¹⁷³). This second resource is highly important for two major reasons: 1) Whole-Genome Sequencing (WGS) allows for the study of the entire genome variability of the samples, including SNVs to large Structural Variants, and 2) The high coverage allows for a more robust variant discovery process and genotyping, making the SV detection more feasible.

In this section, we classify the results in four main blocks: 1) Sample filtering based on population and genetic features in order to obtain a set of representative Iberian samples, and the importance of high coverage in SV detection, 2) Relevance of integrating different variant caller outputs and deplete the variants using a Logistic Regression Model (LRM), 3) A detailed description of SVs in humans and their functional impact, and finally, 4) Validation of all variants recovered in the variant calling pipeline.

4.2.1. Filtering of GCAT samples and features of BAM files

The characterisation of the genetic architecture by different ethnic/population origins is essential for two reasons: 1) Each population has genetic particularities, mainly due to low-frequency ($0.01 < \text{MAF} \leq 0.05$) and rare ($\text{MAF} \leq 0.01$) variants, because those variants appeared recently in the population or natural selection affected the variant negatively in a specific population region^{2-4,50,110}. Besides, rare variants could increase the prevalence of rare or complex diseases in specific populations^{4,6,7}, which indicates the necessity to improve the genetic

characterisation of each population individually. In addition, 2) Generating population-specific panels of genetic variability improves the imputation of rare variants, in contrast to global panels, which impute mainly common ($0.05 \leq \text{MAF}$) and low-frequency variants accurately^{2-4,110}.

On the other hand, the sequencing coverage is determinant to perform accurate variant discovery and genotyping, improving variant detection of low-frequency and rare variants^{51,63,196}. Thus, sequencing at high coverage improves variant detection, mainly for Structural Variants (SV), where the mapping errors are frequent^{8,46}.

In this section, we explained the sample filtering of GCAT in order to obtain an Iberian-specific cohort. Next, we study the relevance of coverage in SV detection. Finally, we analyse the importance of generating a reference panel without chromosome Y in female samples.

4.2.1.1. Filtering the non-Iberian representative samples

The samples in the GCAT project were obtained in different areas such as rural, coastal or mountain, and in big cities of the Northeast region of Spain (Catalonia). We classified our samples as Iberian (IBS) following the definition in the 1000G project, which referenced this population as the “Iberian Population in Spain”¹.

To characterise the Iberian samples in the GCAT project, we first filtered the samples by quality to decrease false discovery rates, and select the Iberian population representatives. From 808 GCAT samples, **one sample** was discarded for not meeting all quality controls (further details in section 3.5.3). Secondly, we discarded all samples which were genetically different from the Iberian population. This was done by applying a Principal Component Analysis (PCA), in addition with the samples from the 1000G and Genomes and the Population Reference Sample (POPRES) samples (further details in section 3.5.3.3). Figure 21 shows the observed genetic variation between all analysed populations.

As previously mentioned in the features of the GCAT samples, 16% of participants were self-reported as non-Caucasian, mainly of American-Hispanic origin. Figure 21A and Figure 21B confirmed this information, showing that the vast majority of discarded samples was from Latin American populations. Besides, we also discarded participants with the birthplace in Spain, showing the importance to analyse the genetic background of participants, even when their self-reported as Iberians. Otherwise, we could include noise in the population-specific studies. After applying the PCAs, **we discarded 18 samples from another population**. Figure 21C confirmed that the remaining GCAT samples overlapped with the Iberian samples from the 1000G and POPRES projects, indicating that sample selection was made correctly.

Finally, two additional filtering steps were performed. First, we evaluated the family relatedness between the GCAT samples, by applying the Identity by Descent (IBD) test (detailed documentation in section 3.5.3.4), identifying two relationships, one full-sibling, and other first-cousin, respectively. **One of each pair was discarded** (further details in section 3.5.3.4) (Figure 21D). Second, we applied a PCA with all remaining GCAT samples, and **we filtered out two extra samples**, according to the mean $\pm 4\text{sd}$ criteria (section 3.5.3.5) (Figure 21E).

After these filtering steps, **785 of 808 samples** were used to characterise the genetic variability of the Iberian population and perform a panel of genetic variability Iberian-specific.

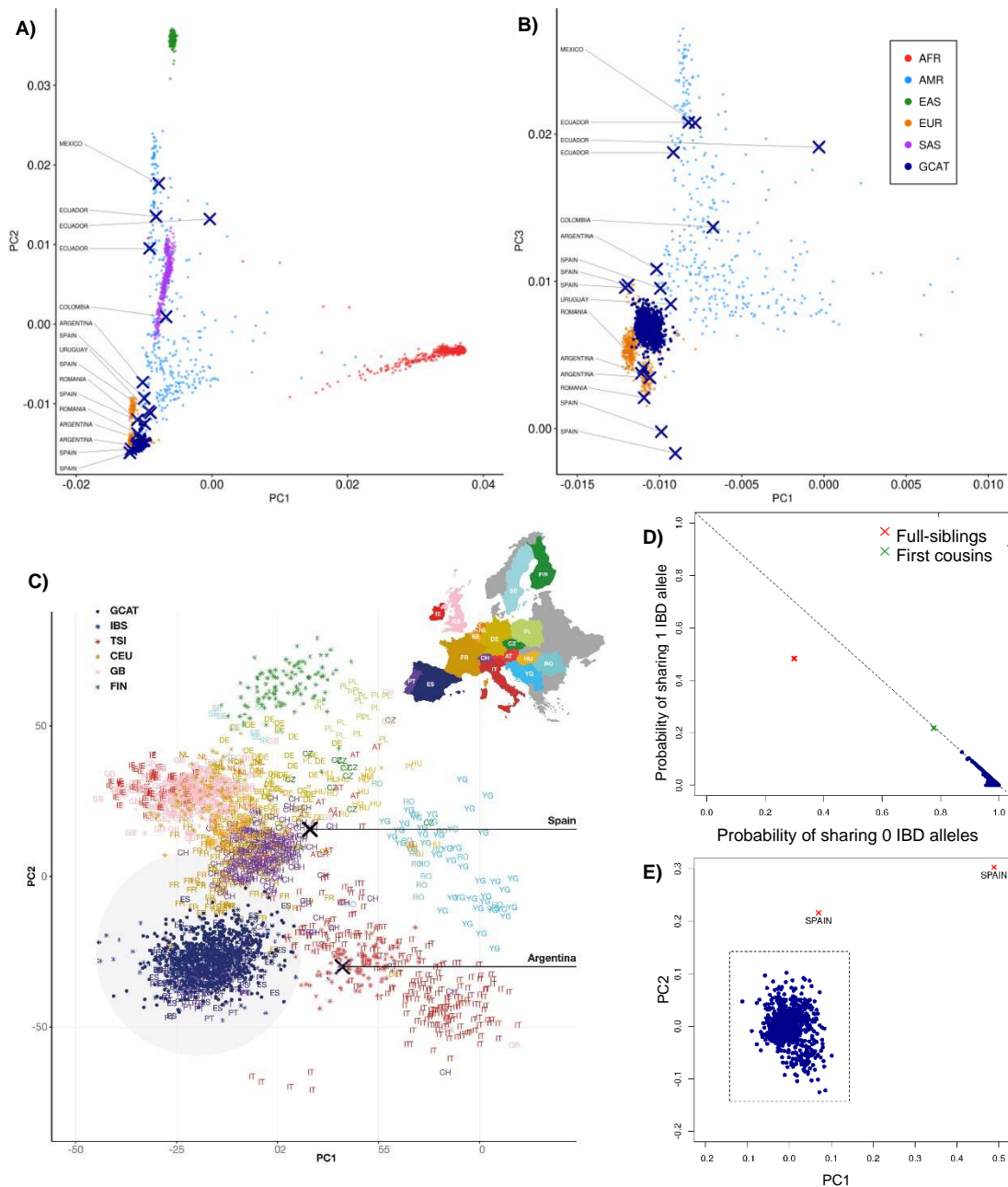


Figure 21. Evaluation of the homogeneity in the GCAT cohort. A) Distribution of the GCAT samples among 1000 Genomes, with PCA1 and PCA2. **B)** Distribution of the GCAT samples among 1000 Genomes, with PCA1 and PCA3. **C)** GCAT samples distribution among 1000 Genomes European samples and PROPES project using PCA1 and PCA2. **D)** Identity by Descent test (IBD). **E)** GCAT homogeneity, with PCA1 and PCA2.

X: Discarded Sample; *: 1000G European samples

4.2.1.2. Importance of coverage in variant detection

Library properties, such as insert size, read length, or coverage, influenced SV detection^{9,46}. Particularly, high coverage allows for a better genotype in heterozygous variants^{46,63,64} (section 1.2.4). Besides, *de novo* assembly algorithms require high coverage to be executed correctly, showing the relevance of coverage in variant and genotype calling approaches^{8,105}.

In this context, the effect of coverage on SV detection has been evaluated. This study was performed using 10 GCAT samples selected randomly, that was later downsampled at different coverages: 5X, 10X, 15X, 20X and 25X (Figure 22) (further details in section 3.5.4).

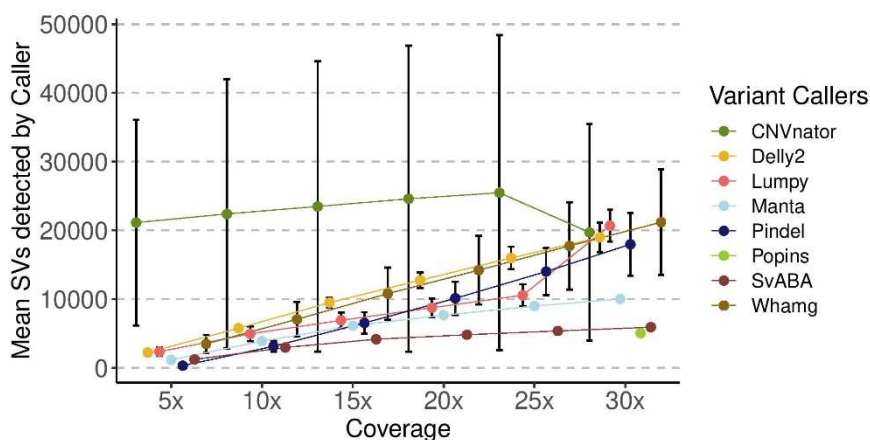


Figure 22. Variant calling performance at different coverages. High coverage (30X) allows detecting more SVs due to the number of signals available. Specific strategies, such as Read depth (RD), used by CNVnator, detected fewer variants at 30X, mainly due to their recall on coverage changes, discarding more potential false-positives. On the other hand, *de novo* assembly strategies (AS) used by Manta, SvABA and Popins, stabilize their detections at 30X.

Variant caller capacities are influenced by coverage, **up until seven times more variants can be found using a coverage of 30X in comparison to one of 5X** (Figure 22). However, **not all variant caller strategies were equally affected by coverage**. For example, *de novo* assembly tools, such as Manta, SvABA and Popins, did not benefit from coverages higher than 30X to detect more variants, as opposed to the Discordant-Read (DR) or Split-Read (SR) strategies, used by Delly2, Lumpy, Pindel, and Whamg. Besides, using **high coverages, improves the accuracy of variant detection by Read Depth (RD) strategies** (ex: CNVnator), filtering potential false-positives (Figure 22).

Finally, **the detection of different SV types was differently affected by the coverage**; for example, the performance of Whamg in detecting inversions was profoundly affected by coverage. The same happened with the detection of translocations by Lumpy, or *de novo* insertions with Pindel, evidencing a possible room of improvement in SV detection (Supplementary Figure 5). However, increasing the number of SV discoveries is not correlated to true-positive (TP) variants. For this reason, to obtain an accurate SV catalogue, combining different variant callers could decrease false-positives, enabling the correction of FP due to high coverages.

4.2.1.3. Improving variant detection in chromosome X

The study of sex chromosomes has been a challenge in next-generation sequencing studies. One of the major problems is the short-read alignment to the reference genome given that the X and Y chromosomes have high similarity in some regions such as PseudoAutosomal Regions (PARs). This can produce technical artefacts, affecting downstream analyses on variant calling⁷³.

By default, all reference genomes include both the X and Y chromosomes in sequencing studies⁷⁴. However, including the Y chromosome in read alignment from female samples, can

have a negative impact on the variant detection of the X chromosome, due to the scavenger effect of the Y chromosome^{73,74}. In order to verify these statements, the X chromosome coverage was evaluated in four GCAT samples (two female and two male samples), by constructing the BAM files with two hs37d5 reference genome versions, one including the chromosome Y and another without it (Table 17). In addition, the performance of variant detection in X chromosome was evaluated with Haplotype caller using three female samples (further details in section 3.5.2.1).

| Total Coverage in X Chromosome | | | |
|---------------------------------------|------------------|---------------------|-------------------|
| Sample and gender | With ChrY | Without ChrY | Difference |
| JID250 (female) | 40,510,267 | 40,558,146 | 0.11 % |
| JID259 (male) | 22,805,867 | 23,786,713 | 4.12 % |
| JID297 (female) | 40,741,933 | 40,784,267 | 0.10 % |
| JID439 (male) | 20,244.035 | 21,141,948 | 4.24 % |

Table 17. Distribution of X chromosome coverage by sample gender.

The coverage distribution in the X chromosome was not equal across samples gender. The female samples included nearly double of reads than male samples (Table 17). Besides, the **scavenger effect of the Y chromosome on female samples was not critical**; only 0.1% of reads were aligned in another genome region. In contrast, if we discarded the Y chromosome in the alignment of the male samples, the X **chromosome increased the number of reads by around 4.2%** (Table 17).

| Total Variants in X Chromosome | | | |
|---------------------------------------|------------------|---------------------|-------------------|
| Sample and gender | With ChrY | Without ChrY | Difference |
| JID250 (female) | 123,785 | 124,185 | 0.32 % |
| JID297 (female) | 119,278 | 119,616 | 0.28 % |
| JID436 (female) | 97,080 | 97,312 | 0.23 % |

Table 18. SNV and indel detected by generating a BAM file with a gender-based reference genome.

Thus, **variant detection in female samples improved slightly without the Y chromosome** (Table 18), allowing a better variant characterization on the X chromosome. Due to improvements of the coverage and variant detection of the X chromosome, these analyses suggested that variant calling could benefit from aligning samples to their gender-based reference genome, in the cases where the gender of the sample is known.

4.2.2. A general description of the variants recovered after applying the merge strategy and the Logistic regression model

Variant detection using short-reads data (100-300 bp) has been applied widely to analyse human genome⁶¹. The three major variant groups associated with genome variability are SNVs, Indels (< 50 bp), and Structural Variants (SVs) (≥ 50 bp). Fortunately, the variant callers designed to detect SNVs and Indels are accurate due to the small variant sizes, which do not affect critically the read mappability on the reference genome, in contrast, the larger SVs.

The detection of SVs using short-reads is a challenging problem, producing high False-Discovery Rate (FDR) of 9-89% and recall between 10-70%^{8,61,62,77,99,100}. Different studies recommended integrating multiple variant callers for SV detection to improve both the FDR and recall, showing the possibility to obtain a curated SV set^{5,8,9,46,76,95,161}.

In this context, to improve the variant calling of the GCAT samples, we performed a multi-variant calling. The calling of SNV and small indels (1-30 bp) were performed by sample with Haplotype caller, Strelka2, and Deepvariant (section 3.6.1). We considered that a SNV or indel was consistent between these callers if they shared the same position and reference-alternative alleles. Next, for SNVs, we determined as false-positive all variants detected by one caller. For indels, we applied a Logistic Regression Model (LRM), classifying the variants as true-positive (TP) and false-positive (FP) (further details in section 3.7.2). The calling of mid deletions (31-150 bp) and SVs was performed by the callers described in section 3.6.2. We carried out output integration by sample and SV type. Redundant mid-size deletions and SVs were combined if they fulfilled the following conditions: 1) Same SV type, 2) The breakpoint-error overlapped between software, 3) The variant size had at least 80% reciprocal overlap between callers. Then, to determine if the variant was a TP, we applied a specific LRM for mid dels and SVs (detailed documentation in section 3.7.2).

After integrating all variant caller outputs by sample and variant type, we generated a multi-sample VCF, grouping all samples by each variant type independently. Then, for each variant, if the proportion of TP determined by LRM was $\geq 50\%$ in the GCAT cohort, we classified it as a TP; otherwise, the variant was an FP (further details in section 3.7.3). Finally, we discarded (i) the variants which were not in Hardy-Weinberg Equilibrium (HWE), (ii) all variants with more than 10% of missingness, and (iii) monomorphic variants (section 3.7.4). Figure 23 shows the variant filtering results.

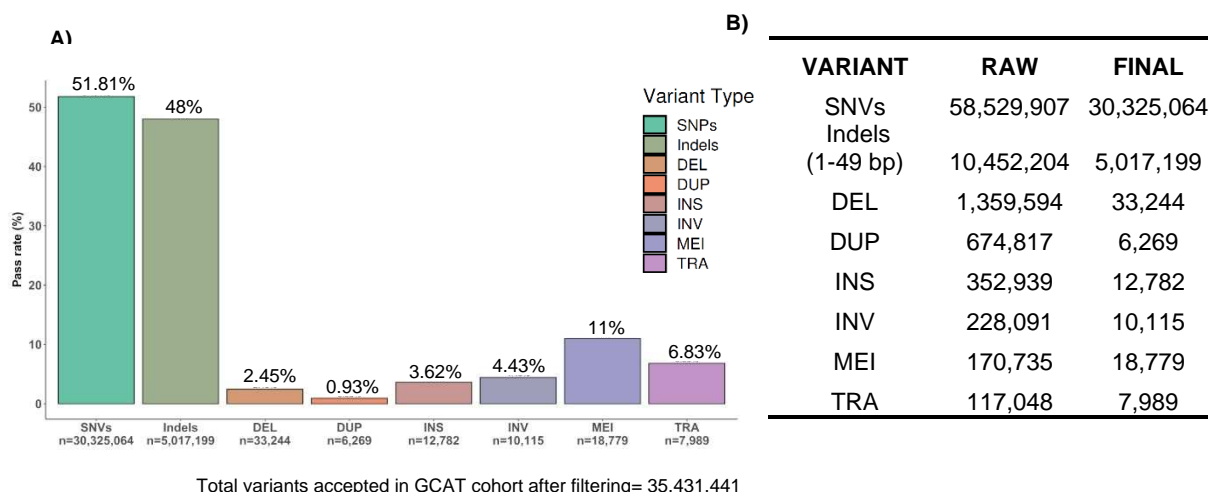


Figure 23. Properties of variant calling in the Iberian cohort. A) Pass rate after variant caller integration and Filtering with a Logistic Regression Model. **B)** All Variants detected by variant callers and recovered after the variant filtering.

Figure 23 was obtained discarding PARs and Y chromosome variants due to the high homology with chromosome X regions, prone to false discoveries. The variants were classified by sizes, with SNVs as the change of a single nucleotide; Indels, all variants of sizes 1-49 bp, short indels; and Structural Variants (SV) ≥ 50 bp.

We detected **71,885,335 variants**. After filtering, **we accepted 49.3%** of them, highlighting a better PASS rate in smaller sizes (Figure 23A). Our approach accepted around 50% of variants discovered in SNVs and indels, showing the necessity to use more than one caller to increase the recall and precision (Figure 23A). Besides, we estimated that **3.5M SNVs and 606K indels** were present in a single typical genome. Particularly, we **accepted 3.07% of detected SVs**, demonstrating a hard filter by our LRMs, which was particularly **strict with DUPs** (Figure 23). Also, the FDR in SV discovery using short-reads could be higher than 89%. Next, **the 0.25% (89,178 variants) of 35,431,441 variants accepted were SVs**, demonstrating less recurrence in the genomes than SNVs and indels.

The assembly errors produced by short-reads are known. In inversions, 24 genome locations of hg19 were described as prone to false-positive detections¹⁹⁷. In **our variant calling, 13 regions of these 24 were detected**. After the filtering process, those 13 regions were discarded, demonstrating that the LRM model was highly accurate to filter false-positives.

Analysing the genetic architecture of population-specific regions allowed us to better understand the genome variability effect on phenotype between different populations⁶. Rare and low-frequency variants are usually associated with population-specific phenotypes, highlighting the importance of characterizing populations in detail³. Concretely, in the GCAT samples, 78.92% of all recovered variants had a MAF < 5%, with **50.18% of singleton and doubleton variants** (Figure 24). This amount of variants could provide new insights into the genetic effect on phenotype in the Iberian population.

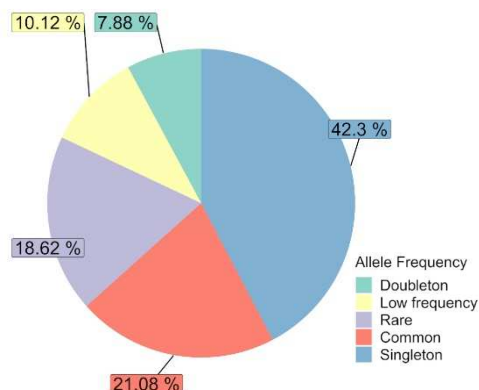


Figure 24. GCAT variants distributed by allele frequency.

4.2.3. A detailed description and characterisation of Structural Variants into the Iberian cohort

In the last decade, variant discovery and genotyping using short-read sequencing has been applied in numerous population projects and studies. Unfortunately, the technology limitations could not accurately detect Structural variants (>50 bp) due to the coverage and read-length limitations. Considering SNVs, the genomic variation between two human genomes is around 0.1% (4M SNVs per genome¹⁰), a difference that increases to 1.5% with SVs^{8,50,198}. Also, the SVs affected between 3-10 times more nucleotides than SNVs^{9-11,105}, showing their potential effect on human phenotypes.

Nowadays, the sequencing costs of samples at high depth (30X) using NGS (short-reads) have decreased, allowing for the improvement of SV detection in populational studies^{9,51,64,196}. However, these methods have some limitations, such as an underrepresentation of SVs discoveries in repetitive regions. These limitations could probably be ameliorated with the long-reads generated with Third-Generation sequencing technologies (TGS)^{8,21,61}. However, these technologies are expensive and cannot be used at large-scale or in population studies. Thus, the use of sequencing technologies has resulted in SV characterisation lagging, in contrast with SNVs and indels.

For this reason, a recent publication in Nature Reviews “the SV in the sequencing era”⁹⁵, suggested to generate complete catalogues of SVs^{11,65} and to find new relations of SVs with diseases. In this context, different initiatives such as gnomAD-SV or Ira M.Hall labs provided novel SV catalogues^{11,65}, describing their effects on the human genome. On the other hand, Almarri et al.²¹ tried to characterize the distribution of SVs in different populations. In this direction, we focused on the characterisation of the 89,178 SVs discovered in the GCAT cohort in order to (i) provide new variants not already discovered by the community, and (ii) to improve the functional insights of the impact of SVs in humans.

4.2.3.1. Comparison of variants detected against different repositories

Since the first Copy Number Variant (CNV) detected in the 2000s, the SV discovery has been limited due to technological limitations⁹⁵. Different public archives such as dbVar⁹⁶ and Database of Genomic Variants (DGV)^{96,97} collected all SVs validated by the scientific community. Besides, new projects such as gnomad-SV, Human Genome Structural Variant Consortium (HGSV) or Ira M.Hall lab characterised more SVs, generating new catalogues.

In order to evaluate the number of new SVs detected by our project, we used the databases mentioned before, considering an SV as the same if positions overlapped in a window of ± 1000 bp, the SV type was coincident and there was at least 80% of length overlap across the projects (further details in section 3.8.1.2). Besides, we used the dbSNP database to evaluate the number of new SNPs and Indels detected by our project (further details in section 3.8.1.1).

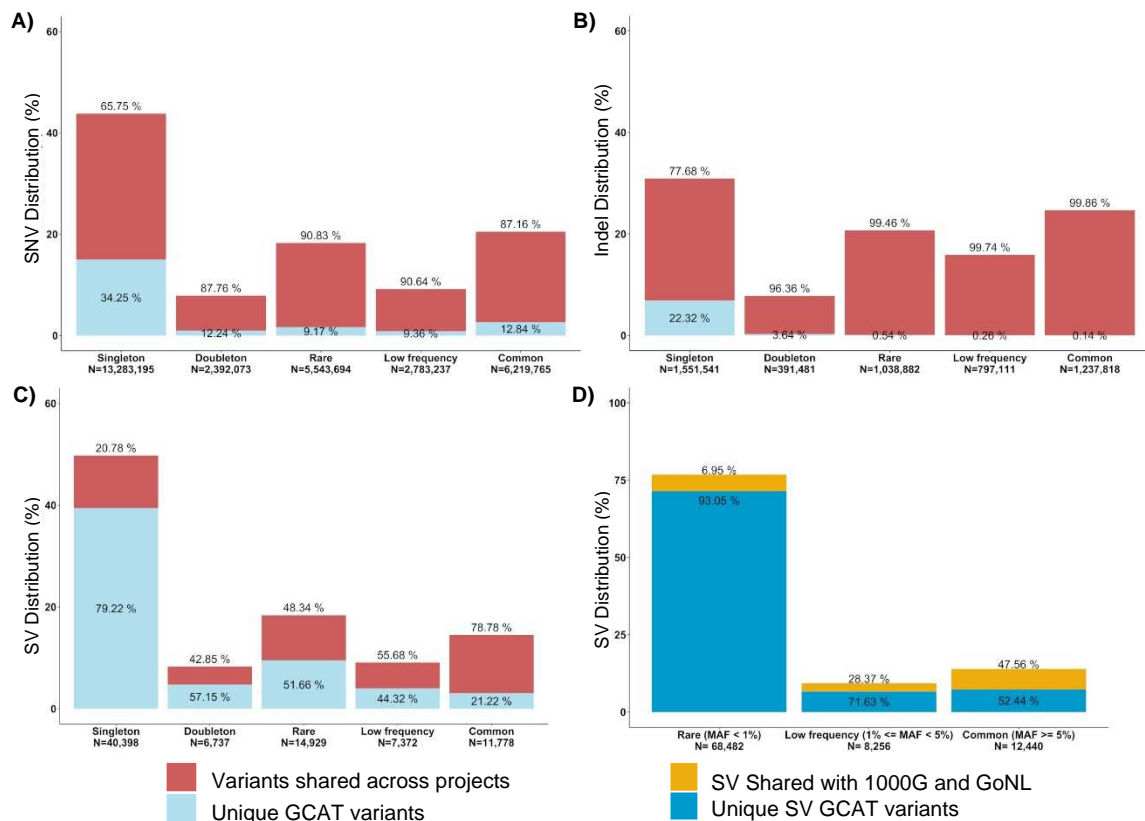


Figure 25. New variant contribution in comparison to popular repositories. A) The SNVs **B)** and indels compared with dbSNP. The majority of new SNVs and indels discovered were Singletons, demonstrating that the variants with MAF > 1% were already discovered. However, 12.84% of common SNVs from the Iberian catalogue were not already catalogued, indicating possible population-specific variants **C)** In SVs, the contribution of new SV discovered was greater than SNVs and indels, being rare SVs (MAF < 1%) as the lesser characterised (Translocations were discarded). **D)** Considering the SVs included in reference panels, 52.44% of common SVs and 71.63% of low-frequency SVs from Iberian catalogue were not already included.

19.18% of SNVs and indels included in the Iberian catalogue (35.3M) were not already classified in dbSNP, 84.32% of which variants (MAF < 1%). Besides, singletons from SNVs and indels, 34.25% and 22.32% respectively, comprised most of the new variants (Figure 25A, B). However, **12.84% of common SNVs were not already included in dbSNP** (Figure 25A), possibly due to population-specific variants. Notwithstanding, evaluating SNVs that overlapped only by position but that not match the alternative allele, resulted in **less than 1% SNVs with MAF > 1%** (Supplementary Table 9), indicating that most of MAF > 1% not described were polymorphic variants.

On the other hand, **61% of SVs from the Iberian catalogue were new**, a majority of which with MAF < 1% (88.31%). **21.22% and 44.32% of common and low-frequency variants were new (Figure 25C)**, demonstrating that SV discovery was skewed compared to SNVs and indels. Nevertheless, all catalogued SVs were not included in reference panels, and thus, were not available for GWAS studies. In this context, 14.60% of SV from the Iberian catalogue were shared in 1000G and GoNL. **52.44% of common and 71.63% of low-frequency SVs were unique from the Iberian catalogue** (Figure 25D), highlighting the interest to convert this catalogue to a reference panel, in order to increase the chances to include more SVs in GWAS studies.

4.2.3.2. Variant size ranges detected by our methodology

Detecting large SVs using short-reads is challenging due to alignment limitations, leading to false-positive detections and misinterpretations. For example, discovering large inversions is not feasible because many of them are inserted between repetitive regions and segmental duplications, hindering their detection⁶⁷. Different projects such as 1000G or gnomAD-SV distributed the SVs by sizes, showing variant discovery limitations in some ranges. Figure 26 shows the SV distribution by sizes in our project.

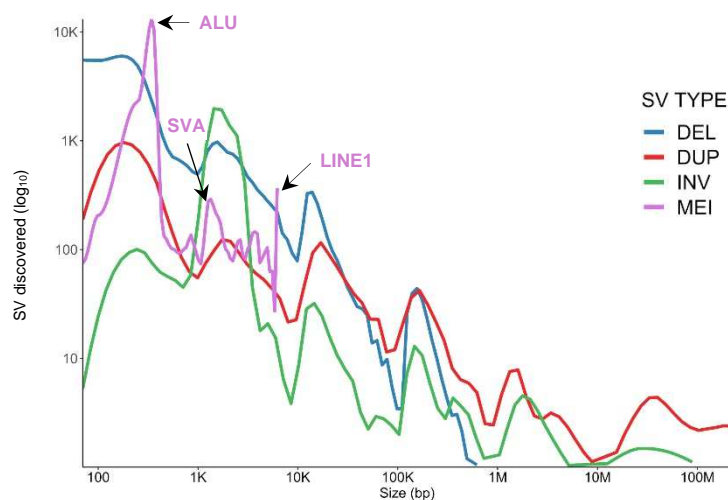


Figure 26. Structural Variant distribution by length. The bulk of SVs detected was between 100 bp and 10K. At large sizes, the SV discoveries decreased. The three peaks observed in MEIs corresponds to ALUs, SVA and LINE1, respectively. MEI: Mobile Element Insertion; DEL: Deletion; DUP: Duplication; INV: Inversion

The median size of SVs discovered in the Iberian cohort was of **291 bp**, which was consistent with gnomAD-SV, suggesting that SV detection using short-reads was favourable at

small sizes. The median size of SVs type discovered was categorised as follows: **312 bp DEL**, **584 bp DUP**, **1,531 bp INV**, and **279 bp MEIs**.

The size distribution between all SVs detected showed a larger number at small sizes (Figure 26), in contrast to other projects, such as 1000G or dataset of Ira M. Hall lab. We detected **more inversions at sizes between 1-1.6 Kbp** (Figure 26) than 1000G, gnomAD-SV and dataset of Ira M. Hall lab, highlighting the relevance to use different variant callers to improve the recall. However, we detected duplications and inversions bigger than 10M bp, with a duplication of 197 MB located in chromosome 2 position 33,141,357, as the largest SV. These uncommonly large sizes suggested that these were false-positives or miss classified variants; however, these represented just **0.086 % of all SVs**. Finally, the length analysis elucidated that **the three peaks in MEI discoveries** were associated with the three main groups of transposons, such as ALU (250-350 bp), SVA (1-1.3 Kbp), and LINE1 (6 Kbp) (Figure 26). Overall, we could detect SVs at different sizes, which allowed us to improve the SV characterisation in the Iberian cohort.

The median of nucleotides affected by SVs (without including INS and TRA) in a genome was around **211M bp**. This represented **6% of all nucleotides in humans**, in contrast to 4M with SNPs. These results reinforced the theory that SV had a large influence on human phenotypes due to the large number of bases altered in a single genome. **Common variants (MAF \geq 5%) affected 66.89% of nucleotides**, with the **only 0.27% of rare variants (MAF < 1%) the group in which fewer nucleotides altered** in a single genome. Besides, 97.87% of 211M bp **were affected by duplications**, followed by **deletions with 1.65%**. This suggested that CNVs could be the SV type with a larger impact on human phenotypes.

4.2.3.3. Structural Variant distribution in the Iberian cohort

The number of SVs per genome and its distribution in a population has been updated as the sequencing technologies evolved. The gnomAD project, using NGS technology estimated around 7,439 SVs per genome. However, studies which used TGS technology estimated more than 20,000 SVs per genome^{66,67,95}. These results highlighted the difficulties in determining the SV number per genome and their distribution in a cohort accurately.

In this direction, we analysed the allele frequency distribution into the GCAT cohort and the median number of SVs per genome using short-reads, considering that SVs in inserted repetitive and homologous regions were under-represented in this dataset due to NGS technology limitations.

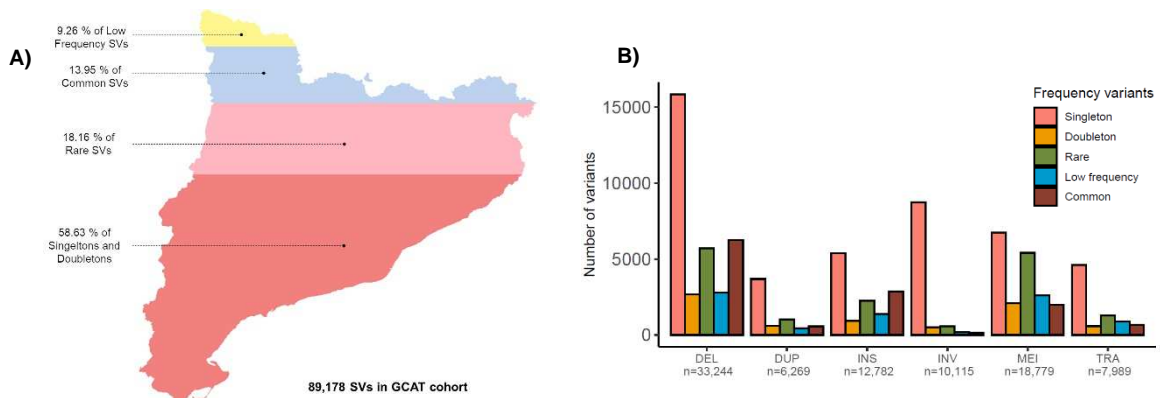


Figure 27. Structural Variant (SV) distribution in the GCAT cohort. A) Allele frequency distribution in the GCAT cohort. **B)** SV categorized by type and distributed by allele frequency in the GCAT cohort.

76,79% of detected SVs were rare variants (MAF < 1%), of which **58,63% were sample-specific** (singletons and doubletons) (Figure 27A). However, gnomAD-SV identified around 92% of SVs as rare. These results suggested the necessity to characterise SVs in a specific population, in contrast to global projects such as gnomAD or 1000G, to improve the allele frequency particularities of each population individually. The predominantly variant type were **deletions with 37.3% of all variants,** followed by **MEI (21.1%), insertions (14.3%), inversions (11.4%), translocations (8.8%), and duplications (7.1%)** (Figure 27B). In all SV types, the singletons were predominant; besides, **91.3% of all inversions were singletons and doubletons** (Figure 27B). These results showed the difficulties of detecting recurrent inversions in human genomes using short-reads. This result was consistent with previous studies, where the bulk of inversions were located in repetitive regions⁶⁶, limiting their detection with NGS technologies.

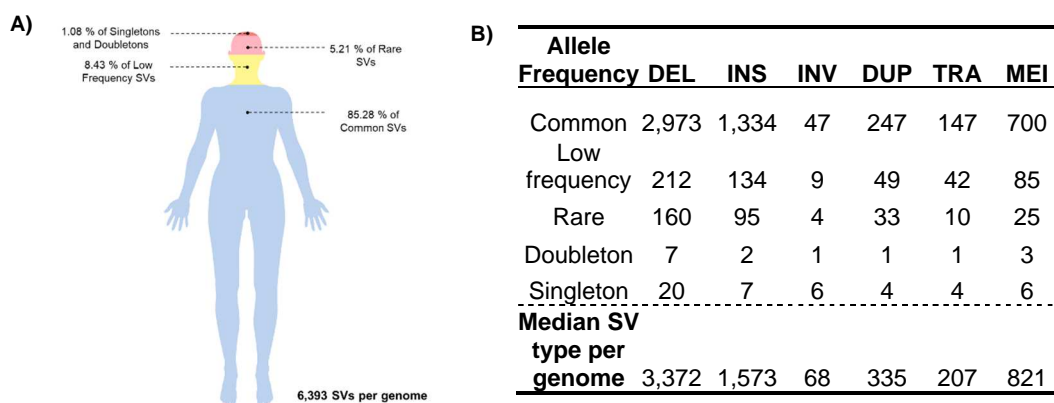


Figure 28. Structural Variant (SV) distribution in a human genome. A) Allele frequency distribution per genome. **B)** Median of SVs per genome, categorized by SV type and distributed by allele frequency.

We estimated that **6,393 SVs were included in a single genome.** The allele frequency in a genome was opposite to its population distribution, with **common variants as the most represented at 85.28%,** followed by **low-frequency variants at 8.43%** and **rare variants (MAF < 1%) at 5.21%** (Figure 28A). This distribution was concordant with the observations of the 1000G, in which 1-4% of variants per genome had a MAF < 5%¹. **Deletions were the most represented SVs and inversions, the less recurrent** (Figure 28B).

4.2.3.4. Functional impact of Structural Variants

Structural Variants (SVs) are the major biological variability source both at a population and at individual level^{9-11,75,95}. Besides, SVs can modify gene expression, topological associating domains (TAD) or disrupt protein-coding genes, producing an impact on gene function or developing different rare or complex diseases and developmental disorders^{5,11,65,66,95}.

In this context, we evaluated the functional impact of the SVs detected in the GCAT project (89,178 SVs) using the AnnotSV⁷⁵ tool, focused on annotating all SVs using different repositories such as Refgene, Online Mendelian Inheritance in Man (OMIM), GeneHancer database³⁷ (all databases detailed in Table 12) and provided the pathogenicity levels of SVs, based on morbid genes described in the literature (section 3.12.2.1). Also, AnnotSV provided the pLI and HI values to evaluate the loss of function intolerance of genes (further details in section

3.12.2.1). We benchmarked the annotation using SnpEff¹⁵², which can annotate DEL, INV, and DUP (section 3.12.2.1).

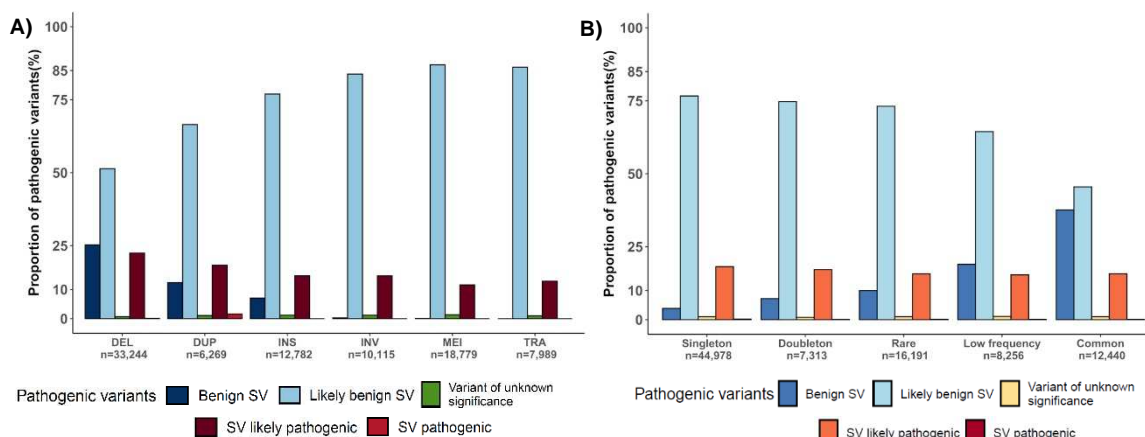


Figure 29. Pathogenic predictions of SVs. A) Pathogenic distribution by SV type. **B)** Pathogenic distribution by allele frequency.

The pathogenic effect of SVs on humans is mostly unknown. The majority of **SVs (87.7%)** are reported as “likely benign” or “likely pathogenic” (Figure 29), demonstrating ambiguity in their classification. Besides, deletions and duplications were better characterised, because more research was done in these SV types (Figure 29A). Similarly, as the allele frequency increased, the SVs knowledge rises too, reflected in an increase of the “benign SVs” predictions for common variants (Figure 29B). It should be noted that **17% of discovered SVs were classified as “likely pathogenic”**, showing opportunities to associate new SVs to different diseases. However, the datasets used to annotate these SVs still have room for improvement. For example, the inversion 11q13.2 is strongly associated with obesity and common diseases¹⁹⁹, and AnnotSV reported this variant as likely benign, demonstrating that some likely benign SVs could, in fact, be related to diseases. We found that **~16% of all SVs per genome, independently of allele frequencies, could be associated with a pathogenic effect**, evidencing a deleterious effect of SVs also at high allele frequencies (ex: common, low frequency) (Supplementary Table 4).

The functional interpretation of genome variability is related to protein-coding genes. However, the variants in intronic or intergenic regions, such as regulatory regions could also affect the gene function. In this context, we annotated the SVs to know their potential effect on gene function (section 3.12.2.1).

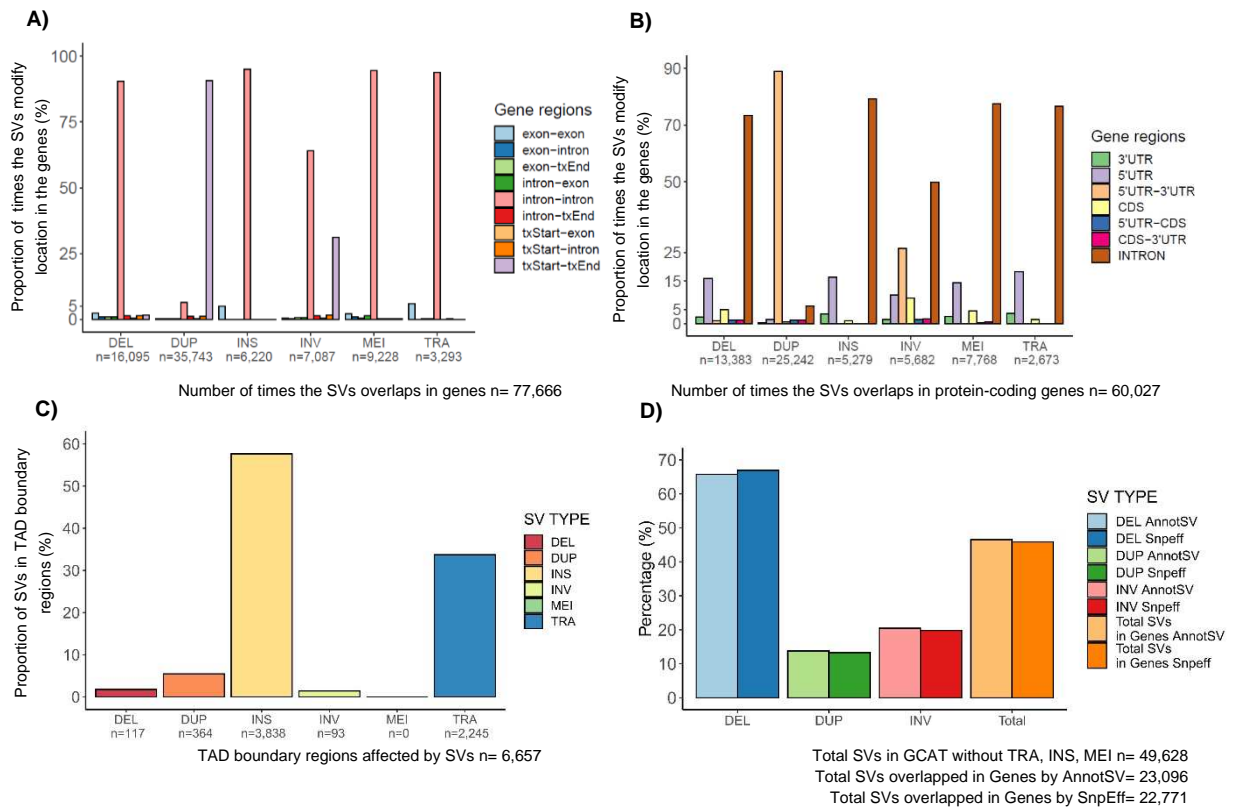


Figure 30. Structural Variant annotation in the human genome. A) Structural Variant location in protein-coding and non-protein-coding genes. **B)** Structural Variant location in protein-coding genes. **C)** Structural Variant distribution in Topologically Associated Domains (TAD) boundaries. **D)** Proportion of SVs annotated in genes, using AnnotSV and SnpEff. By SV type, deletions overlapped more gene regions than other SV types.

Tx = Transcript; UTR = untranslated regions

46% (41,672) of detected SVs in the GCAT project overlapped with genetic regions.

The deletions (36.64%) was the most predominant, followed by Mobile Element Insertions (MEI) (21.2%), *de novo* Insertions (14.4%), inversions (11.34%), translocations (8.8%), and duplications (7.62%). Similar results were obtained with SnpEff, showing the concordance between different annotation tools (Figure 30D). Besides, the detected SVs in the gnomAD-SV and Audano et al⁶⁷ projects overlapped with genetic regions in 47.7% and 46.6%, respectively, in concordance with the GCAT results. Also, for gnomAD-SV, 49.89% of SVs that modified genes were deletions, followed by MEIs (26%), similar to the GCAT catalogue. However, duplications and inversions represented 15.54% and 0.26% respectively, differing from our results, showing the importance of using more variant callers to improve variant discovery. On average, **we estimated that 2,868.36 SVs overlapped with genes per genome.**

The number of genes modified by SV was correlated to the SV length. As we mentioned in section 4.2.3.2, the duplications and inversions were the types with the largest SVs (Figure 26), so one SV could overlap with multiple genes. In this context, 5.42% of SVs (2,260 SVs) were involved in multiple gene modifications. In the whole GCAT project, **the SVs affected 21,003 genes** (protein-coding and non-protein-coding genes), and inversely, **SVs overlapped gene 77,666 times** (Figure 30A). Duplications were the most predominant, affecting 46% of 77,666 times that a SV overlapped a gene, in contrast to insertions (8%) and translocations (4.24%) (Figure 30A). **More than 88% of SVs overlapped only intronic regions**, highlighting that

duplications and inversions, due to their sizes, modified a large number of transcripts (Figure 30A).

Concretely, the modifications of protein-coding genes by SVs could lead to larger phenotypical consequences than those of non-protein-coding genes; for this reason, we further characterised coding genes. Our dataset included **15,246 protein-coding genes modified by 35,359 SVs (39.6% of all SVs from the GCAT catalogue)**. From those, a SV overlapped a gene 60,027 times (Figure 30B). As previously described, duplications modified more genes, and the majority of SVs overlapped intronic regions (Figure 30B). Besides, **7,695 SVs affected UTR regions**, with an **enrichment of 10-18% in 5'UTR**, except for duplications (Figure 30B). On average, per genome, **464 SVs affected UTRs**, which were distributed as 387.73 SVs common SVs (MAF > 5%), 45.54 low-frequency SVs ($1\% \geq \text{MAF} > 5\%$) and 33.34 rare SVs (MAF < 1%). Particularly, the coding gene regions (CDS) were modified less by SVs, only between 5-10% with the exception of duplications (93.52%) and inversions (36.05%) (Figure 30B). In the GCAT project, **3,135 SVs modified CDS regions**, with deletions as the predominant (32.72%), followed by duplications (27.36%), inversions (23.6%), MEIs (13.3%), and a small fraction of the other SV types. Also, disregarding singletons and doubletons, 659 SVs remained modifying CDS regions, showing selective pressure on the SVs in those locations. Finally, on average, **69.75 SVs modified CDS in a single genome**, distributed by **54.25 common SVs, 8.37 low-frequency SVs, 3.72 rare SVs, 0.54 doubletons, and 2.85 singletons**.

The SVs also affected the 3D structure of chromatin by modifying the topologically associating domains (TAD) and their boundaries, potentially leading to different diseases. In the GCAT dataset, **6,657 SVs affected TAD boundaries**, with enrichment of **insertions (57.65%) and translocations (33.72%)** (Figure 30C). Besides, MEIs were not found in overlap with TAD boundaries (Figure 30C), indicating that these events were underrepresented in those genome regions.

To determine the deleterious effect of variants is one of the main biomedicine goals. The predicted loss of function intolerance (pLI) is one of the more widely measure used for this aim. In order to know the deleterious effect of SVs in the human genome and their representation in the Iberian cohort, we evaluated the loss of function in protein-coding genes and its distribution in a single genome, taking into account the pLI and haploinsufficiency (HI) parameters provided by AnnotSV (section 3.12.2.1).

32.96% of 35,359 SVs modified protein-coding genes with high prediction loss of function intolerance (pLI) (Figure 31), corroborating the high impact of SVs on gene function. Deletions and Mobile Element Insertions (MEIs) were more deleterious than duplications and translocations (Figure 31A). Besides, as expected, **SVs with MAFs < 1% represented 79% of all pLI genes**, demonstrating selective pressure to variants with high gene function impact (Figure 31B). However, 21% of SV were $\text{MAF} \geq 1\%$, indicating different pathogenic levels (Figure 31B). On average, **746.6 SVs per genome were related to genes with the pLI effect**. They were divided into 623.50 common variants, followed by 65.70 low-frequency variants, 47.37 rare variants, 2.37 doubletons and 7.66 singletons. These results suggested different penetrance for SVs, with rare variants (MAF < 1%), found less predominantly in a genome, could have larger diseases implications. Additionally, **1,416 heterozygous variants predicted a haploinsufficiency (HI) effect on gene function**, predominantly deletions and MEIs (Figure 31C). As in Figure 31B, the majority of the HI resulted from SVs with MAF < 1%, reinforcing the

hypothesis of their deleterious effect on gene function and their contribution to human diseases (Figure 31D). On average, **per genome, 93.32 SVs were associated with a predicted HI effect, with common variants as the most represented at a count of 77.97, followed by 9.54 low-frequency, 4.55 rare, 0.3 doubletons, and 0.94 singletons.**

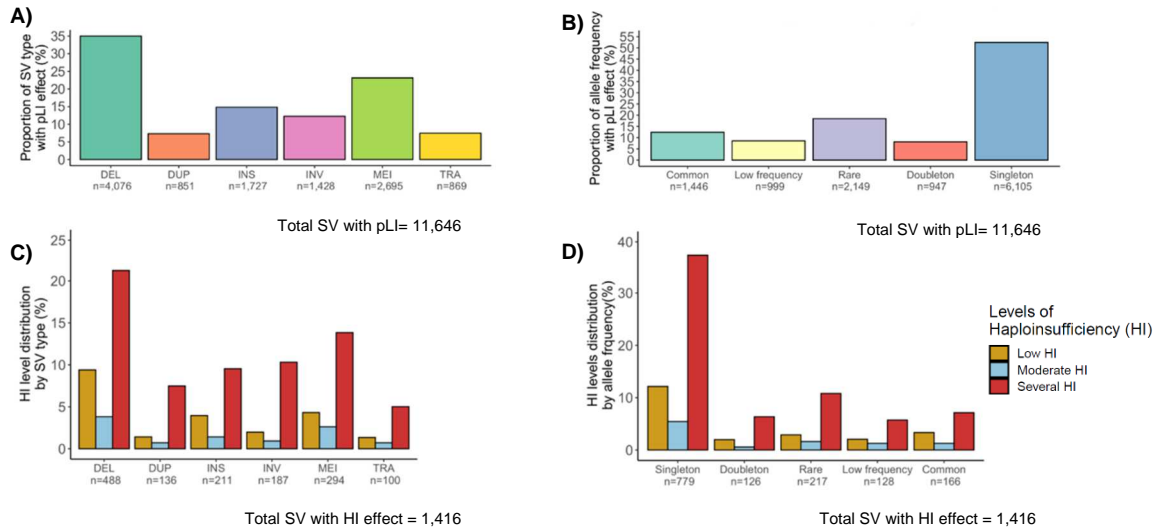


Figure 31. Deleterious effects of SVs in GCAT cohort. A) Prediction of loss of function intolerance (pLI), categorised by SV type. Predictions ≥ 0.9 indicate SVs with a high deleterious effect. **B)** Prediction of loss of function intolerance, categorised by allele frequency. **C)** Prediction of Haploinsufficiency (HI) effect, categorised by SV type. **D)** Prediction of Haploinsufficiency (HI) effect, categorised by allele frequency.

Genes with $pLI \geq 0.9$ and HI were “extremely loss of gene function intolerant” (Supplementary Figure 10). For this reason, associate which SVs affected these genes, we could improve GWAS, facilitating the identification of causal variants in complex diseases and improve the knowledge of SV effects on human complex diseases. In this direction, we evaluated the implication on phenotypes of the SVs with the most deleterious predictions (pathogenicity ≥ 4 , $pLI \geq 0.9$, and HI) on protein-coding genes. Additionally, we analysed which SVs were tagged by SNPs with high LD of $r^2 \geq 0.8$ using the GWAS catalog, in order to find alternative interpretations of variants associated with diseases and traits (Figure 33). Singletons and doubletons were discarded in both analyses (section 3.12.2.2).

Mental and muscular diseases were the top 10 diseases related to deleterious SVs (581 variants) (Supplementary Table 5), with deletions as the predominant at 40.97% (Figure 32A). Further research is necessary to understand the phenotypical effect of all SV types on diseases in order to enrich the OMIM database with new disease-associated variants.

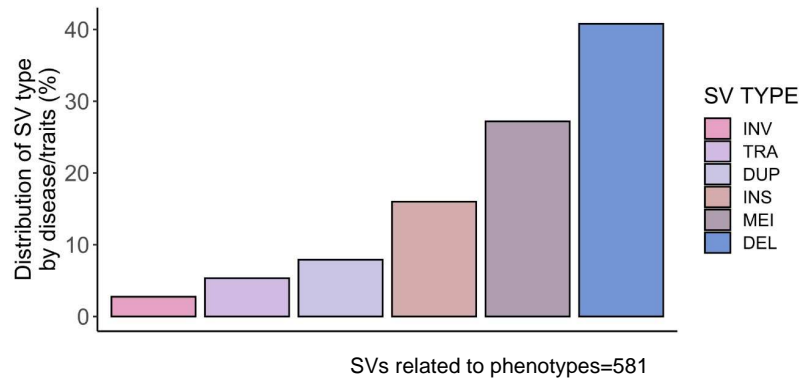


Figure 32. Prediction of the effect of Structural variants on phenotype using OMIM database. A) Structural variants related with a diseases in OMIM database. The SVs selected were those with $pLI \geq 0.9$ and HI and pathogenicity ≥ 4 .

Then, we evaluated the phenotypic effect of SVs tagged by SNPs and their distribution in the genome in order to predict the effect of SVs as casual variants. For this reason, we used the SVs tagged by SNPs with an LD $r^2 \geq 0.8$.

The SNPs from the GWAS catalog tagged with high LD were 3.72% of the SV in the GCAT dataset (36,887 SVs (MAF $\geq 1\%$)) (Figure 33A). The predominant type variants were deletions, insertions, and MEIs. Besides, most SNPs overlapped intronic and intergenic regions (Figure 33B). Finally, **51 (8.77%) of 581 SVs disease-related** (Figure 33A) **were tagged by 51 SNPs** in the GWAS catalog (Supplementary Table 6), showing the importance of using panels of genetic variability including SVs, to improve the performance of GWAS studies.

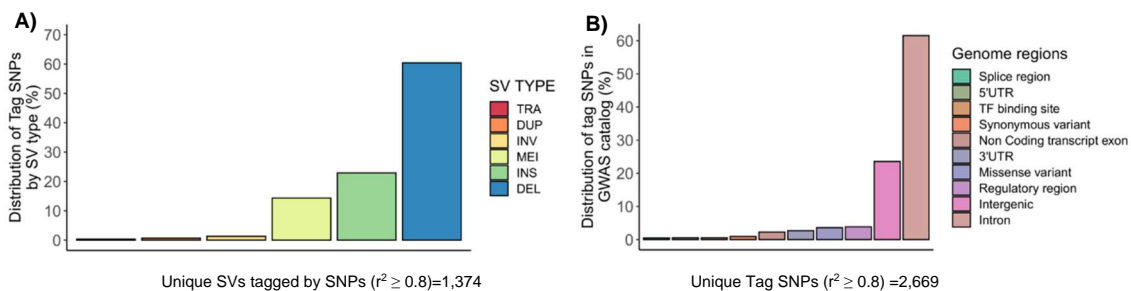


Figure 33. Structural variants tagged by GWAS catalog SNPs with high Linkage disequilibrium (LD ≥ 0.8). **A)** Structural variants tagged by SNPs, categorised by SV type. **B)** Tag SNP distribution by genome regions.

4.2.4. Validation of Iberian dataset

We performed different validations in all variant types to estimate the accuracy of our strategy to detect the genome variability.

4.2.4.1. Validation of SNVs and indels using the GCAT genotyping array data

Validation of SNV and indel detection with our WGS pipeline was performed by comparing WGS variant calling results with SNP-array calls in 570 samples, for which both WGS and GCAT SNP-genotyping array data were validated. Additionally, we assessed the genotype of matched SNVs and indels between both detection methods (section 3.8.2.1).

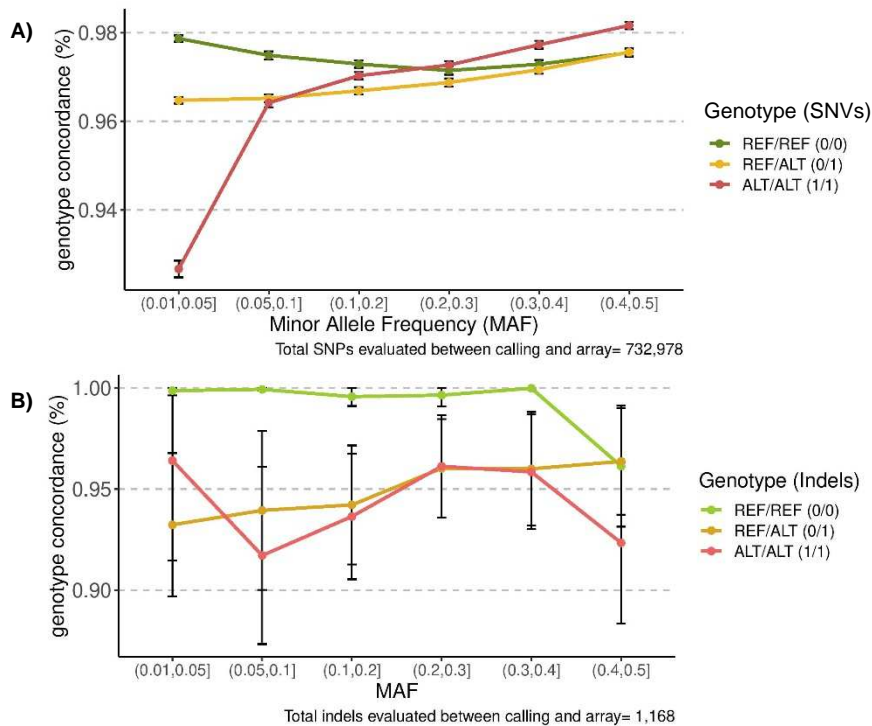


Figure 34. Genotype concordance between WGS variant calling and GCAT-genotyping array for SNVs and indels. We compared the genotypes obtained from our variant calling pipeline with those from SNP-genotyping array, in order to estimate the genotype accuracy of variant callers. **A)** Genotype concordance of SNVs between variant calling and GCAT SNP-array. **B)** Genotype concordance of indels between variant calling and GCAT SNP-array.

96.14% of SNVs from SNP-genotyping array matched with WGS variant calling, indicating that variant calling was highly precise to detect SNVs. Besides, few discrepancies were observed between genotypes, being **concordant for > 96% of all the genotype types** and frequencies, with the exception of low frequency (0.01, 0.05] homozygous alternative SNVs, where the concordance decreased to < 94% (Figure 34A). These results show that our SNV calling strategy, characterised by a high coverage (30X) WGS and an optimised variant calling pipeline, was able to produce accurate and precise variant detection.

86.8% of 1,168 indels (≤ 30 bp with MAF > 1%) from SNP-genotyping array matched with WGS variant calling. Further, **more than 90% of alleles** reported by variant calling were concordant with array genotypes, independently of allele frequency (Figure 34B). These results demonstrate that variant calling was highly efficient to detect and genotype indels.

4.2.4.2. Experimental validations of Structural Variants

4.2.4.2.1 Validation of deletions and duplications using Comparative Genomic Hybridization array (CGH array)

Large deletions and duplications (>20 Kb) detected by variant callers were validated using a Comparative Genomic Hybridization array (CGH array). The NA12878 sample from GIAB project was used as a reference sample to find probe intensity changes in 5 GCAT samples (further details in section 3.8.2.2). The results are presented in Table 19.

| Sample | All detected Del and Dup (>20 Kb) CGH | All validated Del and Dup | Deletions from variant calling | Validated Deletions | Duplications from variant calling | Validated Duplications |
|----------------|---------------------------------------|---------------------------|--------------------------------|---------------------|-----------------------------------|------------------------|
| JID054 | 53 | 22 (41%) | 13 | 12 (92%) | 40 | 10 (25%) |
| JID258 | 40 | 11 (27%) | 14 | 11 (79%) | 26 | 0 (0%) |
| JID398 | 36 | 18 (50%) | 13 | 11 (85%) | 23 | 7 (30%) |
| JID404 | 36 | 9 (25%) | 13 | 8(61%) | 23 | 1 (4%) |
| JID486 | 41 | 13 (32%) | 14 | 9 (64%) | 27 | 4 (15%) |
| Average | 41 | 14 (34.1%) | 13 | 10 (77%) | 28 | 4 (19%) |
| Total | 206 | 73 (34.5%) | 67 | 51 (76%) | 139 | 22 (19.5%) |

Table 19. Large Duplications and Deletions (>20kb) validated in 5 samples using a Comparative Genomic Hybridization array (CGH array). The CGH array assay allowed us to validate large (>20 Kb) Duplications and Deletions, where at least five probes are giving the signal for variant detection.

On average, we validated 34.5% of large duplications and deletions, and specifically, 76% of large deletions and 19.5% of large duplications. This difference between the number of validated deletions and duplications could be ascribed to known CGH array limitations in detecting duplications.

4.2.4.2.2 Validation of inversions using the InvFEST dataset

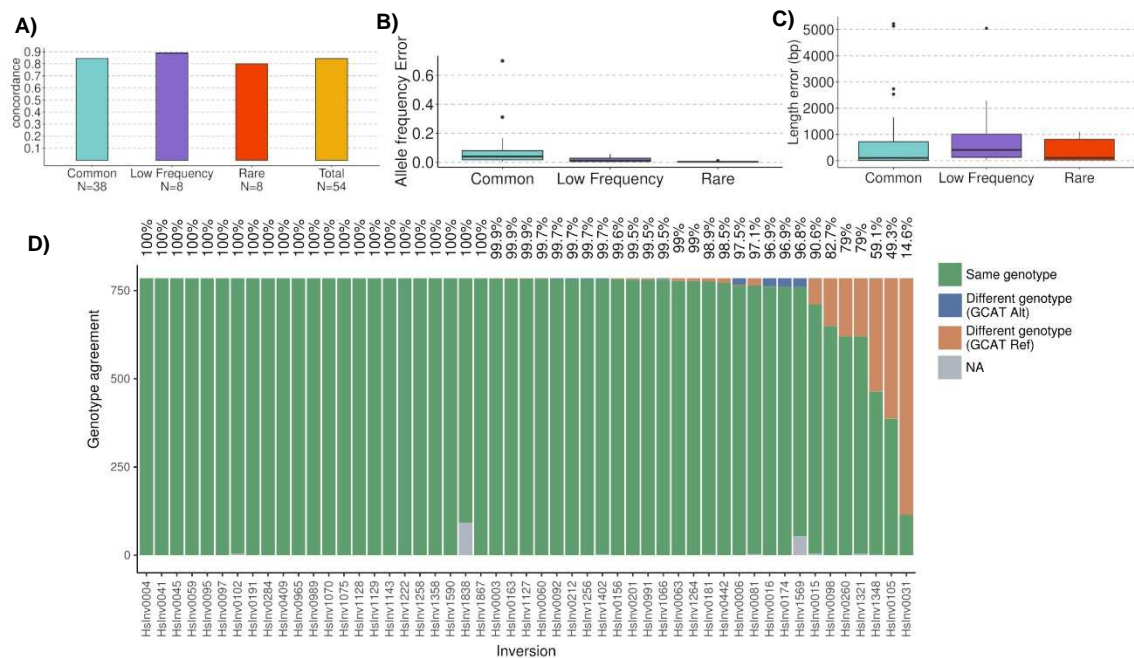


Figure 35. Inversion validation. A) Inversions shared between InvFEST and Iberian-GCAT catalogue. The non-homologous (NH) inversions from InvFEST were 64, classified into 45 common (MAF > 5%), 9 low frequency (1% > MAF > 5%) and 10 Rare (MAF < 1%) variants. We detected 84% of them, showing a high capacity to detect inversions by our variant calling. **B)** Allele frequency differences between the InvFEST and Iberian-GCAT catalogues. The allele frequencies between catalogues are similar in European populations, indicating that allele frequency estimations reported by our pipeline were precise. **C)** Length discordance between the InvFEST and Iberian catalogues. **D)** Genotype comparison between variant calling and imputation calls with InvFEST experimental panel (Lerga-Jaso et al., in preparation). Genotype agreement between the 51 NH inversions in common with the InvFEST dataset of validated inversions for the 785 sequenced GCAT individuals (with percentage indicated at the top). Individuals that have extra alleles with the same orientation as the reference genome in the GCAT calls are labelled as GCAT Ref (orange) whereas extra alternative orientation alleles in GCAT are indicated as GCAT Alt (blue). For the seven inversions that showed most discrepancies, the reference genome orientation was assigned to virtually all allele discordances, whereas InvFEST imputation calls the alternative. NA, individuals with unknown imputation call from InvFEST panel (not considered in the comparison).

The validation of inversions was performed using the InvFEST dataset, which experimentally validated inversions from 1000G. The coincident inversions between the GCAT and InvFEST dataset could provide an idea about the Iberian-GCAT catalogue's inversion accuracy (section 3.8.2.3). We evaluated the length and allele frequency concordance between the two projects, and the recall of our variant calling using the imputation results from 785 GCAT samples imputed using the InvFEST as a reference panel.

On average, **84% of Non-Homologous methods (NH) inversions in the InvFest project were detected by our variant calling**, and >80% in all allele frequency categories (Figure 35A). Besides, the **allele frequency and length reported by our calling was concordant** with InvFEST for the majority of inversions, demonstrating high precision in inversion characterisation (Figure 35B, C). Finally, considering the NH imputation results of InvFEST as a reference, genotypes in imputation and variant calling were **the same for 94.7% of the comparisons** (Figure 35D), with a 100% genotype agreement in all individuals for 23 inversions and less than 5% genotype discrepancy for 21 inversions. Only seven inversions gathered 91.4% of the genotype discordances, consisting of 99.8% of the cases of the reference genome in the GCAT samples (Figure 35D). This excess of missed inverted alleles leads to an underestimation of the inversion frequency, although these errors do not have a clear cause.

4.3 A Haplotype-resolved panel of the Iberian Cohort

Current Whole-Genome Sequencing (WGS) projects are unable to distinguish the parental origin of produced sequences, as both the homologous chromosomes are analysed simultaneously^{159,160,162}. For this reason, phasing process is required, consisting of resolving the phase of variant genotypes to produce haplotypes¹⁵⁹. Haplotypes are of paramount interest, to answer questions about human evolution, or improve GWAS resolution through imputation^{159,160,200}.

In this direction, different projects such as 1000G⁵, GoNL² and HRC¹⁵⁷ estimated haplotypes using different phasing algorithms, among which Shapelt¹⁸⁸ and Beagle¹⁶⁴ are the most recognised, to generate haplotype-resolved panels (reference panels), mainly used in GWAS. Current reference panels have managed to well characterise and phase SNPs and indels; however, they are still limited in Structural Variants (SVs) ($\geq 50\text{bp}$)^{5,161} due to the current sequencing library properties (ex: read length)³⁴. Only the 1000G, GoNL, and 1KJPN⁵¹ projects have included SVs in their sets, among these, 1000G is the unique publicly available set with SVs, even if lacking translocations and *de novo* insertions⁵. Additionally, resolving the haplotypes with SVs is a challenge; for example, 1000G and GoNL firstly generated a haplotype scaffold with SNPs and indels using Shapelt, then used MVNcall¹⁶⁹ to infer SV genotypes in the haplotype scaffold, which was conditioned to a flank set of phased SNP sites^{5,161,162}. Hence, the SVs were not appropriately phased, calling for the development of improved strategies.

SVs play a pivotal role in genetic diseases, and steady improvements in sequencing technologies and phasing algorithms have better characterised these variant types. For this reason, we generated an Iberian-GCAT reference panel with special emphasis on SVs, using the dataset obtained from the variant calling of GCAT samples (section 3.7). In this section, we evaluated the performance of different phasing strategies to generate a reference panel including SVs; even more, we evaluated and validated the imputation performance of the Iberian-GCAT reference panel, using samples from different ancestries. Finally, we carried out an imputation

benchmarking using different reference panels to show the advantages of using the Iberian-GCAT reference panel in GWAS studies.

4.3.1. Evaluating different phasing strategies to create the Iberian-GCAT reference panel

As mentioned in section 1.3.1, various phasing algorithms are available to obtain haplotypes from genotyping data. Unfortunately, no algorithm has been specifically evaluated to phase SVs. To address this limitation, we evaluated the SV phasing performance of Shapetl, using different versions and in combinations with other algorithms.

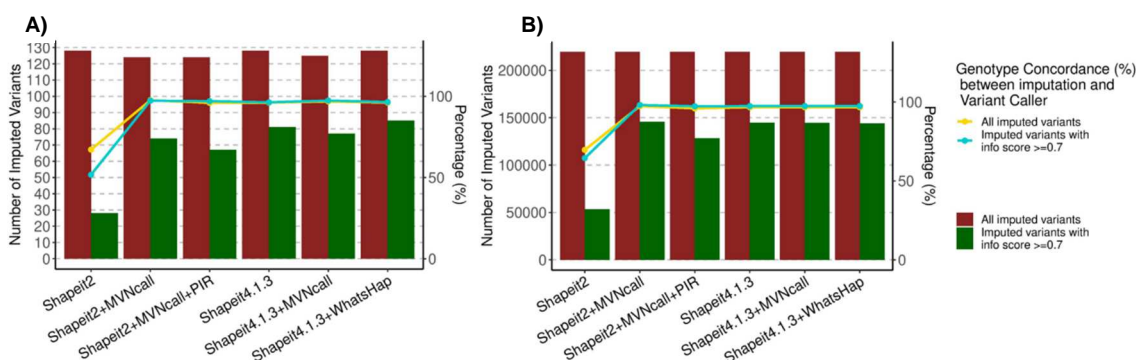


Figure 36. Phasing benchmarking. The left Y axis showed the number of imputed variants, the right Y axis showed the percentage of concordant genotypes between WGS and imputation. **A)** Large deletions (≥ 50 bp) on chromosome 22 imputed with an info score ≥ 0.7 , obtained from pilot reference panels built with different phasing strategies. **B)** SNVs and indels on chromosome 22 imputed with an info score ≥ 0.7 , obtained from pilot reference panels built with different phasing strategies. PIR: Phasing Informative Reads

We generated different pilot reference panels of chromosome 22, including SNVs, indels, and large deletions (≥ 50 bp), using the following phasing strategies: Shapeit2, Shapeit2+MVNcall, Shapeit2+PIRs+MVNcall, Shapeit4 (version 4.1.3), Shapeit4+MVNcall and Shapeit4+WhatsHap (section 3.9.1.2), to evaluate the best strategy to phase the Iberian-GCAT catalogue. Then, we imputed the GCAT SNP-array data of chromosome 22 using IMPUTE2. Finally, we evaluated the reference panel efficacy, in terms of the number of SNVs, indels, and large deletions imputed with a high quality (info score ≥ 0.7) (further details in section 3.9).

As shown in Figure 36A, **the phasing strategies affected the imputation of large SVs**, with Shapeit2 generally producing fewer high-quality SVs with info scores ≥ 0.7 . **The best imputation results were obtained using phasing informative reads (PIRs) with WhatsHap**, where 85 out of 128 SVs reached an info score ≥ 0.7 (Figure 36A), and increasing the imputation quality of rare variants (Supplementary Figure 4A). Finally, the genotypes reported by calling and imputation were concordant, determining that the phasing strategy does not influence the imputation genotype (Figure 36A). Also, 93% of all common SVs (59 common SVs in pilot reference panel on chromosome 22) were imputed with Shapeit4.1.3+WhatsHap. This percentage decreased with the allele frequency, being 80% for low-frequency variants (20 low-frequency SVs in pilot reference panel on chromosome 22) and 28.57% for rare variants (49 rare SVs in pilot reference panel on chromosome 22) (Supplementary Figure 6A). On the other hand, **the Shapeit2+MVNcall strategy improved SNV and indel imputation only 0.7% compared with Shapeit4.1.3+WhatsHap; the differences between both strategies were thus negligible** (Figure 36B).

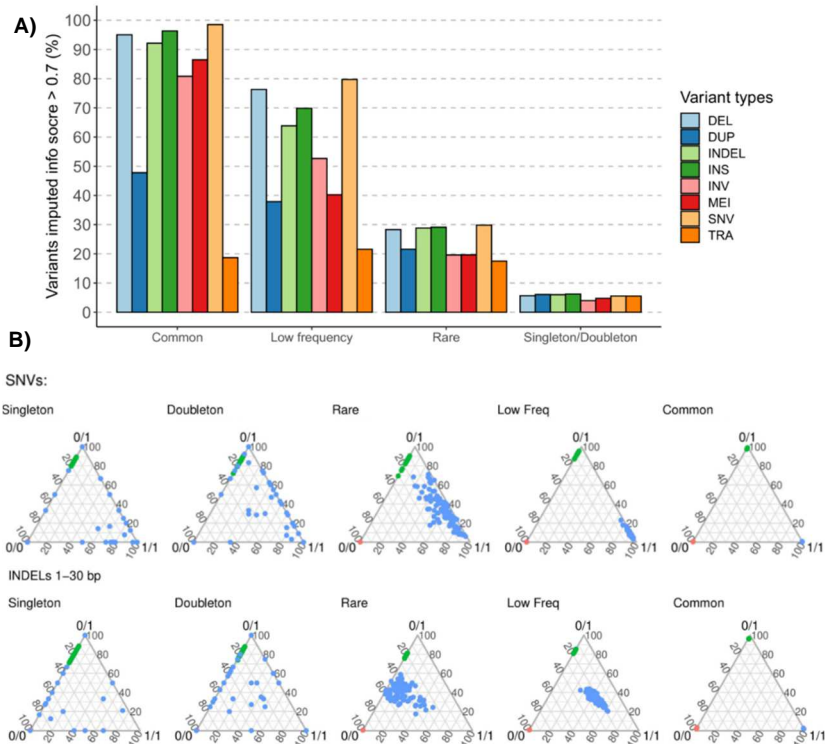
Furthermore, **Shapeit4.1.3 and Shapeit4.1.3+WhatsHap imputed more rare variants, for SVs as well as for SNVs and indels**, demonstrating that the new Shapeit4 version could include more rare variants in GWAS (Supplementary Figure 6B). These results suggested that the Shapeit4.1.3+WhatsHap strategy was the best combination to improve the imputation efficacy for SVs. We thus selected it to build the Iberian-GCAT reference panel.

4.3.2. Imputation performance using the Iberian-GCAT reference panel

To evaluate the Iberian-GCAT reference panel performance for imputation, we imputed SNP genotyping data obtained from both the GCAT (95 samples) and the 1000G project (1,880 samples), using IMPUTE2. We discarded all imputed variants with an info score < 0.7. The remaining variants were used to assess the imputation accuracy of the reference panel, including SVs (further details section 3.10).

4.3.2.1. Evaluation of imputation on the GCAT genotyping array

To assess the imputation accuracy of the Iberian-GCAT reference panel, especially for SVs, we selected a subset of 95 GCAT samples for which SNP-genotyping array and NGS data were available. We built a reference panel following the phasing strategy of Shapeit4+WhatsHap (including PIRs information) using the 690 remaining GCAT samples to avoid imputation bias. Then, with this panel, we imputed the SNP-genotyping data from the 95 GCAT samples. Finally, we determined the SV imputation performance and genotype concordance for each variant type, by comparing the imputation genotypes with those from NGS variant calling, considering as reference the variant calling genotypes (further details in section 3.10.1). Furthermore, we evaluated the effect to include A/T and C/G variants obtained from SNP-genotyping array and the effect of PIRs to generate the reference panel on imputation performance (see section 3.10.1).



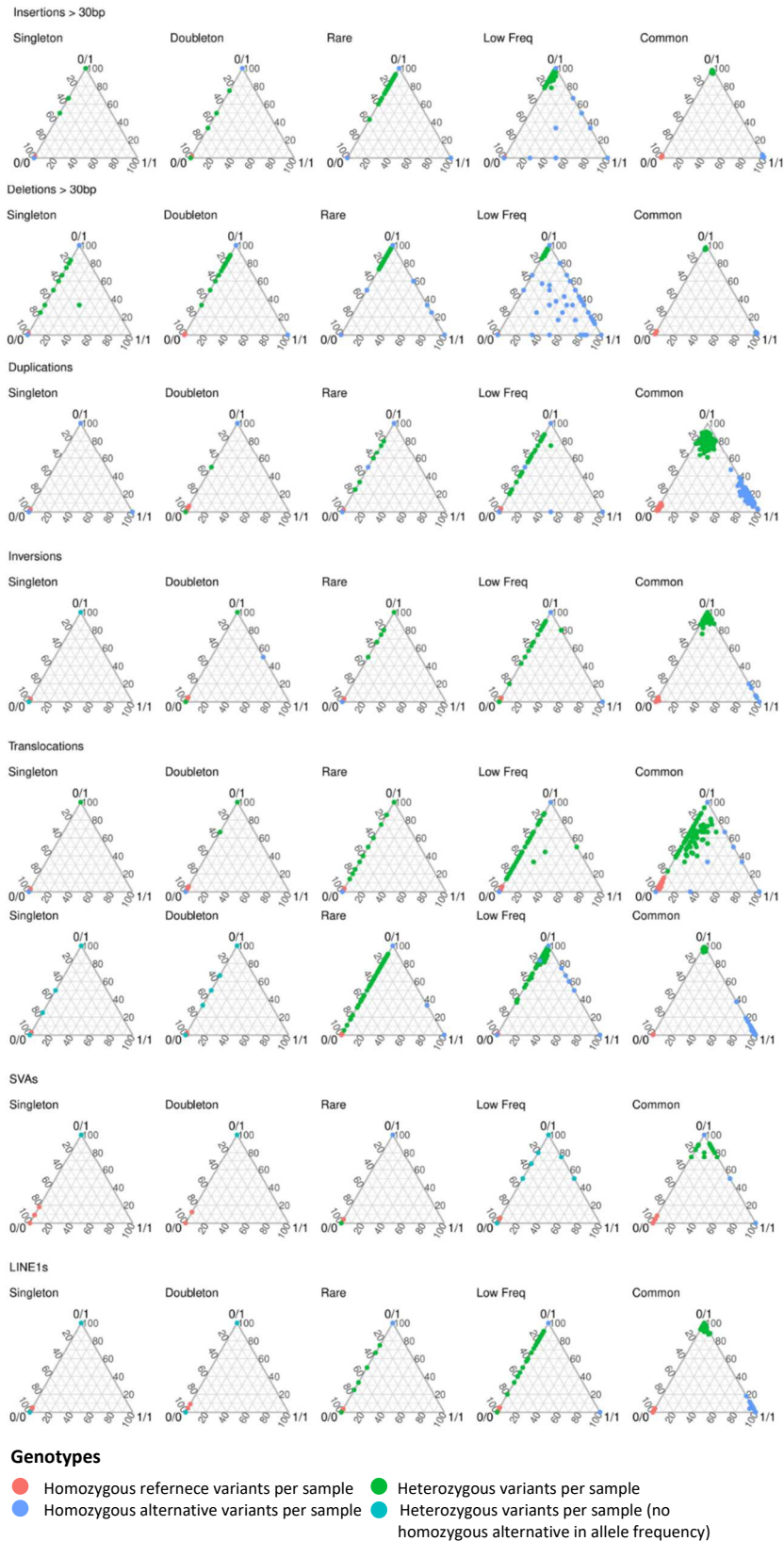


Figure 37. Imputation performance of the Iberian-GCAT panel for different variant types. A) Variants imputed with info score ≥ 0.7 grouped by frequency. **B)** Genotype concordance for 95 samples, grouped by variant type and allele frequency. Each dot is the percentage of genotype concordance between genotype imputation and genotype reported by Iberian-GCAT dataset, considering all variants per sample. The genotype concordance was calculated for each genotype state independently. When a dot is in a vertex, it means high genotype concordance.

Increasing the number of SNVs and indels in the SNP-array data is relevant to improve the SV imputation; however, using PIRs to create a reference panel did not improved SV imputation significantly, in contrast with SNVs, where the quality improved slightly (Supplementary Figure 8). Imputation performance strongly depended on the variant frequency, with **common variants (MAF \geq 0.05) showing the best imputation results.** Besides, only ~5% of all singletons and doubletons were imputed with high info scores (Figure 37A). On the other hand, duplications and translocations were the SV types with lower imputation (Figure 37A).

Similarly, genotype concordance between variant calling and imputation indicated **high concordance for common variants, being ~100% in all variant types,** with the exception of duplications and translocations, for which dropped to ~80% and ~60%, respectively (Figure 37B). Again, **genotype concordance decreased with the allele frequency.** For low-frequency variants ($0.01 \leq \text{MAF} < 0.05$), **the homozygous alternative genotypes showed a concordance of ~40-60%,** with the exception of SNVs, where concordance was ~90% for both heterozygous and homozygous alternative genotypes. Finally, for rare variants ($\text{MAF} < 0.01$), concordance for heterozygous genotypes **was ~60-86%, and dropped to ~35%** for homozygous alternative genotypes (Figure 37B). These results suggest a correlation between the haplotype frequency of homozygous alternative variants and imputation accuracy, where the homozygous alternative variants were less frequent than heterozygous variants, as well as happens with rare variants. Taken together, these results indicate that **imputation would allow enriching GWAS with common SVs at high quality;** however, imputation for duplications and translocations, as well as for homozygous alternative genotypes at $\text{MAFs} < 0.05$, is less accurate, and would require larger samples sizes for building reference panels.

Then, we assessed the impact of SVs in phasing algorithms, in order to determine if the inclusion of SVs affected the subsequent imputation quality of SNVs and indels (methodology in section 3.10.1). Besides, we evaluated the relationship between SNVs in high linkage disequilibrium (LD) ($r^2 \geq 0.8$) with SVs and the imputation quality of SVs, in order to decipher the reasons of the lower genotype concordance in common duplications and (further details in section 3.10.1).

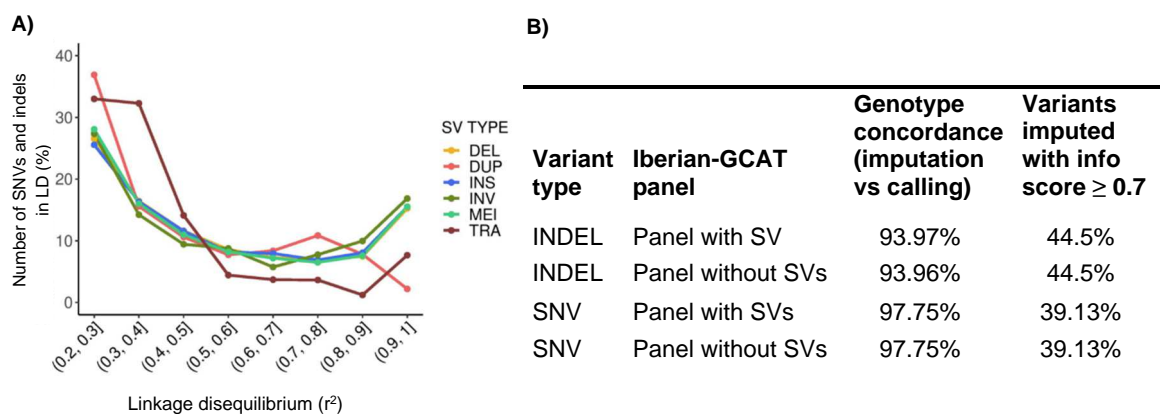


Figure 38. Impact of structural variants on imputation. A) Proportion of SNVs and indels in LD with common SVs. **B)** SV effect on SNV and indel imputation quality.

The imputation quality of SNVs and indels was not influenced by the inclusion of SVs in the Iberian-GCAT reference panel (Figure 38B), neither in terms of genotype concordance. Additionally, the info score of SNVs and indels around SVs was not decreased (Supplementary Figure 9). On the other hand, duplications and translocations were the SV types with the lowest

proportion of SNVs and indels in $r^2 \geq 0.9$ (Figure 38A), demonstrating that these low proportions could be the reason of the lower imputation performance, difficulting their imputation using this SNP-genotyping array data. Besides, for duplications, the imputation genotypes of common variants were grouped in their respective allele states with an error in genotype concordance of 20% (Figure 37), showing that **at least ~10% of SNVs and indels with $r^2 \geq 0.7$ were necessary to impute SVs accurately** (Figure 38A). However, we needed **~15% of SNVs and indels with $r^2 \geq 0.9$** to decrease the imputation errors below 1% (Figure 38A).

4.3.2.2. Evaluation of imputation on the 1000G genotyping array

The Iberian-GCAT reference panel was built with 785 GCAT samples and following the Shapeit4+WhatsHap strategy (section 3.9.2), this resource was used to evaluate the imputation performance across different populations. We used the SNP-genotype array data from 1000G, which includes 1,880 samples from 19 populations worldwide (full list of populations in Supplementary Table 7) (section 3.10.2.1). We imputed each population individually; then, we used all SVs with an info score ≥ 0.7 for downstream analyses (further details in section 3.10.2).

We assessed the precision and recall of SV imputation, by comparing imputation results with validated SV sets from 1000G samples, obtained from Audano et al.⁶⁷ (nine samples), and from Hickey et al.¹²¹ (three samples) (section 3.10.2.2). We evaluated length, position, and SV type errors of imputed SV using the Audano et al. dataset (section 3.10.2.2). Also, we assessed genotype concordance using Hickey dataset (section 3.10.2.2).

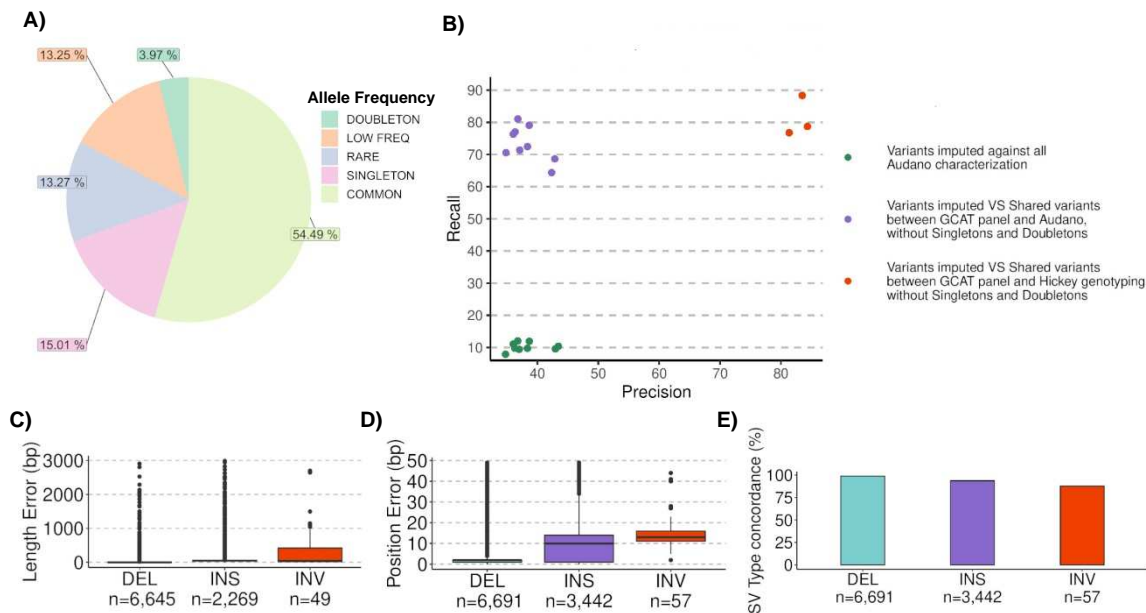


Figure 39. Imputation performance of the Iberian-GCAT reference panel. A) Allele frequency distribution of the 16,704 SVs shared between the Audano et al and GCAT SV dataset. **B)** Imputation accuracy results using Iberian-GCAT reference panel for SVs, compared to third-generation sequencing (TGS; Audano et al.) and genotyping (Hickey et al.). **C)** Length discrepancies between the GCAT SV dataset and Audano et al. **D)** Breakpoint resolution discrepancies between the GCAT SV dataset and Audano et al. **E)** Structural Variant type concordance between the GCAT dataset and Audano et al.

Audano et al. characterised 93,852 SVs for nine samples, classified into deletions, insertions, and inversions. **17.8% of them were shared with the GCAT SV dataset, the majority of which were common (54.5%), followed by singletons+doubletons (18.9%), rare (13.3%), and low-frequency (13.3%)** (Figure 39A). First, we evaluated the imputation performance

considering the whole Audano et al. dataset, which showed **low precision (~35%) and recall (~10%)**. Thus, we discarded singletons and doubletons, due to their known low imputation accuracy, and compared once again the variants shared between both projects; **recall showed a substantial increase (64-81%)** (Figure 39B). Three of the nine samples analysed by Audano et al. were genotyped using short reads by Hickey et al. ¹²¹. A comparison with this second dataset showed better accuracy for imputed SVs, with a **precision of ~80% and a recall between ~79-89%** (Figure 39B). These results drive to relevant conclusions, such as that **imputation allows detecting about ~10% of whole SVs of non-Iberian samples**, using an Iberian-GCAT reference panel. **SV detection with long-reads followed by genotyping with short-reads could improve the SV detection, increasing to ~80% the precision results**. Additionally, breakpoint resolution, length, and variant type reported by the Iberian SV catalogue and Audano et al. was consistent in all categories (Figure 39C, D, E).

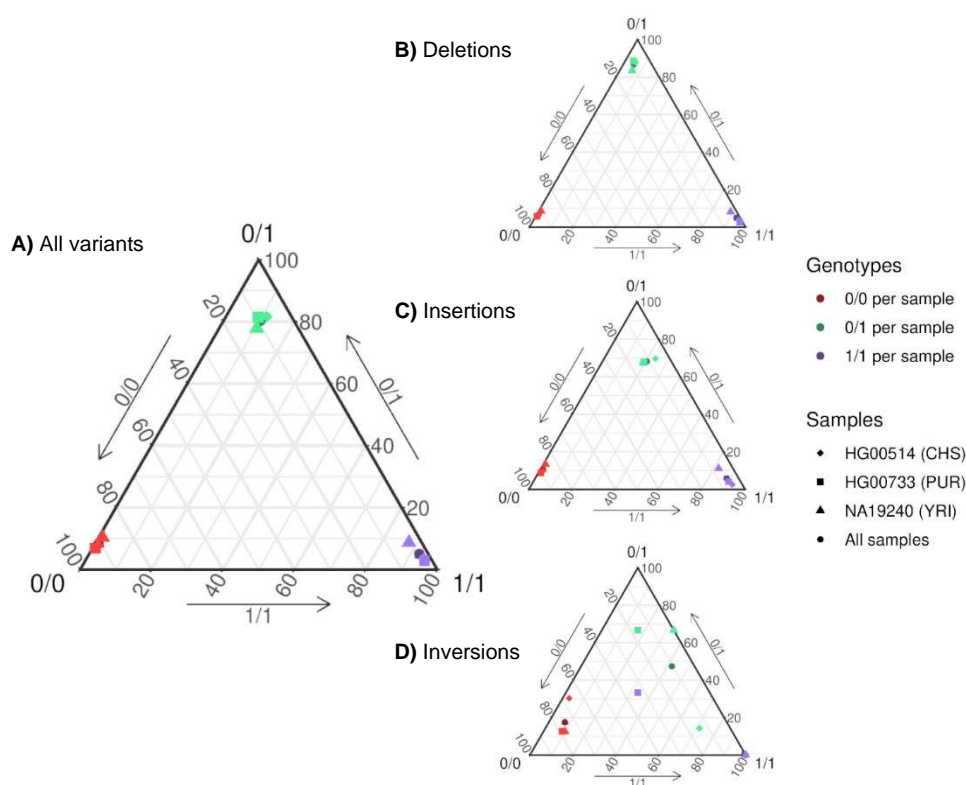


Figure 40. Genotype concordance of structural variants between imputation and Hickey et al genotyping. Genotype concordance was calculated using all variants for each sample, obtaining proportion of concordance between SV imputation results and Hickey et al. **A)** Genotype concordance of all SVs imputed in three samples from different populations. **B)** Genotype concordance of Deletions (n=17,205). **C)** Genotype concordance of Insertions (n=8,903). **D)** Genotype concordance of Inversions (n=141).

Finally, we evaluated the SV genotype concordance using three 1000G samples, imputed with the Iberian-GCAT reference panel and genotyped by Hickey et al. **Genotype concordance across all samples was ~80%** (Figure 40A), indicating that the Iberian-GCAT reference panel could be used in different populations, reporting accurate genotypes of imputed SVs. More in detail, **deletions were the SVs imputed with highest concordance (~90%)** (Figure 40B). The insertion genotypes obtained from imputation with the Iberian-GCAT reference panel, comprising TRAs, DUPs, TRPs and INs, **were ~70% concordant in heterozygous and ~95% in homozygous** with those reported by Hickey et al. (Figure 40C), showing a good

genotype concordance. Unfortunately, only a small number of inversions (n=141) were shared between the two projects (Figure 40D), hampering the possibility of obtaining deeper insights into this poorer concordance. Additionally, **the sample with lowest SV genotype concordance was of African (YRI)**, suggesting that the Iberian-GCAT reference panel includes fewer variants with high LD in African populations (Figure 40).

4.3.2.3. Structural variant worldwide distribution

Imputation quality is relevant to determine if a reference panel can be used to impute a cohort originally genotyped with SNP-genotyping array data. For this reason, we evaluated the imputation quality reported by IMPUTE2 in different populations, using the Iberian-GCAT reference panel (section 3.12.1). Additionally, we determined the SV distribution across all populations of 1000G (further details in section 3.12.1), in terms of the allele frequency distribution and, SVs sharing between populations, to obtain an overview of SV distribution worldwide.

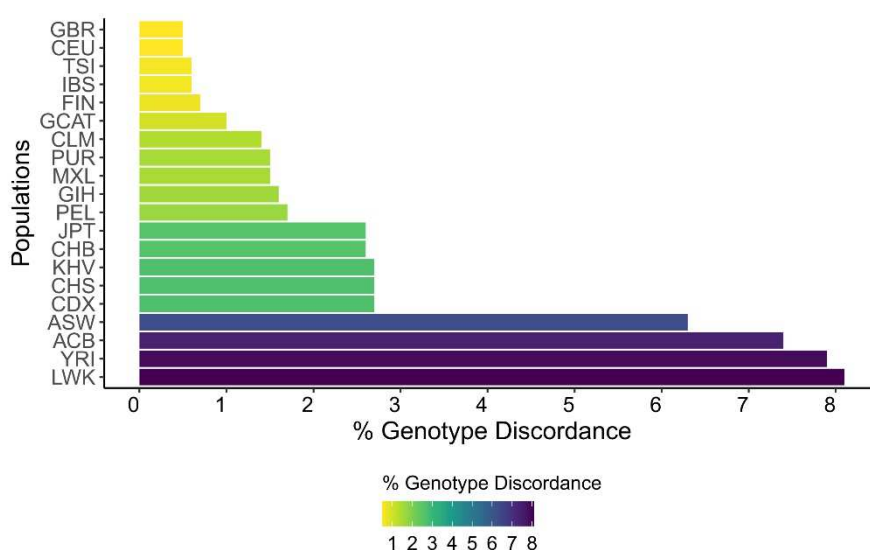


Figure 41. Imputation quality grouped by populations. The imputation quality was determined comparing the input SNP-genotype array with with the best-guess imputed genotypes and reporting the concordance. IMPUTE2 reported this metric in summary table (further details in section 3.12.1).

Imputation quality was overall high in all populations, suggesting **that the Iberian-GCAT reference panel could be used in all populations for imputation analysis**, even in African populations, where the imputation quality is lower (Figure 41). As expected, **imputation quality was strongly correlated with the genetic similarity between populations and Iberians**. Indeed, European ancestries showed a genotype discordance $\leq 1\%$ in European ancestries, followed by Latin American ancestries (and Indian population (GIH)) ($< 2\%$ discordance), Asian ancestries ($< 3\%$ discordance) and African ancestries, which genotype discordance increased to $> 6\%$ (Figure 41). Surprisingly, the highest genotype discordance among European populations was observed in the GCAT cohort. This result could be explained by the fact that, for the GCAT cohort imputation, we built a reference panel with 690 samples only (section 3.10.1), while for all the other populations, the Iberian-GCAT reference panel included 785 samples (section 3.9.2). However, genotype discordance was only 1%, demonstrating high imputation quality.

After imputation in each population, we recovered a total of **49,724 SVs** with an info score ≥ 0.7 . Below, we describe the SV distribution in each population.

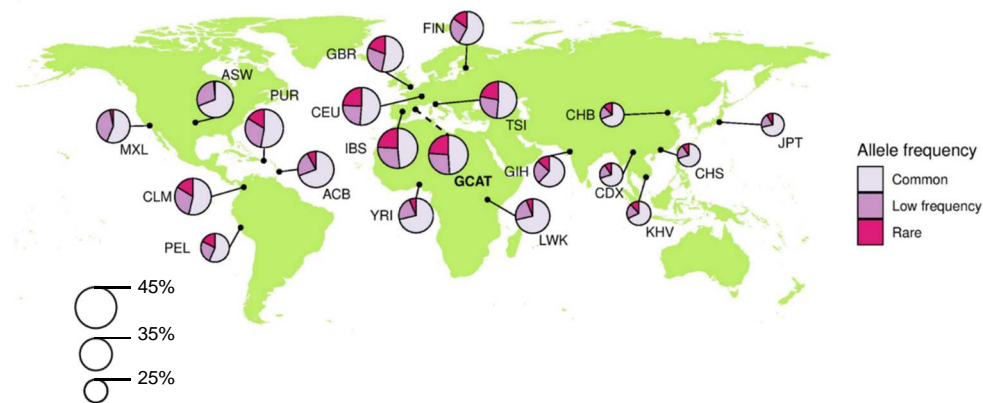


Figure 42. Structural variant distribution by population and allele frequency.

For the Asian populations (with the exception of the Indian ones (GIH)), we were able to recover fewer SVs than other continents (~25%). Between **35 and 40% of all the imputed SVs were instead present in the African and Latin American populations** (Figure 42). **The European populations carried (with the exception of Finnish (FIN)) > 40% of the imputed SVs** (Figure 42). Additionally, **the majority of imputed variants were common** (MAF \geq 5%), with ~70% in Asian and African populations, and ~50% in Europeans and Latin Americans. Also, in both Latin American and European populations more low-frequency and rare variants with info scores \geq 0.7 were imputed, showing the Iberian-GCAT reference panel efficacy to recover more low-frequency variants in closer ancestries (Figure 42). Finally, **deletions were the most imputed SVs across all populations (~50%), followed by insertions (~22%) and ALUs (~17%)**, where the other SVs represented ~10% of all imputed SVs per population (Supplementary Figure 7). These results indicate that imputation performance is correlated with ancestries.

The SV distribution shows different patterns of recurrence. For example, **30% of all SVs imputed were shared across all continents** (Figure 43A), **with 5,055 common SVs** (Figure 43B), **demonstrating that SVs could have an inheritance component** (Figure 43A). Besides, 21% of all imputed SVs were representative in European populations, suggesting that **the Iberian-GCAT reference panel recovered more SVs with high quality in Europeans** (Figure 43A). On the other hand, 14,013 SVs were imputed in a single population; **these private SVs were predominantly rare** (Figure 43C).

Summarizing, the Iberian-GCAT reference panel could be used for imputation analysis in all the analysed populations due to the high-quality imputation scores. The ancestry component is relevant to improve the imputation of SVs, with the Asian populations the community with lower info scores, in contrast to Europeans and Latin Americans. The SVs could have an inherited nature, where 30% of all SVs are shared across all populations, showing the importance of including the Iberian-GCAT reference panel for GWAS to find new disease-variants associations.

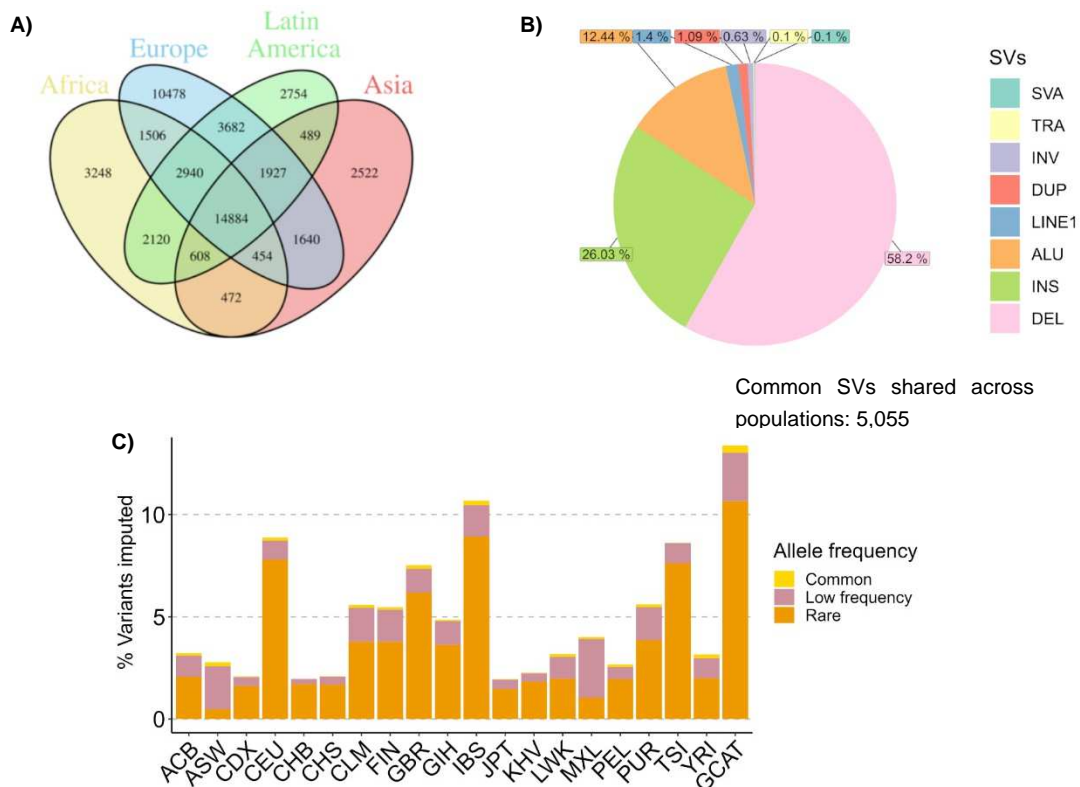


Figure 43. Structural variants (SVs) particularities across all populations. A) Structural variants shared across all continents. We selected SVs with an info score ≥ 0.7 . **B)** Common SVs (MAF $\geq 5\%$) shared across all populations grouped by SV type. **C)** Structural variants imputed in a single population divided by allele frequency.

4.3.3. Comparing imputation performance of multiple reference panels

Currently, different reference panels were generated from WGS data (Table 1). In this context, we compared the imputation performance of the Iberian-GCAT reference panel, with the most popular reference panels. However, not all reference panels included SVs, for example, GoNL has two versions, one including SV (GoNL-SV) and discarding indels, and another including only SNVs and indels (GoNL), HRC only includes SNVs or 1000G phase 3 covering SNVs to SVs. To compare the imputation performance between reference panels, we used GUIDANCE, a tool developed in our group that is able to impute with multiple reference panels in a single execution (further details in section 3.11). We imputed SNP-genotyping array data from 4,448 GCAT samples using five reference panels, and we then selected, for each panel, variants with and info score > 0.7 and MAF > 0.001 (further details in section 3.11).

The imputation performance for SNVs and indels were similar between 1000G and Iberian-GCAT panels, with the Iberian-GCAT reference panel recovering more indels, and 1000G more SNVs (Figure 44A). HRC and 1000G recovered more rare SNVs than GCAT due to their larger sample size (Table 1). However, **the majority of rare indels were recovered by GCAT** (Figure 44A). Population-specific panels such as GoNL and UK10K were able to impute fewer SNVs and indels than Iberian-GCAT panel, suggesting that the ancestry component of GCAT SNP-genotyping array facilitated to recover more variants in the Iberian-GCAT reference panel. Finally, **combining all the reference panels allowed us to obtain more SNVs and**

indels than each of the panel individually (Figure 44A), demonstrating that even smaller panels can contribute to increasing the imputation performance, and this justifies the building of new and population-specific panels. Besides, **increasing the haplotype number is crucial to improve the imputation, specifically in rare variants**, as shown in Figure 44C, where HRC imputed high-quality SNVs and indels.

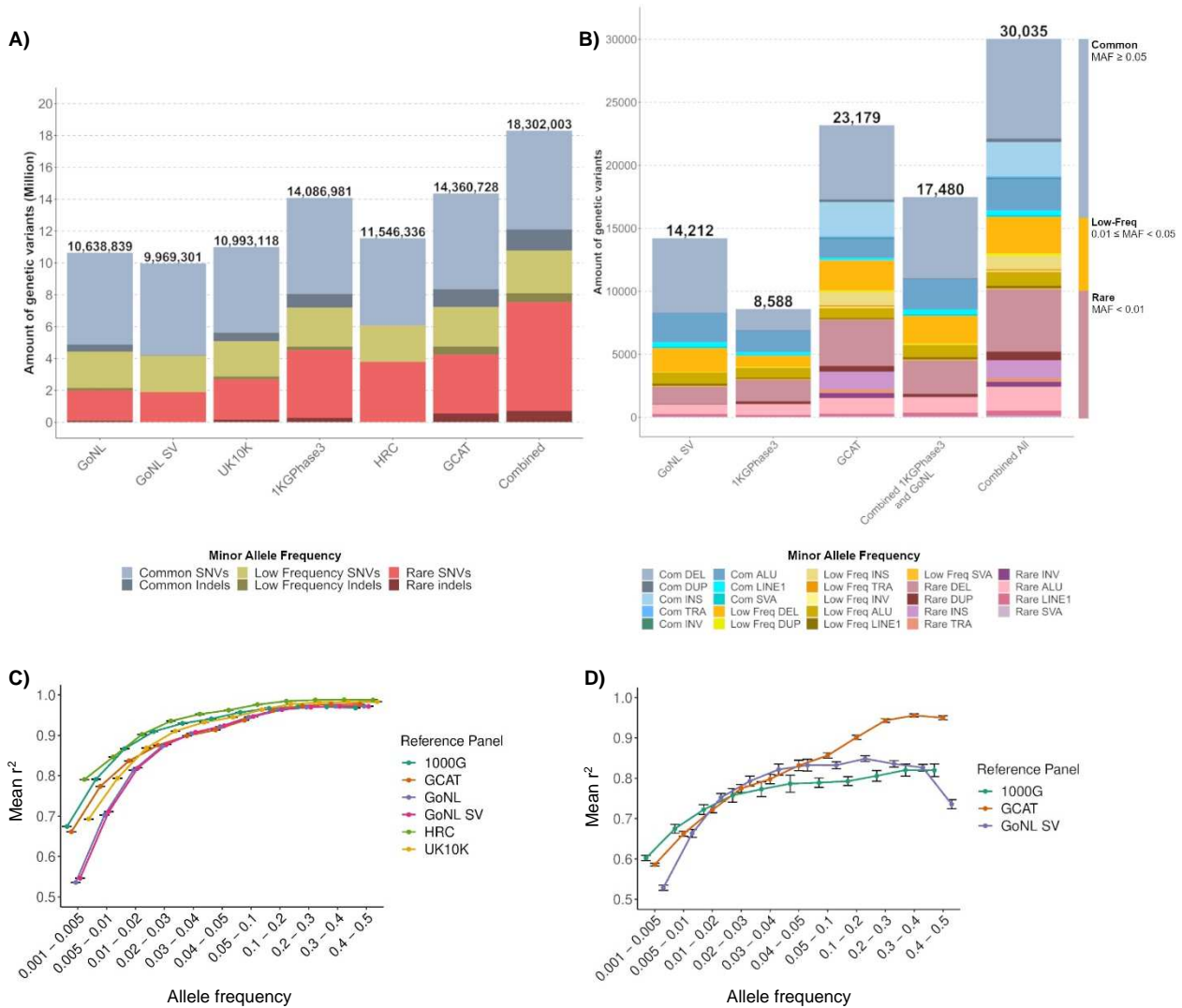


Figure 44. Imputation performance using different reference panels. A) SNVs and indels imputed using reference panels, covering different allele frequencies (info score ≥ 0.7). Combining all reference panels increased the number of imputed variants well-imputed. Particularly, Iberian-GCAT panel imputed more indels than others, mainly due to high coverage used, as mentioned Byrsika-Bishop et al.²⁰⁶. **B)** Structural variants recovered using different reference panels, considering all SV types and allele frequencies (info score ≥ 0.7). **C)** Imputation accuracy of SNVs and indels at different allele frequencies. HRC obtained the best results due to the panel sample size, allowing to impute also rare variants with high quality. **D)** Imputation accuracy of Structural variants for different allele frequencies. The imputation accuracy among different panels were slightly different. However, the Iberian-GCAT panel showed a higher imputation quality for common SVs in comparison to other panels.

On the other hand, the **Iberian-GCAT was the reference panel that recovered more SVs, 2.7 and 1.6 times more than 1000G and GoNL-SVs, respectively. Besides, the GCAT-reference panel overcome by 1.3 times the SV recovery, compared with the combination of those two panels** (Figure 44B). However, the introduction of the **Iberian-GCAT reference panel ameliorated SV recovery nearly of 50% compared to the combination of 1000G and GoNL alone** (Figure 44B). Additionally, the imputation quality between panels was similar, indicating **the necessity to include more haplotypes to raise the quality of rare SVs** (Figure

44D). The **Iberian-GCAT panel improved the imputation accuracy of common SVs** (Figure 44D), paving the way for their inclusion in GWAS. However, these results could be upwardly biased due to the same ancestry between SNP-genotyping array and Iberian-GCAT haplotype reference panel, where the similarity of haplotypes between array and Iberian-GCAT reference panel facilitated the imputation.

Briefly, the Iberian-GCAT reference panel is a great resource to impute variants, especially SVs ($\geq 50\text{bp}$). The imputation allowed to recover of more SVs at high quality, so including this resource in GWAS will improve their statistical power and resolution, allowing them to discover new disease-association variants, and deciphered the SV effect on human diseases.

5. DISCUSSION

Panels of genetic variability (reference panels) are widely used in Genome-Wide Association Studies (GWAS), because they can help to improve their resolution and statistical power, by enabling imputation approaches. Current reference panels cover SNVs and indels, but they are limited in Structural Variants (SVs) (≥ 50 bp). The main challenges in SV discovery are linked to library properties, such as coverage, insert size and read length, and SV particularities, hampering the discoveries by the variant callers. In this context, this thesis focused on building the first haplotype reference panel of the Iberian population, including an accurate characterisation of SVs, which is still lacking in our knowledge. Besides, generating population-specific reference panels opens new opportunities to find variants related to diseases in specific ethnic groups, improving the imputation of rare and low-frequency variants, of paramount interest in precision medicine, as to optimise the diagnosis, prevention and treatment of complex diseases.

In this section, results are discussed in three main blocks. 1) Variant caller benchmarking, covering from the strategy to generate the BAM files to the comparison of the variant caller and Logistic Regression Model accuracy. 2) The genetic characterisation of Iberian-GCAT catalogue, describing the recovered variants and their distribution within the cohort, and the functional implications of SVs. Finally, 3) the Iberian-GCAT Haplotype reference panel performance, especially discussing the strategy to create the reference panel, and the results obtained from imputation.

5.1 Variant caller benchmarking

Currently, more than 150 variant callers⁷⁵ are used to detect the genome variability from WGS data. However, no single algorithm is able to detect the whole spectrum of variants accurately⁹, increasing the chances to produce false-positives in variant detection. Thus, evaluate the strengths and weaknesses of variant callers is necessary to perform a variant discovery precisely. The precision, recall and F-score are metrics to evaluate the variant callers performance. These metrics are determined by several factors, such as library properties (read length, coverage or insert size)^{9,46}. For this reason, before applying variant calling in real samples, we used golden samples such as NA12787 (only for SNVs and indels) and an *in-silico* sample, to evaluate the best pipeline to perform accurate variant and genotype calling.

Different approaches were designed to construct the aligned BAM files, for instance, applying the GATK best practices and using the hs37d5 or hg19 reference genome (Figure 19 and Figure 20). The results of testing these approaches demonstrated that **building the BAM files following the GATK best practices, and aligning the reads using the hs37d5 reference, improved SNVs detection accuracy**. However, we did not appreciate relevant differences between SVs, only low recalls for inversion discoveries by Delly2 using only hs37d5 reference genome, and a better precision in translocations was appreciated using hs37d5 reference genome and GATK best practices in all variant callers. Several hypotheses could explain these results. One was the high coverage (30X) of these golden samples, since more signal could be used to detect the variants, facilitating the discoveries. Another was that the most recent variant callers updates, enabled a more precise depuration of false-positive. In summary, although the differences in variant detection were not relevant, we constructed the BAM files applying GATK best practices recommendations and using the hs37d5 reference genome, to obtain BAM files of a better quality.

Besides, breakpoint resolution was influenced by variant length, and by the strategies used by variant callers. For example, the GoNL project is a unique resource that distinct indels by their

sizes, determining as indels all variants at sizes ≤ 20 bp¹⁶¹, because the resolution of breakpoint was at base-pair resolution, such as SNVs, in contrast to other projects (i.e. 1000G) where trivially indels are < 50 bp. Besides, GoNL project, for variants of a length that could be within the read length (100 bp), only deletions were classified as mid deletions, because their breakpoint resolution was higher than other SVs. Following this criteria, we found similar results than the GoNL project (Figure 15), where precision was around 90% until indel sizes of 20 bp. However, the read length of the *in-silico* sample (100 bp) was smaller than that of the GCAT samples (150 bp), concluding that breakpoint resolution and variant characterisation was highly accurate if the variant size did not overcome 20% of read length, allowing us to increase the definition of indels to 30 bp. These results allowed grouping the variants by length, as follows: SNVs, indels (1-30 bp), mid-deletions (31-150 bp) and SVs (>30 bp) (in exception of deletions > 150 bp). This classification enabled the design of specific filtering and merging strategies as well as benchmarking analyses, considering each variant type independently.

5.1.1. Benchmarking of SNVs and indels

NGS technologies allowed us to detect SNVs and indels accurately. Currently, SNV detection can be done precisely (99%) with a coverage of $\sim 15\times$ ⁸³. Different benchmarking studies used the NA12787 sample as a reference set, determining that, for SNVs, the Haplotype caller, Strelka2 and Deepvariant had a $>99\%$ precision and $>96\%$ recall^{128,129}. In our hands, the results were lower using the same sample, with a precision of $\sim 96\%$ and recall between 82-95% (Table 14). On the other hand, using the *in-silico* sample, the values were similar to the studies. These discrepancies in precision and recall between both samples could be produced by the filters applied. In some cases, the reference allele from the GIAB sample and variant callers did not match, decreasing the metrics slightly. For this reason, better comparisons had to be performed, to improve the variant filtering. Overall, these results showed **high accuracy of SNV detection across the majority of variant callers. Thus, there was no need to generate a Logistic Regression Model (LRM) to filter potential false-positives** since it would lead to similar values (Table 14). For example, the LRM2 only predict as true-positive the variants detected by Deepvariant, showing inconsistencies in these predictions (Table 14). For this reason, in order to clean-up the potential false-positive detections from the GCAT samples, and to decrease the bias obtained from anyone variant caller, we decided to filter out all SNVs detected only by one caller, because the recall and precision metrics indicated that all callers had to detect the same SNVs.

On the other hand, indel detection required higher coverage to achieve an accurate detection. For this reason, Deepvariant, Haplotype caller and Strelka2 had lower values in comparison to SNVs, with precisions around 93%-96% and recalls between 85%-92%^{128,129}. Our analyses showed precisions of $\sim 96\%$ and recalls of $\sim 88\%$ using the same tools and sample (Table 15). However, the precision decreased to 88-92% and recall to 82-84%, when using the *in-silico* sample (Table 15). These results demonstrated that the **NA12787 GIAB sample could overfit the metrics. This could likely be due to an abundance of variants located in conservative genome regions**, which were easy to detect in comparison to repetitive or polymorphic genome regions¹²⁰. Besides, **the LRM2 created using the Deepvaraint, Haplotype Caller and Strelka2 outputs, improved the accuracy of indel detection slightly** (Table 15). Thus, we used the LRM2 to filter potential false-positives for indel detection.

In conclusion, **the LRM was not effective to improve variant detections when variant callers were already highly precise, as demonstrated in SNV detection.** However, to reduce variant caller bias and potential false-positives, combining different variant callers could still improve variant detection. Besides, knowing the reference sample features, a better benchmarking could be designed. For example, using the NA12787, we only evaluated the conservative genome regions, limiting the insights of variant caller performance in low complexity and polymorphic regions. Considering this information and until no real sample will be fully characterised, we could perform other approaches to benchmark polymorphic or low complex regions, by building an *in-silico* sample, where we could evaluate the performance of variant callers in those regions.

5.1.2. Benchmarking mid Deletions and large Structural Variants

The detection of SVs using NGS technologies is a challenge, due to library and variant particularities. Thus, no single variant caller can efficiently detect all SV types and lengths, producing errors in SV identification.

In this context, several strategies have been applied by variant callers to detect SVs, such as Split-read (SR), Discordant read (DR), *de novo* assembly (AS), or Read Depth strategies (RD) (Figure 4). These strategies have their strengths and weaknesses, SR and AS with higher accuracy in breakpoint identification and DR in SV length identification. RD strategies are useful to detect large deletions and duplications, in decrement to worst breakpoint and length resolution⁷⁶. However, recent variant caller updates combined those signals improving SV detection, breakpoint and lengths resolution^{18,76,77}. **The Variant callers used in this project combined different strategies (Table 13), showing high breakpoint accuracy**, allowing an error of ± 10 bp, with the exception of CNVnator, which used the RD strategy and had a breakpoint error > 100 bp (Table 16) (Supplementary Figure 1). Besides, large deletions > 150 bp and translocations had worse breakpoint resolution, showing that while the size and complexity of a variant increased, its breakpoint accuracy decreased (Table 16).

This information was used by the LRM, which converged to selecting the position of the most accurate variant caller for each SV type (Table 6). This strategy produced better results than simply using the median of all variant callers (Supplementary Figure 1). On the other hand, the length reported by the LRM was the median between all variant callers, due to the consistency of their results (Supplementary Figure 2). **Using this approach, we expected to correct the position and length bias due to short reads misalignments, obtaining an accurate SV catalogue.**

SV detection by NGS technology produces high FDR (9-89%) and low recall (10-70%), depending on the size and SV type⁸. Better SV discoveries can be performed using TGS technologies, with increased read lengths to 10-20 Kb. However, due to their costs and high sequencing error rate (8-20%)⁸ their applicability in population studies remains a challenge, leaving NGS as the unique realistic technology to discover SVs. **For this reason, in order to increase the precision and recall of SVs, one possible approach consists in combining different variant callers, preferably those which use different detection methods^{9,95}.** In this direction, two main approaches were designed to combine variant caller detections. One combined the SV calls detected by at least two variant callers (logical rules), which increased the chances of merging two “bad” pairs of algorithms, giving a small precision increase, but reducing

the recall⁹. This approach was used by the GoNL project, which filtered all SV detected by one variant caller or one strategy. A second approach consisted in creating machine learning methods, giving different discriminative power to all variables introduced in those algorithms, increasing the strengths of each variant caller to filter the potential false-positives¹³². For example, 1000G generated a Support Vector Machine (SVM) method to depurate the calls with questionable quality.

In this direction, we performed an exhaustive variant caller benchmarking for each SV type and contrasted the results against a Logistic Regression Model (LRM) trained to detect each SV type. Figure 17 **showed that the LRM improved SV discovery, with an F-score of 0.9, recall of 0.85 and precision of 0.95**. However, not all variant callers detected all SV types and sizes equally, which was a feature considered in the LRM decision process. In some SV types, the metrics differences between LRMs and logical rules or variant callers individually were no higher than expected. For example, for duplications, the LRM and the ≥ 2 callers strategy performed accuracy similarly (Supplementary Figure 3E). In *de novo* insertions, the LRMs outperformed all logical rules and variant callers (Supplementary Figure 3C), showing the relevance of using machine learning algorithms for SV filtering. This approach allowed to filter the potential false-positive detections in real samples for each variant type, providing an avenue to obtain an accurate catalogue of SVs in the Iberian population.

Besides, a depurated catalogue of variants is not the unique barrier to build a resolutive haplotype reference panel. An accurate genotype is also crucial, in order to find more variants in linkage disequilibrium, and thus, enable the imputation of more variants with high quality¹¹⁶. The **genotyping strategy of the LRM had a genotype error of 5.6%**, overcoming the genotype performance of all variant callers individually (Figure 18), demonstrating that combining the genotypes from different outputs is the best strategy to reduce the errors. For deletions, insertions and inversions, **the genotype concordance was highly accurate (92%), probably due to the high coverage (30X) of the *in-silico* sample** (Supplementary Figure 4).

However, for duplications and translocations, the genotypes obtained by different variant callers were highly discordant of those of the *in-silico* sample. For example, in duplications, Manta and Pindel reported only heterozygous variants, with $> 99\%$ errors in homozygous (Supplementary Figure 4E). These discrepancies drove us to perform custom genotyping, using the BAM file (section 3.4.2.3). **Our strategy outperformed the genotype concordance for duplications and translocations**, with 20% of genotype error for duplications and 5.16% for translocations (Supplementary Figure 4E, F). This disparity indicated that variant callers used different assumptions to determine the genotype of variants, and maybe should reconsider specific genotyping strategies for each SV type, once the variant is discovered. A more promising strategy for future variant callers, would thus consider the SV particularities for genotyping, instead to genotype all SV types following the same strategy. For example, the coverage for duplications is at least two times higher than other genome regions, and this amount of signal could produce the homozygous alternative alleles bias in Manta or Pindel.

Finally, as previously mentioned, not all variant callers could detect all SV sizes accurately⁹. For this reason, we included the variant size as a discriminative variable in LRM. Figure 16 and Supplementary Figure 11 show that the F-score of variant callers fluctuated across variant sizes. **This information allowed us to filter potentially false-positive variants, increasing the**

accuracy of the LRMs. These findings indicated that accurate variant calling selection was of paramount importance, in order to cover all SV size ranges efficiently.

In conclusion, this exhaustive benchmarking allowed us to design a pipeline to improve the strengths and decrease the weaknesses of all variant callers, generating different LRMs, one for each SV type. Besides, we decreased the genotype error reported by variant callers, customising and using all information available to be accurate in our decisions. **These LRMs were used to classify the variant detections as potentially true-positive or false-positive in real samples, indispensable for a comprehensive catalogue of variants and ultimately necessary for building the Iberian-GCAT haplotype reference panel.** The LRMs were designed to solve the reads misalignments produced by their short sizes (100-150bp), which increased the false-positive detections. In the future, when TGS technology decreases its costs and sequencing error rates, better SV detections will be done, enabling the discovery of more SVs, mainly in repetitive regions.

5.2 Processing 808 Whole-Genome Sequencing samples from GCAT biobank

The GCAT biobank sequenced 808 samples using whole-genome sequencing. The sampling was performed in different geographic regions within Catalonia, including volunteers with ages between 40-65 years. However, 16% were non-Caucasian, mostly from American-Hispanic origin^{173,174}. For this reason, we discarded 20 samples with non-Iberian representative genetic background (Figure 21A, B, C, E). This was of paramount interest, because variants from other ethnicities in the reference panel could include noise in downstream imputation analyses, resulting in variants not present in the main ethnic group. Besides, the remaining 788 samples overlapped with other Iberian samples across PROPES and 1000G projects, reaffirming the Iberian origin of our cohort (Figure 21C).

Additionally, we discarded two samples due to first and second grade of family relatedness (Figure 21D). This analysis allowed us to obtain approximate allele frequencies in our cohort, avoiding bias produced by particular family variants. Finally, we discarded one sample due to irregularities in variant callers executions. **At the end of sample filtering, we used 785 of the 808 GCAT samples to perform the Iberian-GCAT reference panel.** Although this sample size is smaller than other reference panels, such as HRC or 1000G, which hampering the imputation of rare variants. This resource could elucidate the genetic architecture of the Iberian population, and in turn, improve the imputation of variants from this population, which in our knowledge is still lacking.

The 785 GCAT samples were sequenced at high coverage (30X), opening new opportunities to increase the recall of SVs⁹ and genotype accuracy¹⁵⁶. In Figure 22, we evaluated the number of SVs detected at different coverages, appreciating **that 30X coverage allowed for the discovery of seven times more SVs than 5X coverages.** This provided the means to find more variants and create a haplotype reference panel with more SVs than both 1000G and GoNL, which used 7.4X and 14.5X coverages, respectively. However, not all detection methods were affected equally at different coverages. At 30X, *de novo* Assembly strategies (AS), used by callers such as SvABA, Manta and Popins, were able to detect most SVs. In contrast, the variant detection of CNVnator decreased while the coverage increased, mainly due to at high coverages. This tool was able to detect more accurately deletions and duplications, enabling the improvement of calling accuracy (Supplementary Figure 5). Despite an increase in SVs recall (Figure 22), high

coverages resulted in low precisions (in exception of Read-Depth methods), due to the increasing number of read misalignments⁹. For this reason, in order to take advantage of the high coverage of the project, we combined the variant caller outputs in the LRM, filtering out potential false-positive detections without affecting the recall.

However, high coverage implied computational challenges, such as data storing and processing. For example, the **785 BAM files required 100 Terabytes of space**, requiring a supercomputer such as MareNostrum4. Besides, with an increased coverage, variant callers needed more computational resources to manage all read information, highlighting those software that re-genotype whole variants into the cohort (ex: Lumpy, Delly2), which took more time to execute their pipeline (Supplementary Table 8). Currently, performing variant calling in hundreds of samples is not trivial, and requires a supercomputer to parallelise several executions and solve the computational requirements. Therefore, just considering time, **we needed ~766,663.37 hours to obtain variants from 785 samples using 12 variant callers, spending the equivalent of ~820,329€ in electricity** (Supplementary Table 8). This results demonstrated that a supercomputer was necessary to detect the genome variability of multiple samples sequenced at high coverage. For this reason, future variant callers will need to address these challenges with more efficient strategies to implement these analyses routinely.

5.3 The Iberian-GCAT catalogue description

Initially, the Iberian-GCAT catalogue included 71,885,335 variants. After applying the LRM model and all filtering steps (section 3.7.4), **we obtained a final set of 35,431,441 variants, accepting 49.3% of all discoveries**. As expected, 85.58% of variants were SNVs, followed by indels (14.16%) and finally, Structural Variants (0.25%) (Figure 23B). Also, 78.92% of all variants had a MAF < 5%, with 28.74% of rare and low-frequency variants (Figure 24) (discarding doubletons and singletons). Low-frequency and rare variants are more likely to be associated with diseases. Thus, this catalogue provides new opportunities to find risk variants, and understand the Iberian genetic architecture.

5.3.1. SNV and indel description in the Iberian-GCAT catalogue

Previous studies demonstrated that SNV detections were accurate using single variant callers, obtaining precisions higher than 99% and recalls higher than 96%^{128,129}. However, **we accepted 51.81% of all detected SNVs**, given that nearly half of all variants were detected by at least two variant callers (Figure 23A). We choose to follow this rule based on state of the art and results obtained in benchmarking, where all callers had precisions and recalls above 95%, so we were conservative, deciding that if two callers detected the same SNV was more likely a true-positive than only one. Evaluating this number in detail, we discovered that Deepvariant included the majority of single SNVs calls, maybe due to the re-genotyping step (section 3.6.1.3), increasing the potential false-positive detections. Besides, **we accepted 48% of all indels (1-49 bp)**, showing that the LRM was conservative (Figure 23A). Finally, **we estimated a median of 3.5M SNVs and 606K indels per genome**, which was consistent with previous estimations, of per genome ranges between ~3.3-4M SNVs and ~492K-851K^{10,82,83} indels. The consistency of these results demonstrated that the filter of two variant callers for SNVs, and LRM designed in indels were effective in cleaning-up the potential false discoveries.

Currently, different reference panels included in their sets SNVs and indels. The most recent panels, such as Estonian and Iceland, recovered 16.5M and 31.1M of SNVs and indels, at similar sequencing coverages as our samples but higher sample sizes (Table 1). Overall, we detected 35.3M of SNVs and indels in 785 samples, which is similar to the Iceland reference panel, and nearly the double of variants in comparison to the Estonian reference panel. These results could be explained by the methodology used to detect SNVs and indels by these projects: the Estonian panel used Haplotype caller⁴, and the Iceland reference panel used the Unified GATK genotyper²⁰¹, limiting the variant detection to a single variant caller. For this reason, **we increased the chances to detect more SNVs and indels by using multiple variant callers, highlighting the necessity to use multiple variant callers even in SNVs and indels.**

Finally, after comparing all SNVs and indels of the Iberian catalogue against dbSNP, **19.18% of them were unique, with 84.32% of them were rare variants (MAF < 1%)**. Besides, the majority of new SNVs discovered were singletons (Figure 25A), indicating that variants at high MAFs were well-covered by previous datasets. Particularly, without considering the alternative allele, 9.36% of SNVs were new (Supplementary Table 10), with only 0.11% with a MAF \geq 1%, suggesting that even though it looks like SNVs are widely characterised, there is still room to find novel variants at MAFs < 1%, and polymorphic variants in specific genome positions were not already characterised. Besides, our validation tests confirmed this assumption, showing that >96% of SNVs and their genotypes obtained directly from the SNV-genotyping array (Figure 34A) were concordant with variant calling. For indels, 86.8% of those directly reported by the SNP-genotyping array were concordant with the output of variant calling, with more than 90% of genotype coincidences (Figure 34B). **These results indicated that our variant calling and filtering were applied correctly, obtaining an accurate catalogue of SNVs and indels, recovering novel variants at MAFs < 1%.**

5.3.2. Structural variant description in the Iberian catalogue

As previously mentioned, variant callers produced high FDRs in SV detection⁸. This was corroborated in our calling of SVs, with our LRMs **accepting only 3.07% of all detected SVs**. For this reason, **the LRMs improved SV filtering, resulting in an accurate SV catalogue**. These results suggested that false-positive detections using NGS could be greater than expected.

Particularly, the LRM generated for duplications was the most conservative, accepting less than 1% of all detected duplications (Figure 23A). This was due to CNVnator, which detected a high amount of deletions and duplications compared to other callers (Supplementary Figure 5), many of which were potential false-positives. Besides, several misalignments and artefacts that produce false-positives have been previously described¹⁹⁷ (section 4.2.2). For example, our initial dataset detected 13 of these inversions, filtering all of them after our clean-up procedure (section 3.7.4). These results suggested that our SV catalogue was well-curated.

After the filtering step, **we obtained a dataset of 89,178 SVs, larger than other reference panels such as 1000G or GoNL (Table 1), highlighting the benefits of high-coverage for SV discovery, and of using our LRM filtering method**. Deletions were the most recurrent with 37.3%, followed by MEI (21.1%), insertions (14.3%), inversions (11.4%), translocations (8.8%), and duplications (7.1%) (Figure 27B). This distribution was consistent with 1000G, GoNL and gnomAD-SV projects, except inversions were we recovered a higher number. However, 91.3% of all inversions were singletons and doubletons, which are mostly irrelevant to incorporate in

GWAS. Besides, **76.79% of all SVs were rare variants (MAF < 1%), in contrast to 92% of gnomAD-SV**. This result suggested the necessity to increase the sample size of specific populations, instead of global projects such as gnomAD or 1000G, where the low number of samples from different ethnic groups, could modify the allele frequency particularities of each population. Besides, characterise specific populations, opens new opportunities to find particular common genetic variants that could be hidden in global projects. However, the proportion of singletons and doubletons in the Iberian-GCAT catalogue (58,63%) was consistent with the gnomAD-SV catalogue (49.8%), indicating that 50% of variants in a cohort of these sample sizes were sample-specific. On the other hand, **23.21% of Iberian catalogue SVs had a MAF \geq 1%, and could be relevant for future imputation analyses** (Figure 27A).

Nevertheless, **common variants (MAF \geq 5%) constituted the majority of observed variants in any single genome (85.28%)** (Figure 28A), consistent with 1000G¹, who reported that between 1-4% of variants per genome wherein MAFs < 5%. Besides, deletions (3,327 dels per genome) were the most representative SV type in a single genome, as opposed to inversions (68 inversions per genome) (Figure 28B), consistent with gnomAD-SV¹¹, who reported a median of 3,505 deletions and 14 inversions per genome. **The enrichment of deletions discoveries could be related to the methodological particularities, such as the alignment process and the properties of their genomic regions**. In contrast, inversions were mainly inserted in repetitive regions, hampering their detection. Chaisson et al⁶⁶ reported around 156 inversions per genome, thanks to using long reads on their detections. Therefore, our median inversion estimation per genome could increase using TGS technologies.

The median number of detected SVs per genome has increased with the improvement of sequencing technologies (Figure 5). Our study **estimated ~6,393 SVs per genome**, nearly the double than 1000G (3,441 SVs per genome). However, gnomAD-SV estimated 7,439 SVs per genome, due to the higher sample size, increasing the chances of finding more SVs. Besides, the filtering strategy also affected the number of detected SVs per genome. For example, 1000G and gnomAD-SV use machine learning algorithms. In contrast, GoNL estimated 7,006 SVs per genome, using simple logical rules (more than two callers (\geq callers) detect the same SV) to filter them¹⁶¹. This was above their expected count of SVs per genome, given their project's coverage (14.5X). Hence, GoNL SVs could probably include some false-positives, due to the combination of "bad" variant callers, as we can see in Supplementary Figure 3, where the precision of combining \geq 2 callers was lower than LRM, inflating the SV number in their catalogue. Currently, TGS technologies estimated >20,000 SVs per genome^{66,67}, highlighting the lack of proper characterisation of SVs in the human genome.

The SV median size detected in the Iberian cohort was of 291 bp, consistent with gnomAD-SV (331 bp)¹¹. Further, the bulk of SVs was between 100 bp to 10 Kbp (Figure 26), suggesting that **SV detection using short-reads was favourable for small sizes**. The median size distributed by SV type was 312 bp for deletions, 584 bp for duplications, 1,531 bp for inversions, and 279 bp for MEIs. For 1000G, the median size was 2,455 bp for deletions, 35,890 bp for duplications, 1,697 bp for inversions and 297 bp for MEIs⁵. Overall, **the median size of inversions and MEIs was highly concordant between both projects**, showing for MEIs the three peaks of size distributions, corresponding with ALUs, SVAs and LINEs (Figure 26); thus, corroborating that the calling has been performed correctly. Besides, the short-read improved the inversion discoveries at sizes < 2 Kbp⁶⁶, which is consistent with our results, where the bulk of inversions was at sizes between 1 Kbp - 1.6 Kbp (Figure 26). However, the median size of

deletions and duplications was discrepant when compared with 1000G, mainly due to **the low coverage used in the project, which could result in limitations in the accuracy of size reporting**. Finally, the sizes of *de novo* insertions and translocations could not be evaluated due to the technical limitations of variant callers. For example, for *de novo* insertions, the reads that come from exogenous sequences cannot be mapped in the reference genome, hampering the size estimation. The challenge for translocations was determining both the start and end breakpoints.

Then, **we estimated that 211 MB per genome were affected by SVs**, representing 6% of the genome. Audano et al.⁶⁷ estimated 11 MB without considering duplications, which in our study where the biggest ones (Figure 26), explaining why the higher number of bases affected. However, as mentioned in Manta's documentation, intrachromosomal translocations could be misclassified as large duplications due to read signals errors. Therefore, in future variant caller updates, more signals or strategies would be necessary to use, in order to classify the duplications better.

Finally, we compared the Iberian-GCAT SV catalogue against the most popular SV databases (section 3.8.1.2). **~60% of the SVs were novel**, highlighting that SVs are not yet well characterised. Notwithstanding the efforts to obtain complete SV catalogues, such as gnomAD-SV¹¹ or Abel et al.⁶⁵, further SVs analyses and better sequencing technologies will be needed to obtain more comprehensive SV catalogues. **Besides, 21.22% of all common and 44.32% of all low-frequency SVs described in the Iberian-GCAT catalogue were new** (Figure 25C), demonstrating that SV discovery is skewed compared to SNVs and indels.

Although 21.22% of common and 44.32% of low-frequency SVs were not catalogued, this proportion was not equal if we consider only haplotype reference panels. GoNL and 1000G were the two other projects that included SVs using low-medium coverage. Thus, the proportion of imputable SVs was lower than all SVs discovered by the scientific community. In this context, when compared against the 1000G and GoNL reference panels, **85.40% of the SVs from the Iberian-GCAT catalogue were new, distributed as 52.55% common, 71.63% low-frequency and 93.05% rare SVs** (Figure 25D). Thus, by generating a new reference panel with those SVs, we could enrich the future GWAS analyses with novel SVs, enabling the search for new associations between genetic variants and complex diseases.

We used different approaches to validate the SV discovered by our pipeline. To validate large deletions and duplications (>20Kb), which normally are harder to detect by short-reads, we used the CGH array, validating **76% of deletions and 19.5% of duplications** (Table 19) (section 3.8.2.2). Recall limitations of CGH arrays could explain the low percentage of duplications validated⁵⁴, thus, considering the high percentage of deletions validated, and hard filters applied to consider duplications as true-positive (we accepted less than 1% of all duplications detected in calling step (Figure 23A)), we expected that duplication catalogue was well defined. Besides, 44 of 54 inversions generated by non-homologous recombinations from the Iberian-GCAT catalogue, showed **94.7% of concordance** with experimentally validated inversions from the InvFEST catalogue (Figure 35D), indicating that the calling was performed accurately. In summary, considering the number of variants validated, we consider that our Iberian-GCAT catalogue was well generated.

Overall, most SNVs and indels from our catalogue were already characterised, indicating that our calling was performed correctly, enriching the previous datasets with SNVs and indels with

MAFs < 1%. Besides, we increased the variant detection by performing a variant calling of SNVs and indels using more than one algorithm, improving the genetic characterisation of our cohort. On the other hand, **we detected 60% of new SVs**, as a result of the deep WGS and applying multiple variant callers, covering whole SV types and sizes. **85.40% of SVs were not already included in previous haplotype reference panels, with 52.55% and 71.63% of new common and low-frequency SVs**. Thus, by generating the Iberian-GCAT haplotype reference panel using this catalogue, we could increase the chances of finding more associations between variants and human traits and help explain more of the missing heritability of complex diseases.

5.3.3. The impact of the Structural Variants on human traits

Although Structural Variants (SVs) alter more nucleotides than SNVs and indels, little is known about their functional impact on human traits. In our study, **87.7% of SVs were predicted as “likely benign” or “likely pathogenic”** (Figure 29), showing the limitations of annotation tools. For example, the pathogenicity value developed by ACMG was designed to evaluate highly penetrant variants in rare disorders²⁰². Thus, variants that contribute in small fraction to disease are hard to interpret, due to different levels of penetrance and gene expressivity, which are associated with environmental factors or epistatic mechanisms²⁰³. Those variants are involved in complex diseases, and could be catalogued as benign. This could explain why the inversion associated with metabolic disorders (11q13.2), included in our dataset, was catalogued as likely benign SV¹⁹⁹. Besides, most studies are focused on the functional interpretation of deletions and duplications due to the confidence of their detections, in contrast to other SV types, which is still a challenge to call and validate²⁰⁴. For this reason, deletions and duplications were the most interpretable variants in our catalogue (Figure 29A). However, thanks to high coverage of NGS and TGS sequencing approach, the calling of all SVs will improve, enabling to define better the functional role of all SVs in the future.

SVs could functionally impact genes through mainly two mechanisms, gene expression or gene function loss (predicted loss of function (pLoF)). **A general overview of our dataset demonstrated that 46% (41,672) of all SVs overlapped a gene**. This result was consistent with gnomAD-SV catalogue, where 47.7% of SVs overlapped in gene regions. This number can be partially explained by the current NGS technology based on short reads, which performs poorly in low-complexity regions. Therefore, we were better at characterising SVs outside those regions, where 92.7% of all known autosomal protein-coding nucleotides are localised¹¹. This could thus introduce a bias towards the discovery of SVs overlapping genes. On average, **2,868 SVs overlapped protein and non-protein-coding genes, highlighting the potential impact of SVs on gene function**. However, most SVs overlapped intronic regions (88%), suggesting that SVs implied on loss of gene function could be selected negatively (Figure 30A), resulting in a high number of SVs in introns than exons. This hypothesis was corroborated in Coding Sequencing regions (CDS), where 659 SVs overlapped in those regions (discarding singletons and doubletons). On average, **~70 SVs modified CDS regions per genome**, the majority being common variants (MAF ≥ 5%) (average 54 SVs per genome). This result suggested that common SVs which overlapped in CDS regions could have low penetrance to diseases in comparison to rare variants, explaining this high proportion of common variants in a single genome.

Not all genes tolerate equally sequence alterations. In this context, we evaluated which SVs overlapped genes with a predicted loss of function intolerance (pLI), finding that **32.9% of 35,359 SVs were affecting pLI protein-coding genes**. However, 79% of these SVs had a MAF

< 1% (Figure 31B), indicating that deleterious SVs could be under selection. Overall, it is expected that SVs which modify pLI genes could be penetrant variants in diseases, resulting in rare in populations due to selective pressures, in contrast to common SVs, where their contribution to diseases could be smaller. For this reason, the imputation of rare variants by reference panels is important. However, the challenge remains to increase the sample sizes or combine different reference panels, in order to obtain more haplotypes, which will enable to impute rare variants at high quality, increasing the chances to associate more variants to diseases by GWAS approaches.

Considering the most deleterious SVs (pathogenicity ≥ 4) and their overlap in genes extremely pLI¹⁴⁵ (pLI > 0.9 and Haploinsufficiency), we found that **581 SVs were associated with diseases using the OMIM database**, being especially deletions (Figure 32A). These results indicated that further research is needed to increase the insights of SVs on human diseases, because deletions are the SVs most studied, maybe due to they are easier to detect by current approaches^{18,204}. Besides, **top 10 diseases were related to mental and muscular diseases** (Supplementary Table 5), highlighting the potential role of SVs in human diseases. These results make sense in the context of the studies developed in mental and muscular diseases, where large deletions and duplications play a key role in Autism and Schizophrenia^{204,205}. Besides, on average, the genes expressed in brain are longer and evolutionary conservatives than others. This is a particular feature of Haploinsufficient genes^{147,203}, explaining why the effect of SVs on developmental diseases was better documented than other diseases.

Besides the direct effects on coding sequences, SVs can affect the 3D structure of chromatin, modifying the TADs and their boundaries²⁰⁶. Therefore, to evaluate this hypothesis, we attempted to perform more extensive SV interpretation, beyond simple gene functions. Our results showed that **6,657 SVs potentially modified TAD boundaries, with an enrichment for insertions and translocations**. As TAD regions are located in unfolded chromatin regions, this could increase the chances of producing translocations²⁰⁷ or even could be one way to integrate viruses into DNA²⁰⁸. Surprisingly, we did not detect any MEIs in these genomic regions (Figure 30C), suggesting that MEIs did not insert in those regions or due to a variant caller bias.

Finally, taking advantage of our catalogue, we evaluated if the SVs were in LD with SNVs from the GWAS catalog, to understand if they could be the causal variants underlying an association signal. **3.7% SVs of 36,887 SVs (MAF \geq 1%) were in strong LD ($r^2 \geq 0.8$) with SNVs from the GWAS catalogue**, demonstrating that an SV could be the causal variant from a human trait. Besides, most tag SNPs were located in intronic and intergenic regions, suggesting that the causal variants could be associated with SVs instead of SNVs (Figure 33B), due to their size, which affected more DNA than SNVs, increasing the chances to modify regulatory regions or genes. In addition, 51 of 581 SVs related to diseases (Supplementary Table 6) were tagged by SNPs from the GWAS catalog. **These results demonstrate the necessity to impute SVs in GWAS, increasing their statistical power and finding new disease-associated variants.**

In conclusion, 46% of SVs included in the Iberian-GCAT catalogue overlapped gene regions. However, the vast majority overlapped introns, hampering their functional interpretation. Besides, 35,359 SVs affected protein-coding genes, increasing the chances to find new variants associated with diseases.

5.4 The Iberian-GCAT haplotype panel

One of the central goals of biomedicine is to understand the genetic variability effect in humans, finding risk variants which increase the predisposition to develop a disease. A complete characterisation of risk variants would open a new era in personalised medicine, enabling to design specific treatments for each patient. Genome-Wide Association Studies (GWAS) identified thousands of risk variants, increasing the insights of the genetic architecture on complex diseases¹⁵⁴. Besides, imputing unobserved variants using haplotype reference panels has allowed to include more variants in GWAS, increasing the chances of finding risk variants for diseases¹⁵⁶. However, despite their hypothesised importance on diseases, **SVs are underrepresented in reference panels**, and only 1000G and GoNL included all the SV types (Table 1), but still limited by low-medium coverages.

In this context, we generated the Iberian-GCAT haplotype reference panel using the catalogue of 35,431,441 variants from 785 GCAT samples previously characterised. This reference panel could find variants associated with diseases, increasing the imputation quality of low and rare frequency variants^{4,6}, particularly from the Iberian population, highlighting their relevance in precision medicine. Although SVs can be phased, few studies have shown the performance of phasing tools^{160,162} for these variant types. For this reason, we evaluated the best strategy to phase SVs, **determining that Shapeit4+WhatsHap strategy could recover more SVs of high quality** (Figure 36). Besides, thanks to high coverage, we could use the Phasing Informative Reads (PIRs), improving the imputation quality of rare variants (MAF <1%) (Supplementary Figure 6). When we analysed the effect of PIRs in imputation quality considering all SV types, little improvements were appreciated (Supplementary Figure 8), mainly due to PIRs being currently used to improve imputation of rare SNVs. Overall, the Iberian-GCAT reference panel was the first resource that includes PIR information. However, in the near future, the large read lengths obtained with TGS sequencing technologies will enable the improvement of phasing of variants by using the reads as haplotypes, avoiding the current limitations derived from short reads.

Overall, we took **advantage of high coverage to phase SVs together with biallelic SNVs and indels, obtaining a reference panel with accurate haplotypes, especially for SVs**. For example, 1000G and GoNL obtained SV genotypes using MVNcall, inferring SVs into a haplotype scaffold (constituted of biallelic SNVs and indels), because low coverages do not allow to genotype these genomic rearrangements correctly¹⁶². Then, MVNcall phased the SVs, obtaining the final reference panel. This approach was forced to exclude the non-inferred SVs, causing a loss of SVs. Consequently, it is expected that the imputation quality using these panels to be decreased, due to the low quality of genotypes, mainly for lower variant frequencies¹⁶². For this reason, our strategy offers the possibility to increase the quality of SV imputation, providing a valuable resource in GWAS.

5.4.1. Performance of the Iberian-GCAT haplotype reference panel

We analysed the Iberian-GCAT reference panel imputation capabilities, mainly for SVs, in order to know its strengths and possible impact on GWAS. Although variants imputed with an info score > 0.3 are commonly included in most association tests¹⁵⁶, we filtered out all variants with an info score < 0.7, which corresponds to an allelic dosage (R^2) of 0.5, as recommended GUIDANCE¹⁷² and MaCH²⁰⁹ software, in order to carefully select well-imputed variants^{210,211}. Our

results demonstrate that **more than 80% of common variants included in the Iberian-GCAT reference panel were recovered after imputing the GCAT SNP-genotyping array data, with the exception of duplications (48%) and translocations (19%)** (Figure 37A). However, the imputation quality decreased with the allele frequency, confirming the necessity to increase the sample size of the reference panel, in order to impute at high-quality the variants with MAF < 5%.

The genotypes of imputed variants were concordant (nearly ~100%) with the genotypes reported by our LRM models in common variants, except for duplications (~80%) and translocations (~60%) (Figure 37B). These results demonstrated that our calling and genotyping strategy was highly accurate, because erroneous genotypes and false-positive variants decrease the chances of finding LD patterns between variants¹¹⁶, also affecting the imputation performance¹⁵⁶. However, for low-frequency and rare variants, the genotype accuracy was greater in heterozygous than homozygous alternative variants (Figure 37B), confirming that the allele frequency of homozygous alternative variants were lower than heterozygous, reflecting this decrement in the genotype accuracy.

To understand the low imputation quality for common duplications and translocations, we evaluated the number of SNVs and indels in LD across all SV types, finding that duplications and translocations were those with a lower number of short variants with $r^2 > 0.9$ (Figure 38A). **We estimated that at least 10% of SNVs and indels must be in $r^2 \geq 0.7$ with SVs to obtain good imputation values.** This result demonstrates the importance of performing accurate SNV and indel calling, increasing the chances to obtain more short variants in high LD with SVs, improving SV imputation. Particularly, genotype concordance in imputed duplications was high enough to advise their use in GWAS. However, the genotypes of imputed translocations were not consistent with those reported by WGS calling, discouraging their use (Figure 37B). The low imputation quality for translocations could be produced due to the imputation methodology currently adopted, where all variants evaluated in 1 Mb must be into the same chromosome, losing LD power to impute these events correctly. Besides, the imputation performance of SNVs and indels were not affected by SVs (Figure 38B, Supplementary Figure 9), allowing the use of the Iberian-GCAT reference panel with the majority of variants, with special attention to translocations and duplications.

5.4.1.1. Imputation performance using non-Iberian samples

The genetic structure of populations could affect imputation performance. If the ancestry of both the sample study and the reference panel is the same, the chances to find matching haplotypes increase^{139,156}, improving imputation. Conversely, if ancestries are too distant, the imputation quality could decrease, discouraging the use of the reference panel to impute with an specific population. In this context, we evaluated the imputation quality on populations from different continents (detailed population names in Supplementary Table 7), finding that **the Iberian-GCAT reference panel can be used to impute SNP-genotyping array from different continental groups**, obtaining an imputation quality >95% in European, Latin American and Asian populations (understanding the imputation quality as a percentage of imputed genotypes matched with input SNP-genotyping array), and >92% for African (Figure 41).

The imputation quality grouped different ethnic groups by continents, suggesting that European and Latin American populations were closer genetically from Iberians, in contrast to Asian and African (Figure 41). This was consistent with demographic movements, where Latin American populations, due to colonisation, the genetic background is closer to

Iberians¹⁷⁶. However, Africans have more genetic variants than other populations¹¹ and smaller haplotype blocks due to the high amount of recombinations¹³⁹, hampering the imputation. These results demonstrated that sample filtering was done correctly, showing the consistency between the imputation quality of populations and the ancestry of the Iberian-GCAT reference panel. Conversely, if we added samples representative from other populations, the imputation quality could decrease in European ones.

After imputation analysis, we recovered 49,724 SVs (info score ≥ 0.7), with only ~25% of them included in Asian populations (Figure 42). This result demonstrates that Asian populations require their own reference panels²¹², to include more SVs in GWAS. Besides, **the ancestry distance between populations further affected the imputation of rare variants**^{156,212}, showing that both Iberian populations from 1000G and GCAT imputed the highest amount of SVs with MAFs < 5% (Figure 42). European populations carried more than 40% of imputed SVs (49,724 SVs), with 50% of them having MAF < 5% (Figure 42). In Latin American and African populations, between 35-40% SVs were imputed, with less rare and low-frequency variants recovered in African populations (Figure 42). These results demonstrate that **imputation performance is affected by ancestry diversity**, showing the importance of generating population-specific reference panels, in order to increase the chances to impute at high-quality variants with MAF < 5%.

Additionally, 5,055 of 14,884 SVs shared across all populations were common, with deletions the most representative (Figure 43B). These results suggest that the other SV types could be more population-specific or affected differently by selection pressures; alternatively, this could be a result of variant calling bias derived from short reads, which facilitated the deletion detection, and the limitation to detect variants in segmental duplications, which are enriched in inversions for example⁶⁷. Besides, **most population-specific SVs were rare** (Figure 43C), **highlighting the necessity to build reference population-specific panels**, to increase the chance of imputing rare variants that tend to have functional consequences²¹³.

Overall, considering the genetic background of populations, different SNVs could tag causative/risk variants related to diseases, even different risk/causative variants could be population-specific, showing that not all SVs are equally imputed across populations. For this reason, **population genetic information could be the cornerstone of precision medicine**²¹⁴, detecting genetic particularities of each population, which could be used to improve the diagnosis and treatments of diseases.

5.4.1.2. Accuracy of imputed structural variants

We demonstrated that the Iberian-GCAT reference panel was able to impute SVs in different populations. However, due to ancestry disparities, it does not mean that they are correctly imputed^{6,214}. Therefore, we evaluated the SV imputation accuracy, using nine samples from different populations well characterised with TGS technology from Audano et al. dataset⁶⁷. The majority of SVs shared between projects were common variants because they were the most recurrent across populations (Figure 39A). Imputation recovered ~10% of all the SVs characterised by TGS sequencing (Figure 39B), which was consistent with SV recalls from NGS sequencing⁸ (10-70%), demonstrating that the generation of reference panels using TGS technology could increase the chances to impute more SVs. However, recall increased between 64-81% (Figure 39B), evaluating only the shared variants across projects. This result indicated high imputation performance, because most of variants were recovered.

Surprisingly, the precision of both analyses was critically low (~35%), indicating that the imputation could include false-positives (Figure 39B). However, Hickey et al.¹²¹ using short-reads, genotyped three samples from Audano et al. VCF dataset⁶⁷. This factor enabled to increase the precision up to ~80% and recalls between 79-89% (Figure 39B), demonstrating that the combination of sequencing technologies could increase the accuracy of SV discoveries. This assumption was consistent with results of genotyping tools, where genotyping the SVs with short-reads is a robust method to decrease the false-positive detections^{113,114,121}. Besides, 80% of imputed genotypes matched with those reported by Hickey et al. (Figure 40), reaffirming the Iberian-GCAT reference panel capacities to impute SVs using SNP arrays from different populations. The variant differences between TGS and NGS from Audano and Hickey et al. datasets, could be explained by sequencing errors obtained from TGS technologies, with an estimation of ~1 error every 10 nucleotides, compared to NGS, which it is ~1 error every 1,000 nucleotides⁵⁹. Thus, hard filters in SV discovery for TGS were applied. Overall, **applying TGS sequencing technologies to discover SVs, and genotype using NGS, will result in a better characterisation of SVs.**

In conclusion, the Iberian-GCAT haplotype reference panel will open new opportunities to include SVs in GWAS studies, thanks to their accuracy in imputation analysis, recovering SVs in populations from all continents. However, due to their European ancestry origin, **the Iberian-GCAT reference panel enabled the recovery of more SVs in European populations than in Asian**, recommending the generation of new population-specific reference panels, adding SVs to understand the genetic architecture of diseases in each population.

5.4.2. Benchmarking of multiple haplotype reference panels

We demonstrated that the Iberian-GCAT reference panel is a great resource to impute SVs accurately without losing SNV and indel performance (Figure 37, Figure 38). However, to know the value of this new resource in GWAS, we compared the imputation performance across different reference panels (Figure 44). For SNVs and indels, both the 1000G and the Iberian-GCAT reference panels showed similar imputation performances, where the Iberian-GCAT imputed more indels (Figure 44A), mainly due to high coverage, as reported the new 1000G release²¹⁵. Nevertheless, for rare and low-frequency variants (MAF < 5%), the quality of SNVs and indels was greater in HRC (Figure 44C), showing the value to generate haplotype reference panels with high sample sizes.

On the other hand, **the Iberian-GCAT reference panel outperformed the SV imputation compared to 1000G and GoNL, showing a 2.7 and 1.6-fold increase, respectively** (Figure 44B). Besides, even combining the SV results of 1000G and GoNL did not overcome the SV imputation performance of the Iberian-GCAT reference panel alone (Figure 44B). These results demonstrate the value of the Iberian-GCAT reference panel in SV imputation, being a great resource to improve the resolution of GWAS. **The Iberian reference panel outperformed imputation quality for common variants (MAF ≥ 5%) compared to 1000G and GoNL, highlighting the potential of this resource in GWAS** (Figure 44D). However, for low frequency variants (MAF < 5%), the imputation quality was similar between reference panels, showing the worst values for rare variants (MAF < 1%) (Figure 44D). These results showed the necessity to include more samples in reference panels to obtain good imputation for variants with MAFs < 5%.

In conclusion, **we generated the first reference panel of the Iberian population, including a complete and accurate catalogue of SVs.** This resource will improve future GWAS, adding more SVs than previous reference panels, especially for European and Latin American populations. For this reason, this population-specific reference panel will allow a better understanding of the genetic disease architecture, helping to find new disease-variant associations, and ultimately to improve precision medicine and patient care.

6. CONCLUSIONS

- I. We have first generated a comprehensive strategy to identify and classify all types of detectable germline variants from high-coverage short read sequencing samples, covering from small nucleotide variants to large structural variation. This strategy includes extensive benchmarking and the generation of different Logistic Regression Models that can be used in other similar studies. Estimations using a controlled environment, show overall high precision and recall values for the calling and genotyping of all variant types and sizes.
- II. The application of this strategy to 785 whole-genome sequencing samples from the GCAT-biobank allowed us to identify 35,4 Million variants, corresponding to 30,3 Million Single Nucleotide Variants (SNVs), 5 Million indels (< 50bp) and 89 thousand larger structural variants (≥ 50 bp). This represents a median of 3.5M SNVs, 606K indels and 6,393 SVs per individual in our cohort. This catalogue of variants covers populations frequencies with MAF estimations ranging from singletons to common variants (MAF \geq 5%).
- III. The use of different experimental and comparative validation strategies show an overall high accuracy in the calling and genotyping of our variants, across all variant types and sizes. SNP-genotyping array approaches validated more than 95% of SNVs and 90% indels. The multiple comparisons of our Iberian-GCAT catalogue with existing catalogues of human structural variability, show high consistency and high recall, and demonstrates that the added value of this catalogue, which is centred in SVs, lies on top of consolidated and supported background.
- IV. Of all the SVs from the Iberian-GCAT catalogue, a fraction showed strong gene-related functional impact, compatible with potential roles in diseases and reinforcing their value to identify the missing heritability of complex diseases.
- V. We benchmarked and applied haplotype estimation protocols to build the first genome-wide and haplotype-based reference panel of the Iberian population. This panel was proved to be also useful to impute and predict variants accurately within Iberians, or European populations, but also across a wide range of different ethnicities.
- VI. Finally, as an overall measure of the real value that the Iberian-GCAT reference panel can add to current genetic studies, we have observed a 2.7-fold increase in SV imputation when compared with 1000G.

7. SUPPLEMENTARY MATERIAL

| | | | | | | | | |
|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| HG00096 | HG00099 | HG00101 | HG00103 | HG00106 | HG00109 | HG00113 | HG00173 | HG00176 |
| HG00178 | HG00180 | HG00183 | HG00186 | NA06986 | NA06994 | NA07037 | NA07051 | NA07346 |
| NA07357 | HG00097 | HG00100 | HG00102 | HG00104 | HG00108 | HG00110 | HG00171 | HG00174 |
| HG00179 | HG00182 | HG00185 | NA06984 | NA06989 | NA07000 | NA07048 | NA07056 | NA07347 |
| HG00177 | | | | | | | | |

Supplementary Table 1. Samples form 1000G used to construct the *in-silico* haplotype.

| Sample discarded | BAM low quality | Non-Iberian representative | Familial relatedness |
|------------------|-----------------|----------------------------|----------------------|
| JID047 | | X | |
| JID144 | | X | |
| JID164 | | X | |
| JID239 | | X | |
| JID270 | | X | |
| JID368 | | | X |
| JID399 | | X | |
| JID438 | | X | |
| JID441 | | | X |
| JID466 | | X | |
| JID499 | | X | |
| JID511 | | X | |
| JID533 | | X | |
| JID639 | | X | |
| JID643 | | X | |
| JID645 | | X | |
| JID698 | | X | |
| JID712 | | X | |
| JID744 | | X | |
| JID748 | X | | |
| JID773 | | X | |
| JID777 | | X | |

Supplementary Table 2. Samples discarded from the GCAT cohort after sample filtering.

| | | | | | |
|--------|--------|--------|--------|--------|--------|
| JID025 | JID139 | JID445 | JID556 | JID663 | JID795 |
| JID037 | JID146 | JID455 | JID557 | JID667 | JID797 |
| JID049 | JID151 | JID456 | JID559 | JID669 | JID803 |
| JID051 | JID154 | JID457 | JID565 | JID682 | JID815 |
| JID070 | JID155 | JID463 | JID568 | JID683 | JID836 |
| JID075 | JID171 | JID467 | JID571 | JID686 | |
| JID076 | JID173 | JID470 | JID600 | JID705 | |
| JID081 | JID187 | JID473 | JID602 | JID708 | |
| JID084 | JID193 | JID477 | JID607 | JID717 | |
| JID090 | JID200 | JID481 | JID620 | JID736 | |
| JID093 | JID205 | JID492 | JID622 | JID750 | |
| JID103 | JID208 | JID494 | JID628 | JID753 | |
| JID105 | JID217 | JID495 | JID629 | JID764 | |
| JID110 | JID223 | JID532 | JID632 | JID767 | |
| JID111 | JID226 | JID541 | JID634 | JID768 | |
| JID115 | JID229 | JID543 | JID637 | JID783 | |
| JID119 | JID230 | JID546 | JID654 | JID789 | |

Supplementary Table 3. 95 GCAT Samples used to impute the pilot haplotype reference panel.

| Allele frequency | Benign SV | Likely benign SV | Variant of unknown significance | SV likely pathogenic | SV pathogenic |
|------------------|--------------|------------------|---------------------------------|----------------------|---------------|
| Common | 2299 (42.2%) | 2206 (40.49%) | 62 (1.14%) | 877 (16.1%) | 4 (0.07%) |
| Low Frequency | 157 (29.57%) | 288 (54.24%) | 4 (0.75%) | 86 (16.2%) | 1 (0.19%) |
| Rare | 121 (37%) | 153 (46.79%) | 6 (1.83%) | 50 (15.29%) | 0 |
| Doubleton | 1 (6.25%) | 12 (75%) | 0 | 3 (18.75%) | 0 |
| Singleton | 2 (4.26%) | 37 (78.72%) | 0 | 9 (19.15%) | 0 |

Supplementary Table 4. Number of Structural variants per genome, grouped by pathogenicity grade.

| Phenotypes | num SVs | % |
|---|---------|------|
| Cardiomyopathy, dilated, 3B, 302045 (3)/ Becker muscular dystrophy, 300376 (3)/ Duchenne muscular dystrophy, 310200 (3) | 26 | 4.48 |
| Epileptic encephalopathy, early infantile, 12, 613722 (3) | 25 | 4.30 |
| Mental retardation, AD 33, 616311 (3)/ (Ventricular fibrillation, paroxysmal familial, 2), 612956 (3) | 24 | 4.13 |
| Acrodysostosis 2, with or without hormone resistance, 614613 (3) | 20 | 3.44 |
| Mental retardation, XL 21/34, 300143 (3) | 15 | 2.58 |
| Pitt-Hopkins-like syndrome 2, 614325 (3)/ (Schizophrenia, susceptibility to, 17), 614332 (3) | 15 | 2.58 |
| Mental retardation, AD 39, 616521 (3) | 13 | 2.24 |
| Mental retardation, AR, 6, 611092 (3) | 12 | 2.07 |
| Cerebellar ataxia, nonprogressive, with mental retardation, 614756 (3) | 10 | 1.72 |
| Koolen-De Vries syndrome, 610443 (3) | 10 | 1.72 |

Supplementary Table 5. Top 10 diseases related to SVs using the OMIM database. The SVs evaluated for this analysis where 581 SVs with high pLI, HI and pathogenicity ≥ 4 . The diseases are related to mental and muscular diseases.

| A) | | | B) | | |
|---------|---------|-------|---------------------------------|----------|-------|
| SV type | Num SVs | % | Phenotype | Num SNPs | % |
| DEL | 41 | 86.11 | Heel bone mineral density | 8 | 11.11 |
| INS | 10 | 13.89 | Body mass index | 7 | 9.72 |
| | | | Highest math class taken (MTAG) | 4 | 5.55 |
| | | | Blood protein levels | 3 | 4.16 |
| | | | Height | 3 | 4.16 |
| | | | Menarche (age at onset) | 3 | 4.16 |
| | | | Schizophrenia | 3 | 4.16 |
| | | | Cognitive performance (MTAG) | 2 | 2.77 |
| | | | HDL cholesterol | 2 | 2.77 |
| | | | Monocyte count | 2 | 2.77 |

Supplementary Table 6. 51 unique SVs with disease effect tagged by SNPs in the GWAS catalogue. A) Structural variant type tagged by SNPs. Just deletions and insertions are tagged by SNPs, showing an underrepresentation of other SVs in GWAS catalog. **B)** Phenotypes related to SNPs tagged by SVs in the GWAS catalog.

| Population Code | Population Description | Super Population |
|-----------------|---|------------------|
| ACB | African Caribbeans in Barbados | AFR |
| ASW | Americans of African Ancestry in SW USA | AFR |
| CDX | Chinese Dai in Xishuangbanna, China | EAS |
| CEU | Utah Residents (CEPH) with Northern and Western European Ancestry | EUR |
| CHB | Han Chinese in Beijing, China | EAS |
| CHS | Southern Han Chinese | EAS |
| CLM | Colombians from Medellin, Colombia | AMR |
| FIN | Finnish in Finland | EUR |
| GBR | British in England and Scotland | EUR |
| GIH | Gujarati Indian from Houston, Texas | SAS |
| IBS | Iberian Population in Spain | EUR |
| JPT | Japanese in Tokyo, Japan | EAS |
| KHV | Kinh in Ho Chi Minh City, Vietnam | EAS |
| LWK | Luhya in Webuye, Kenya | AFR |
| MXL | Mexican Ancestry from Los Angeles, USA | AMR |
| PEL | Peruvians from Lima, Peru | AMR |
| PUR | Puerto Ricans from Puerto Rico | AMR |
| TSI | Toscani in Italia | EUR |
| YRI | Yoruba in Ibadan, Nigeria | AFR |

Supplementary Table 7. Populations used in imputation study from 1000G.

| Variant Caller | Total Time (hour) | CPU/h | % |
|--|-------------------|------------------|------------|
| CNVnator | 977.03 | 11,724.34 | 0.34 |
| Whamg +SVTyper | 2,122.87 | 24,384.13 | 0.71 |
| SvABA | 1,867.2 | 29,875.18 | 0.87 |
| strelka2 | 999.05 | 49,414.24 | 1.45 |
| Popins | 4,201.68 | 69,094.31 | 2.02 |
| Manta | 3,194.05 | 153,314.4 | 4.48 |
| Deepvariant | 13,184.25 | 336,301.2 | 9.84 |
| Haplotype Caller | 72,152.83 | 609,258.2 | 17.82 |
| Lumpy | 416,645.28 | 627,228.5 | 18.35 |
| Pindel | 107,925.77 | 647,554.6 | 18.94 |
| Delly2 | 143,393.36 | 860,375.3 | 25.17 |
| Total computational time consumed | 766,663.37 | 3,418,524 | 100 |

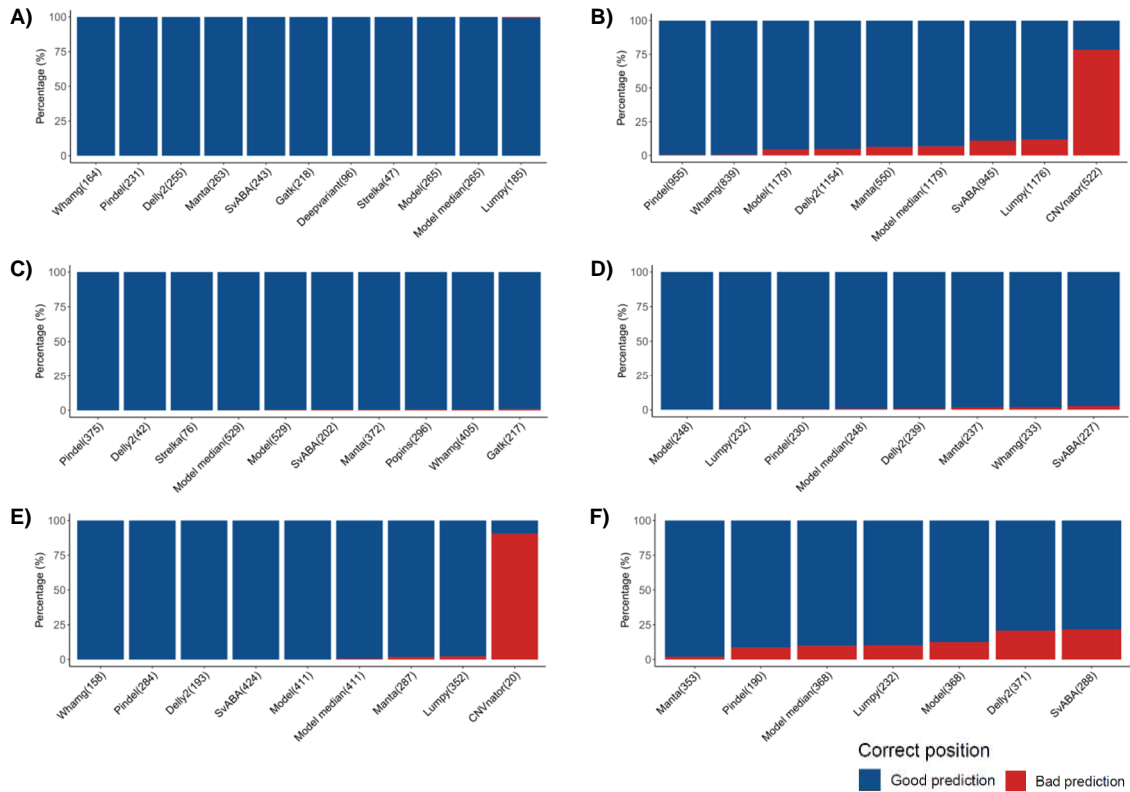
Supplementary Table 8. The Computational time to perform the variant calling of 785 GCAT samples. The computational resources are a bottleneck of variant calling. For example, Delly, Lumpy, Haplotype caller and Deepvariant are executed using all samples together, investing a high amount of time in their executions. The electricity costs needed to perform this variant calling is around 820,329 €, indicating that nowadays, generate a variant calling in current labs is not still available.

| | Unique SNVs | Low frequency | Common | Total SNVs MAF > 1% | Percentage of SNVs MAF > 1% |
|-------------------------------|-------------|---------------|---------|---------------------|-----------------------------|
| Same position | 3,308,128 | 63 | 129 | 192 | 0.01 |
| Same position and alternative | 6,409,906 | 260,645 | 798,536 | 10,59,181 | 16.52 |

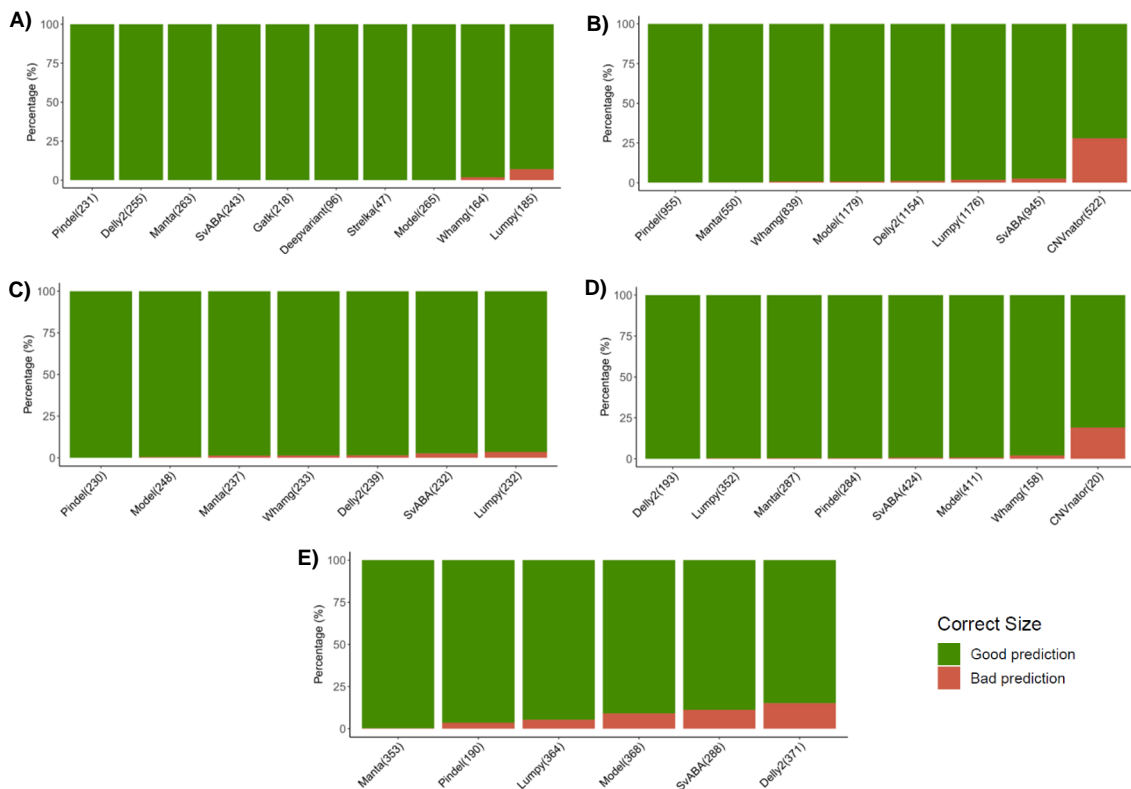
Supplementary Table 9. Number of unique SNVs with MAF > 1%.

| | Total SNV and indels | Unique SNVs | Percentage of unique SNVs (%) | Total unique SNVs and indels | Percentage of unique SNVs and indels (%) |
|-------------------------------|----------------------|-------------|-------------------------------|------------------------------|--|
| Same position | 35,342,263 | 3,308,128 | 9.36 | 3,678,023 | 10.41 |
| Same position and alternative | 35,342,263 | 6,409,906 | 21.14 | 6,779,801 | 19.18 |

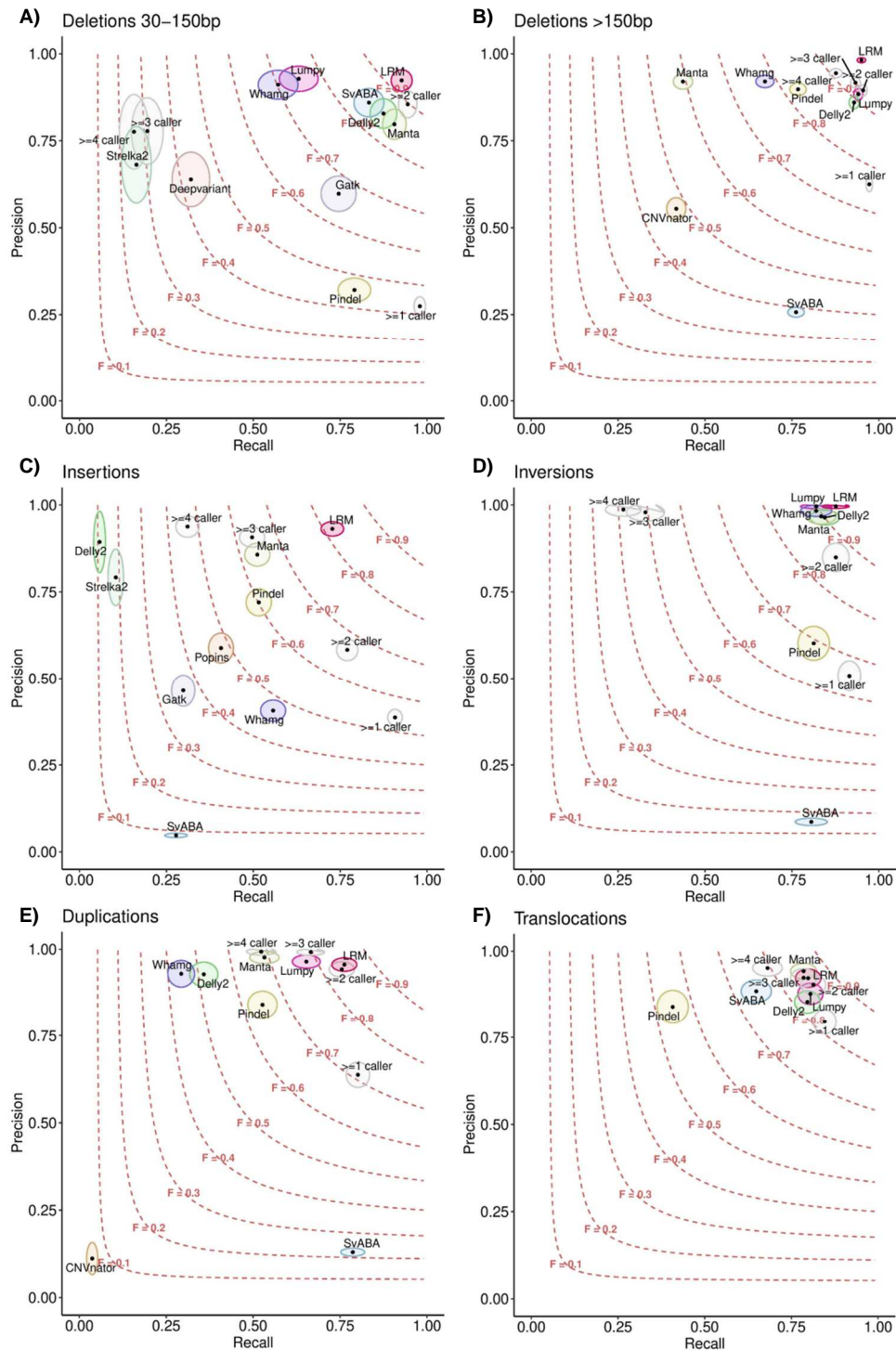
Supplementary Table 10. SNVs and indels in the Iberian-GCAT catalogue. The table shows the number of unique SNVs and indels in Iberian-GCAT catalogue, considering the same position or same position and alternative allele. These results showed that the majority of SNVs and indels are already described, however our panel. Besides, the alternative forms of SNVs are not already included in dbSNP.



Supplementary Figure 1. Position accuracy allowing a breakpoint error of ± 10 bp. The analysis has been performed using the variants detected in the *in-silico*, restricting a breakpoint error of ± 10 bp. Then, we evaluated the percentage of positions in this range. **A)** Mid-Deletions. **B)** Large Deletions. **C)** *De novo* insertions. **D)** Inversions. **E)** Duplications. **F)** Translocations. Model: Report the position using the most accurate variant caller (Table 6). Model median: Using the median value to report the position.

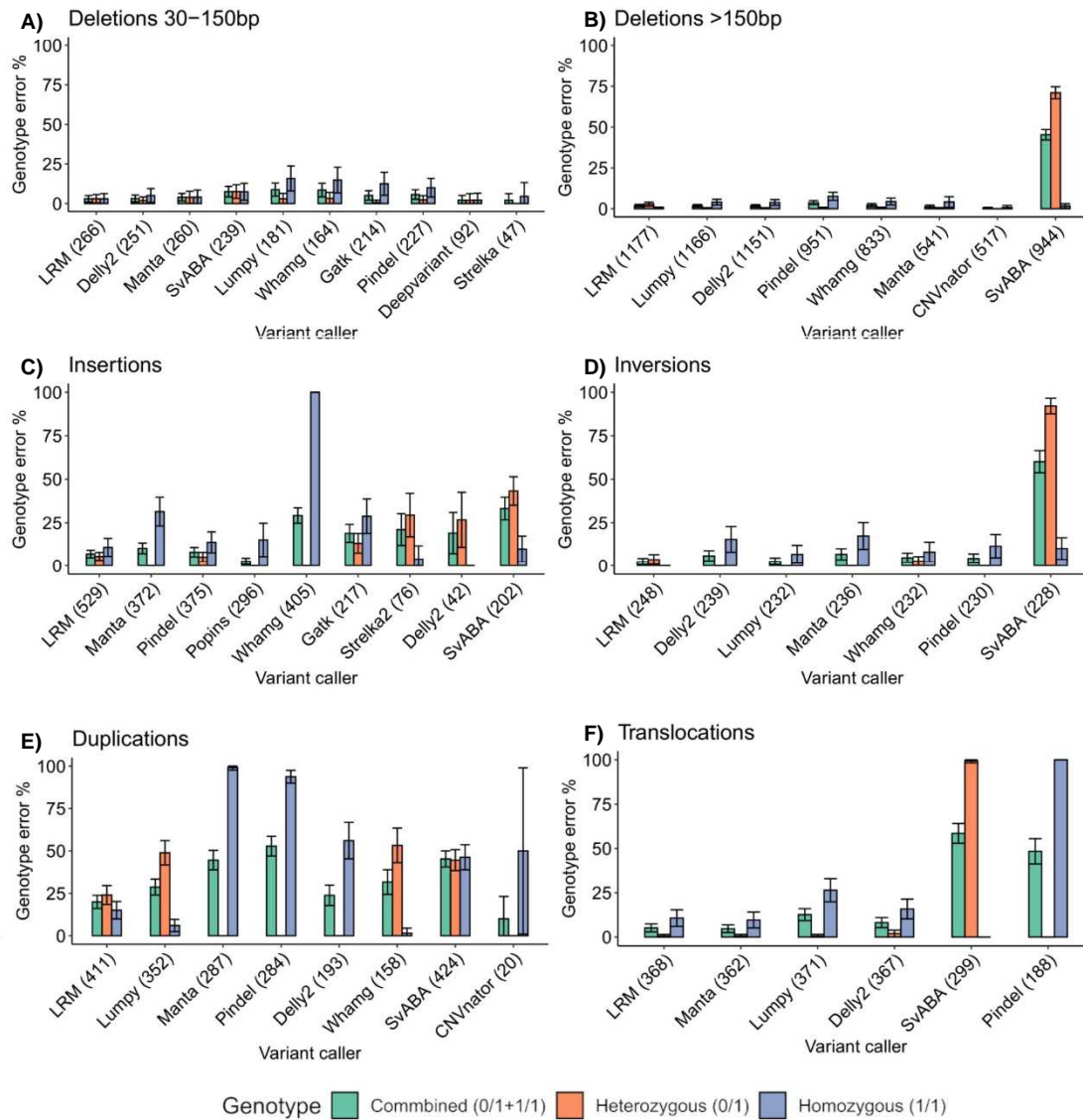


Supplementary Figure 2. Length accuracy allowing 10% of error. To Evaluate the variant caller accuracy to report the length, we allowed 10% of the error to consider the length well predicted. Finally, we calculated the percentage of variants with length well predicted. CNVnator was the worst caller to report the length. Thus, we discarded their predictions to perform the median lengths. **A)** Mid-Deletion. **B)** Large Deletion. **C)** Inversion. **D)** Duplication. **E)** Translocation. Model: Use the median to report the length.

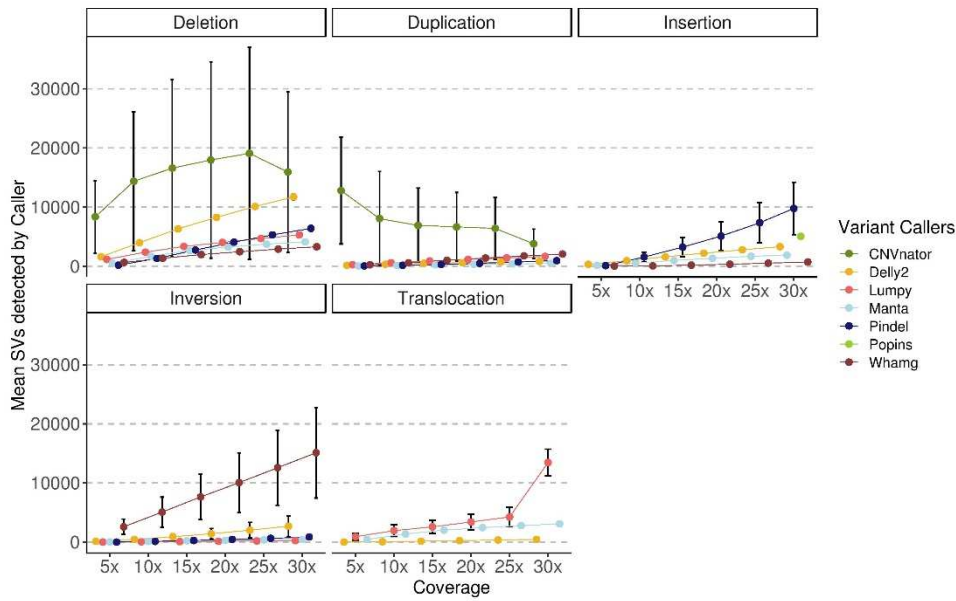


Supplementary Figure 3. Benchmarking analysis classified by variant type. The performance of LRM varies across SV types. **A)** Mid-Deletions (30-150 bp). The LRM outperformed the variant discovery obtaining an F-score 0.93 and precision > 90% compared to logical rules or variant callers individually. **B)** Large Deletions (> 150 bp). The LRM outperformed the variant discovery with an f-score > 0.95 and precision > 98%. **C)** Insertions. The LRM is the most accurate strategy to detect insertions, obtaining an f-score 0.82 with high precision 93%. **D)** Inversions. The LRM improved the discovery of callers slightly individually with 0.93 of f-score. However, the LRM is better than logical rules. **E)** Duplications. No difference is appreciated between LRM and ≥ 2 callers strategies. However, the LRM is 1.3% more precise than ≥ 2 callers. **F)** Translocations. No difference is appreciated between the most accurate variant caller and LRM.

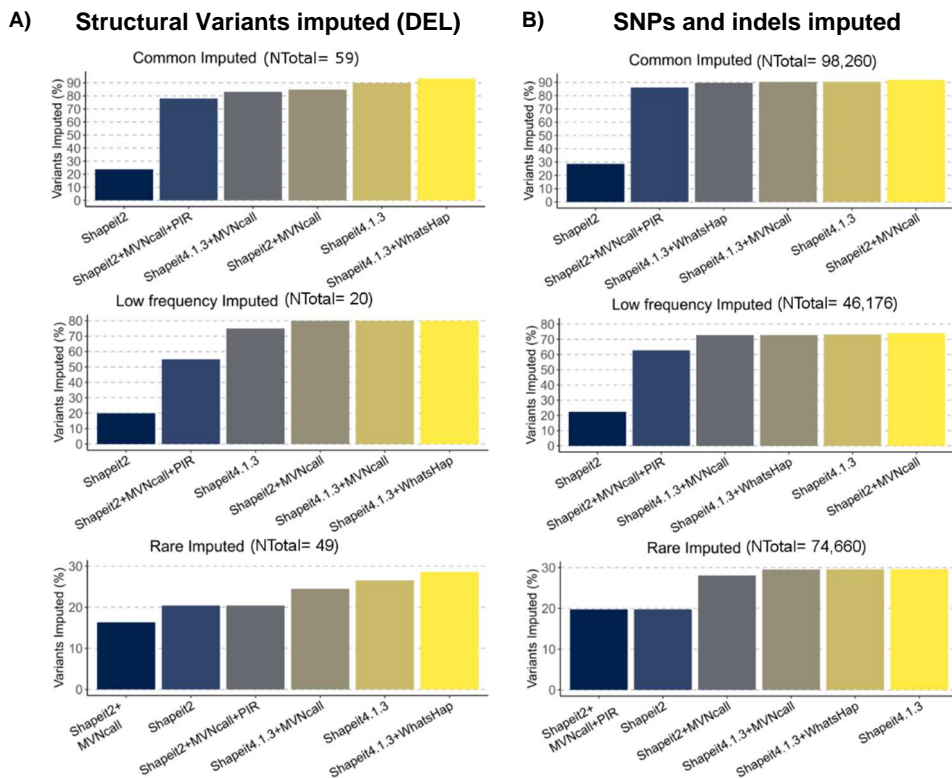
\geq : Logical rules (ex: ≥ 2 callers, at least two callers and methods detect the same variant. This strategy is followed by GoNL project); *LRM*: Logistic Regression Model.; \bigcirc : Confidence interval (CI) area of each algorithm.; $F=$: F-score.



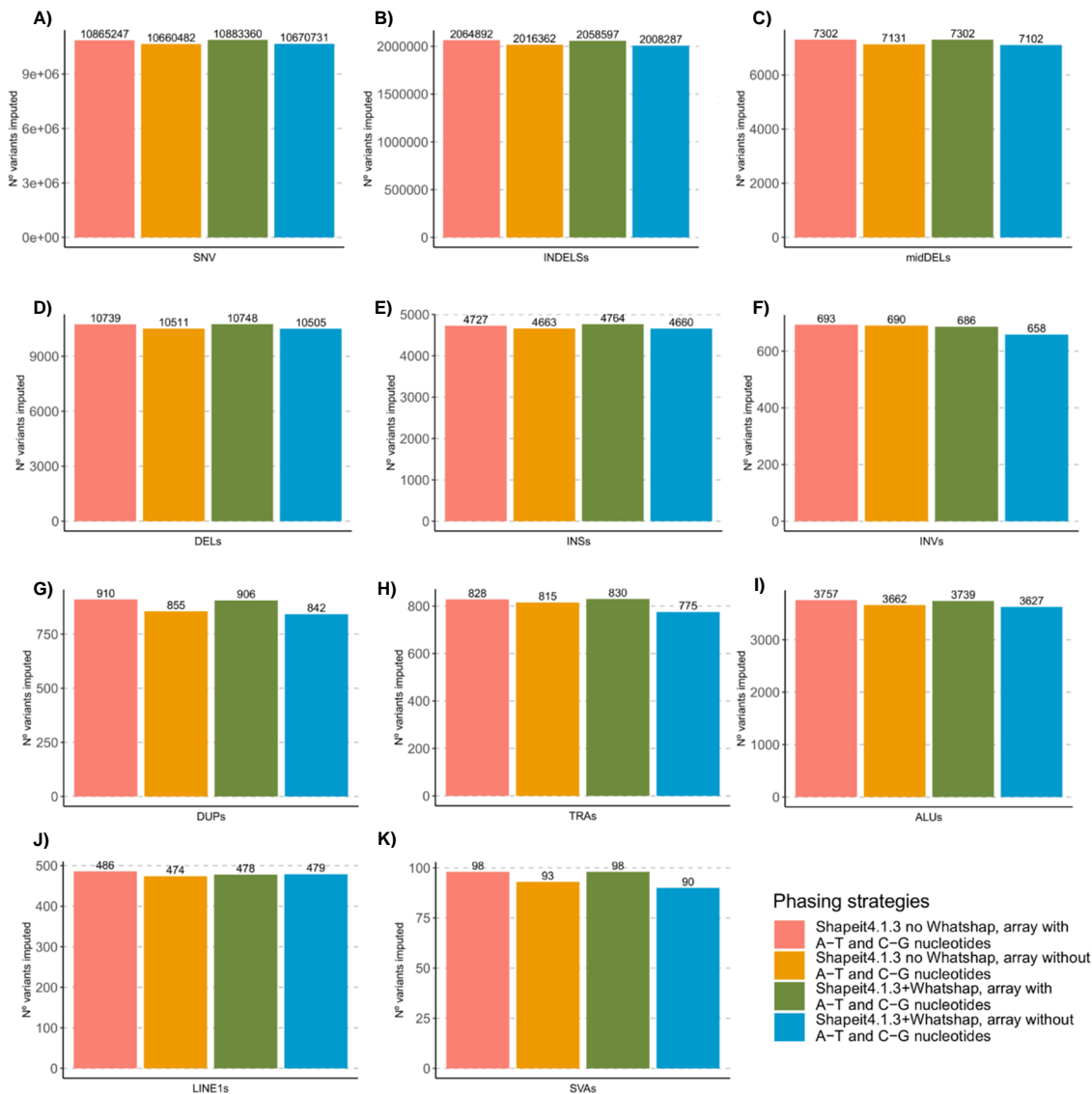
Supplementary Figure 4. The LRM reduced the genotype error in all SV types, highlighting duplications, where the genotyping strategy designed decreased the genotype error < 25%. **A)** Mid-deletions. **B)** Large deletions. SvABA generated errors in heterozygous, being ~75%. **C)** Insertions. Whamg is not able to genotype homozygous variants. **D)** Inversions. SvABA is not able to genotype heterozygous variants being near to ~100%. **E)** Duplications. Manta and Pindel, are not able to genotype homozygous variants. **F)** Translocations. SvABA is not able to genotype homozygous variants and Pindel heterozygous ones.



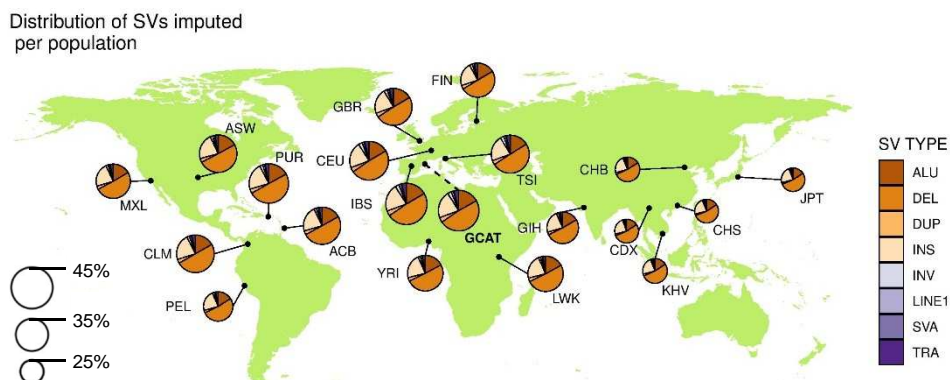
Supplementary Figure 5. Structural variant detection categorised by coverage. While the coverage increases the number of variants discovered rise too. However, CNVnator decreased their detections, mainly in duplications, showing that at high coverages Read Depth (RD) strategies could be more accurate in SV detection. Besides, Popins require at least 30X to be executed, highlighting the importance of coverage in *de novo* insertion discoveries. These result shows the mean of SVs detected at different coverages using 10 samples from GCAT. For this reason, combining all variant callers, the accuracy of SV discoveries could increase.



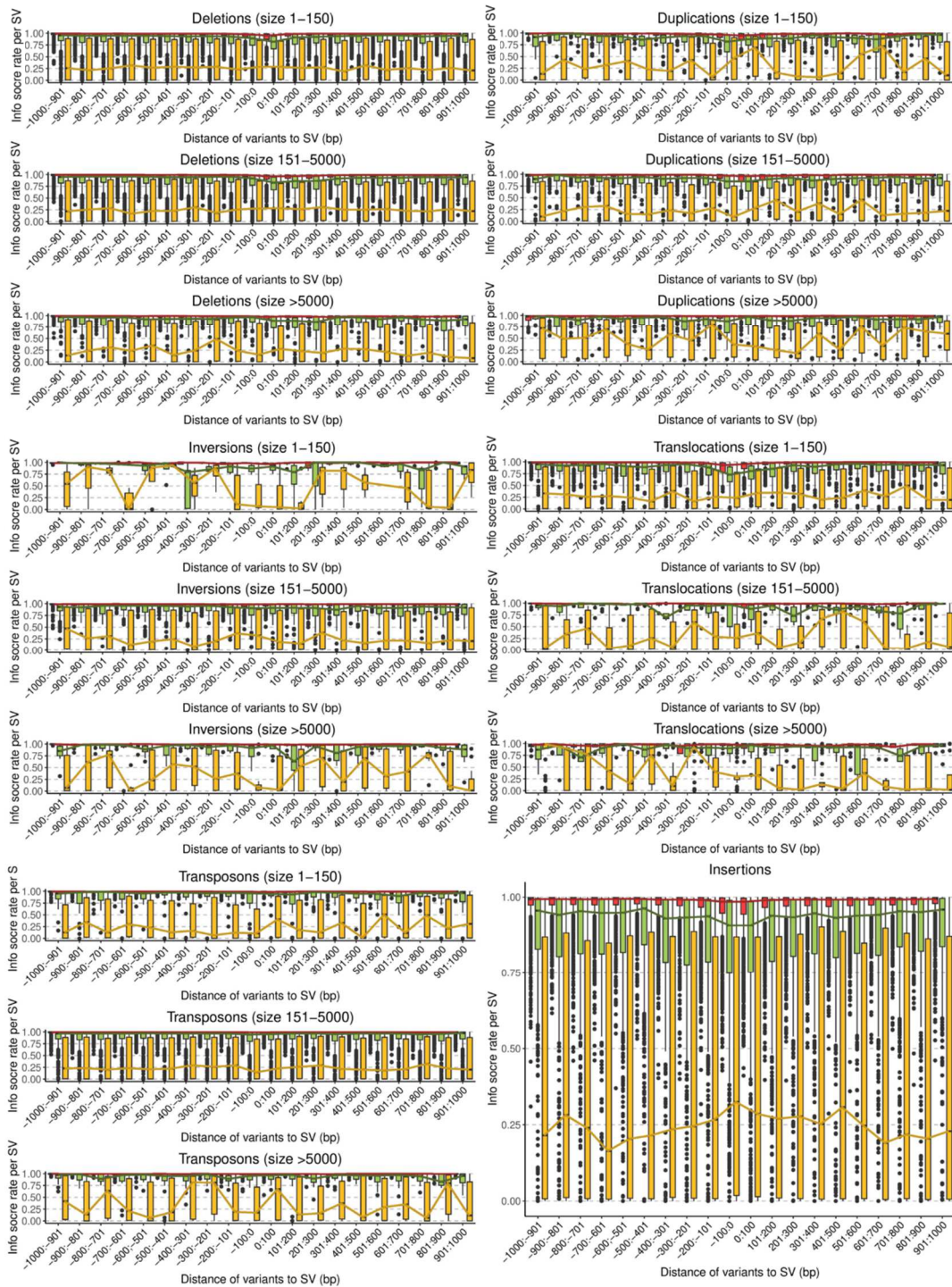
Supplementary Figure 6. Variants imputed in chromosome 22 with info score ≥ 0.7 , using different phasing strategies. A) Large deletions imputed (>150 bp). Using Shapeit4 and phasing informative reads (WhatsHap) improves the imputation of Common variants. Besides, the most improvement is in rare SV variants, where 28.57% of 49 rare SVs were imputed. These results demonstrated that Shapeit4+WhatsHap improves the SV imputation performance. **B)** SNP and indels imputed. There are no remarkably differences between Shapeit2+MVNcall and all Shapeit4 combinations, indicating that the imputation quality of SNPs and indels are not influenced by phasing strategy.



Supplementary Figure 8. Strategies to improve imputation performance. The study evaluated the effect of complementary nucleotides from genotyping SNP array and imputing with a haplotype reference panel generated using phase informative reads (PIRs (WhatsHap)). In conclusion, if the SNP array strand is known, including complementary nucleotides improves the imputation performance. However, generating a reference panel using PIRs do not seem to be related to imputation performance, just improve slightly in SNPs.



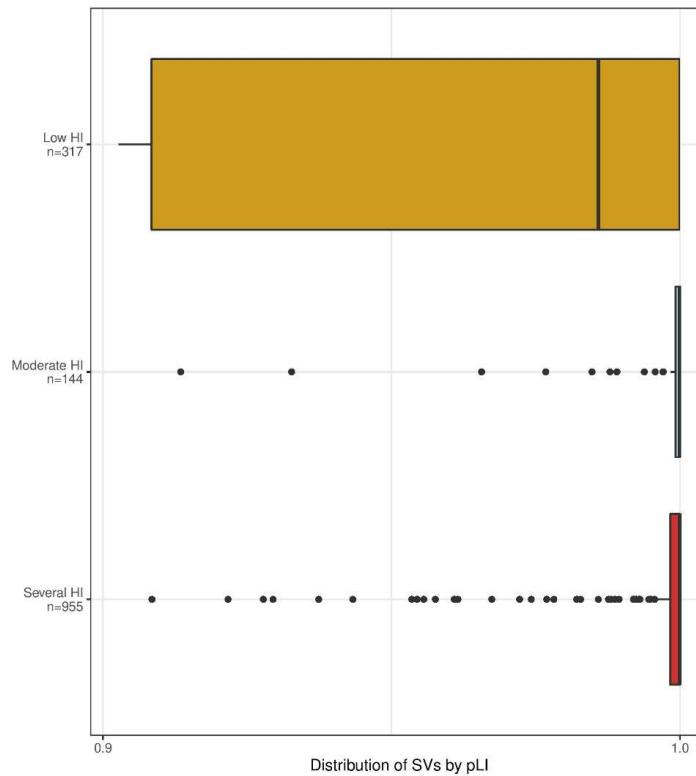
Supplementary Figure 7. Structural Variant (SV) imputed with info score ≥ 0.7 distributed by population and SV type. The majority of SVs imputed are deletions, insertions and Alus. The other SV types are less imputable.



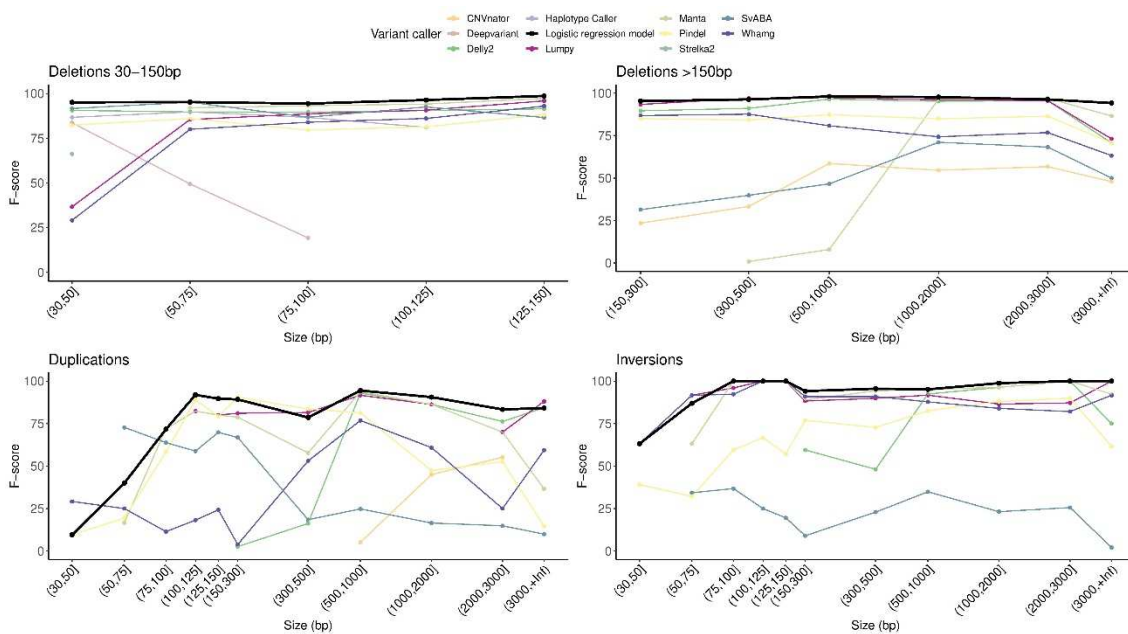
SNPs and indels imputed grouped by MAF

■ Common variant
 ■ Low frequency variant
 ■ Rare variant

Supplementary Figure 9. Structural Variant effect on imputation performance of SNVs and indels. The analysis has been performed evaluating by chunks of 100 bp the SNV and indels quality around different SVs. In figures the center of X axis is the region where the SVs is located, the extremis are the SNVs and indels most far. The imputation of SNVs and indels are not influenced by SVs and their length, as we can see in common (MAF \geq 5%), low-frequency (1% \leq MAF < 5%) and rare (MAF < 1%) variants, just the allele frequency indicated that rare variants are the group with less imputation quality.



Supplementary Figure 10. Structural Variants which modify genes with extremely loss of function intolerance effect.



Supplementary Figure 11. Variant discovery grouped by SV type and classified by SV size. The F-score of variant callers and Logistic Regression Model (LRM) fluctuated across SV sizes, demonstrating different accuracy in specific size ranges. In deletions, the F-score of the Logistic regression model is between 97-99%, depending on the SV size, demonstrating that the variable size improved SV discoveries' accuracy in specific size ranges. In Duplications and Inversions, the F-score of variant callers fluctuated across SV sizes. However, including the SV size in LRM improve the SV discovery of variant callers individually, increasing the SV discovery performance.

8. REFERENCES

1. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
2. Francioli, L. C. *et al.* Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46**, 818–825 (2014).
3. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–89 (2015).
4. Mitt, M. *et al.* Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *Eur. J. Hum. Genet.* **25**, 869–876 (2017).
5. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
6. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).
7. Fisher, E. R. *et al.* The role of race and ethnicity in views toward and participation in genetic studies and precision medicine research in the United States: A systematic review of qualitative and quantitative studies. *Mol. Genet. Genomic Med.* **8**, 1–34 (2020).
8. Mahmoud, M. *et al.* Structural variant calling: The long and the short of it. *Genome Biol.* **20**, 1–14 (2019).
9. Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* **20**, 8–11 (2019).
10. Chiang, C. *et al.* The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).
11. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
12. Gameiro, G. R., Sinkunas, V., Liguori, G. R. & Auler-Júnior, J. O. C. Precision Medicine: Changing the way we think about healthcare. *Clinics (Sao Paulo)*. **73**, e723 (2018).
13. Cristea, M., Noja, G. G., Stefea, P. & Sala, A. L. The impact of population aging and public health support on EU labor markets. *Int. J. Environ. Res. Public Health* **17**, (2020).
14. König, I. R., Fuchs, O., Hansen, G., von Mutius, E. & Kopp, M. V. What is precision medicine? *Eur. Respir. J.* **50**, 1–12 (2017).
15. Ginsburg, G. S. & Phillips, K. A. Precision medicine: From science to value. *Health Aff.* **37**, 694–701 (2018).
16. Campbell, C. D. & Eichler, E. E. Properties and rates of germline mutations in humans. *Trends Genet.* **29**, 575–584 (2013).
17. Séguérel, L., Wyman, M. J. & Przeworski, M. Determinants of mutation rate variation in the human germline. *Annu. Rev. Genomics Hum. Genet.* **15**, 47–70 (2014).
18. Escaramís, G., Docampo, E. & Rabionet, R. A decade of structural variants: Description, history and methods to detect structural variation. *Brief. Funct. Genomics* **14**, 305–314 (2015).

19. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era - Concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).
20. Nielsen, R. *et al.* Tracing the peopling of the world through genomics. *Nature* **541**, 302–310 (2017).
21. Almarri, M. A. *et al.* Population Structure, Stratification, and Introgression of Human Structural Variation. *Cell* **182**, 189-199.e15 (2020).
22. 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).
23. Consortium, I. H. G. S. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
24. E pluribus unum. *Nat. Methods* **7**, 331 (2010).
25. Guo, Y. *et al.* Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics* **109**, 83–90 (2017).
26. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
27. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
28. Pan, B. *et al.* Similarities and differences between variants called with human reference genome HG19 or HG38. *BMC Bioinformatics* **20**, (2019).
29. Duan, Z. *et al.* HUPAN: A pan-genome analysis pipeline for human genomes. *Genome Biol.* **20**, 1–11 (2019).
30. Sherman, R. M. *et al.* Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* **51**, 30–35 (2019).
31. Wong, K. H. Y., Levy-Sakin, M. & Kwok, P. Y. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nat. Commun.* **9**, 1–9 (2018).
32. Sherman, R. M. & Salzberg, S. L. Pan-genomics in the human genome era. *Nat. Rev. Genet.* **21**, 243–254 (2020).
33. Church, D. M. *et al.* Extending reference assembly models. *Genome Biol.* **16**, 2–6 (2015).
34. Goldfeder, R. L. *et al.* Medical implications of technical accuracy in genome sequencing. *Genome Med.* **8**, 1–12 (2016).
35. Paul, D. S., Soranzo, N. & Beck, S. Functional interpretation of non-coding sequence variation: Concepts and challenges. *BioEssays* **36**, 191–199 (2014).
36. Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.* **21**, 71–87 (2020).
37. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)*. **2017**, 1–17 (2017).
38. Consortium, T. E. P. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Genomics* **9**, 1159–1161 (2012).
39. Claudia M. B. Carvalho & Lupski, J. R. Mechanisms underlying structural variant

- formation in genomic disorders. *Nat. Rev. Genet.* **17**, 224–238 (2016).
40. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project The GTEx Consortium* Abstract. *Database Natl. Cent. Biomed. Inf.* **45**, 580–585 (2013).
 41. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
 42. Safran, M. *et al.* GeneCards Version 3: the human gene integrator. *Database (Oxford)*. **2010**, 1–16 (2010).
 43. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
 44. Spataro, N., Rodríguez, J. A., Navarro, A. & Bosch, E. Properties of human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology. *Hum. Mol. Genet.* **26**, 489–500 (2017).
 45. Geeleher, P. & Huang, R. S. Exploring the link between the germline and somatic genome in cancer. *Cancer Discov.* **7**, 354–355 (2017).
 46. Guan, P. & Sung, W.-K. Structural variation detection using next-generation sequencing data: A comparative technical review. *Next Gener. Seq. Applic* **102**, 36–49 (2016).
 47. Sanchis-Juan, A. *et al.* Complex Structural Variants Resolved by Short-Read and Long-Read Whole Genome Sequencing in Mendelian Disorders. *Genome Med.* **7**, 95 (2018).
 48. Quinlan, A. R. & Hall, I. M. Characterizing complex structural variation in germline and somatic genomes. *Trends Genet.* **28**, 43–53 (2012).
 49. McPherson, A. *et al.* NFuse: Discovery of complex genomic rearrangements in cancer using high-throughput sequencing. *Genome Res.* **22**, 2250–2261 (2012).
 50. Mersha, T. B. & Abebe, T. Self-reported race/ethnicity in the age of genomic research: Its potential impact on understanding health disparities. *Hum. Genomics* **9**, 1–15 (2015).
 51. Nagasaki, M. *et al.* Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat. Commun.* **6**, 2–10 (2015).
 52. Teri A., M. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
 53. Dasari, S. & Alex, L. Microarray Based Genotyping: A Review. *J. Cancer Sci.* **1**, (2014).
 54. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
 55. Lamy, P., Grove, J. & Wiuf, C. A review of software for microarray genotyping. *Hum. Genomics* **5**, 304–309 (2011).
 56. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
 57. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1–8 (2016).
 58. Bunnik, E. M. & Le Roch, K. G. An Introduction to Functional Genomics and

- Systems Biology. *Adv. Wound Care* **2**, 490–498 (2013).
59. Giani, A. M., Gallo, G. R., Gianfranceschi, L. & Formenti, G. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Comput. Struct. Biotechnol. J.* **18**, 9–19 (2020).
 60. Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science (80-.)*. **318**, 420–426 (2007).
 61. Mantere, T., Kersten, S. & Hoischen, A. Long-read sequencing emerging in medical genetics. *Front. Genet.* **10**, 1–14 (2019).
 62. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269–276 (2011).
 63. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
 64. Sims, D., Sudbery, I., Illott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: Key considerations in genomic analyses. *Nat. Rev. Genet.* **15**, 121–132 (2014).
 65. Abel, H. J. *et al.* Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**, (2020).
 66. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1–16 (2019).
 67. PA. Audano, A. Sulovari, TA. Graves-Lindsay, S. Cantsilieris, M. Sorensen, AE. Welch, ML. Dougherty, BJ. Nelson, A. Shah, SK. Dutcher, WC. Warren, V. Magrini, SD. McGrath, YI. Li, RK. Wilson, EE. Eichler. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**, 663–675 (2019).
 68. Thankaswamy-Kosalai, S., Sen, P. & Nookaew, I. Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. *Genomics* **109**, 186–191 (2017).
 69. Metzker, M. L. Sequencing technologies the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).
 70. Tischler, G. & Leonard, S. Biobambam: Tools for read pair collation based algorithms on BAM files. *Source Code Biol. Med.* **9**, 1–18 (2014).
 71. Rausch, T., Hsi-Yang Fritz, M., Korbel, J. O. & Benes, V. Alfred: Interactive multi-sample BAM alignment statistics, feature counting and feature annotation for long- and short-read sequencing. *Bioinformatics* **35**, 2489–2491 (2019).
 72. Auwera, G. A. Van der & Mauricio O. Carneiro, Chris Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, Eric Banks, Kiran V. Garimella¹, David Altshuler, Stacey Gabriel, and M. A. D. *From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics* **43**, (2013).
 73. Webster, T. H. *et al.* Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data. *Gigascience* **8**, 1–11 (2019).
 74. Olney, K. C., Brotman, S. M., Andrews, J. P., Valverde-Vesling, V. A. & Wilson, M. A. Reference genome and transcriptome informed by the sex chromosome complement of the sample increase ability to detect sex differences in gene

- expression from RNA-Seq data. *Biol. Sex Differ.* **11**, 1–18 (2020).
75. Geoffroy, V. *et al.* AnnotSV: An integrated tool for structural variations annotation. *Bioinformatics* **34**, 3572–3574 (2018).
 76. Lin, K., Bonnema, G., Sanchez-Perez, G. & De Ridder, D. Making the difference: Integrating structural variation detection tools. *Brief. Bioinform.* **16**, 852–864 (2014).
 77. Tattini, L., D’Aurizio, R. & Magi, A. Detection of genomic structural variants from next-generation sequencing data. *Front. Bioeng. Biotechnol.* **3**, 1–8 (2015).
 78. Kronenberg, Z. N. *et al.* Wham: Identifying Structural Variants of Biological Consequence. *PLoS Comput. Biol.* **11**, 1–19 (2015).
 79. Wala, J. A. *et al.* SvABA: Genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28**, 581–591 (2018).
 80. Kehr, B., Melsted, P. & Halldórsson, B. V. PopIns: Population-scale detection of novel sequence insertions. *Bioinformatics* **32**, 961–967 (2016).
 81. Poplin, R. *et al.* A universal snp and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983 (2018).
 82. Shen, H. *et al.* Comprehensive Characterization of Human Genome Variation by High Coverage Whole-Genome Sequencing of Forty Four Caucasians. *PLoS One* **8**, (2013).
 83. Kishikawa, T. *et al.* Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data. *Sci. Rep.* **9**, 1–10 (2019).
 84. Rahman, N. Realising the Promise of Cancer Predisposition Genes Nazneen. *Nature* **15**, 302–308 (2014).
 85. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
 86. Nielsen, R., Paul, J., Albrechtsen, A. & Song, Y. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **12**, 443–451 (2011).
 87. Kumaran, M., Subramanian, U. & Devarajan, B. Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data. *BMC Bioinformatics* **20**, 1–11 (2019).
 88. Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555–560 (2019).
 89. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178 (2017). doi:10.1101/201178
 90. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. 1–9 (2012).
 91. Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z. & Twigg, S. R. F. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet* **46**, 912–918 (2014).
 92. 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).
 93. Daniel C. Koboldt, Larson, D. E. & Wilson, R. K. Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection. *Curr Protoc Bioinforma.* **44**,

15.4.1–15.4.17 (2013).

94. Sangtae. Kim, Konrad. Scheffer, Aaron. Halpern, Mitchell. Bekritsky, Noh. Enhuo, Morten. Källberg, Xiaoyu. Chen, Kim. Yeobin, Doruk. Beyter, Peter. Krusche, C. S. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**, 591–594 (2018).
95. Ho, S. S., Urban, A. E. & Mills, R. E. Structural variation in the sequencing era Nature reviews | Genetics. *Nat. Rev. Genet.* (2019). doi:10.1038/s41576-019-0180-9
96. Lappalainen, I. *et al.* DbVar and DGVa: Public archives for genomic structural variation. *Nucleic Acids Res.* **41**, 936–941 (2013).
97. 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. *Nucleic Acids Res.* **42**, 986–992 (2014).
98. Handsaker, R. E. *et al.* Large multiallelic copy number variations in humans. *Nat. Genet.* **47**, 296–303 (2015).
99. Mills, R. E. *et al.* Mapping copy number variation by population scale genome sequencing. *Nature* **470**, 59–65 (2011).
100. Teo, S. M., Pawitan, Y., Ku, C. S., Chia, K. S. & Salim, A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics* **28**, 2711–2718 (2012).
101. Rausch, T. *et al.* DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, 333–339 (2012).
102. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: A probabilistic framework for structural variant discovery. *Genome Biol.* **15**, 1–19 (2014).
103. Ye, K. *et al.* Split-read indel and structural variant calling using PINDEL. *Methods Mol. Biol.* **1833**, 95–105 (2018).
104. Chen, X. *et al.* Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
105. De Coster, W. & Van Broeckhoven, C. Newest Methods for Detecting Structural Variations. *Trends Biotechnol.* **37**, 973–982 (2019).
106. 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).
107. Gardner, E. J. *et al.* The mobile element locator tool (MELT): Population-scale mobile element discovery and biology. *Genome Res.* **27**, 1916–1929 (2017).
108. Nguyen, N. P. D., Deshpande, V., Luebeck, J., Mischel, P. S. & Bafna, V. ViFi: Accurate detection of viral integration and mRNA fusion reveals indiscriminate and unregulated transcription in proximal genomic regions in cervical cancer. *Nucleic Acids Res.* **46**, 3309–3325 (2018).
109. Wang, Q., Jia, P. & Zhao, Z. VERSE: A novel approach to detect virus integration in host genomes through reference genome customization. *Genome Med.* **7**, (2015).
110. Hehir-Kwa, J. Y. *et al.* A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat. Commun.* **7**, 1–10

- (2016).
111. Roslin, N. M., Weili, L., Paterson, A. D. & J, S. L. Quality control analysis of the 1000 Genomes Project Omni2.5 genotypes. *bioRxiv* 1–27 (2016). doi:<https://doi.org/10.1101/078600>
 112. Kai, Y., George, H. & Zemin, N. Structural Variation Detection from Next Generation Sequencing. *J. Next Gener. Seq. Appl.* **01**, (2016).
 113. Chander, V., Gibbs, R. A. & Sedlazeck, F. J. Evaluation of computational genotyping of structural variation for clinical diagnoses. *Gigascience* **8**, 1–7 (2019).
 114. Lecompte, L., Peterlongo, P., Lavenier, D. & Lemaitre, C. SVJedi: Genotyping structural variations with long reads. *Bioinformatics* 1–8 (2020). doi:10.1093/bioinformatics/btaa527
 115. Eggertsson, H. P. *et al.* GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* **10**, 1–8 (2019).
 116. Hou, L. *et al.* Impact of genotyping errors on statistical power of association tests in genomic analyses: A case study. *Genet. Epidemiol.* **41**, 152–162 (2017).
 117. Pompanon, F., Bonin, A., Bellemain, E. & Taberlet, P. Genotyping errors: Causes, consequences and solutions. *Nat. Rev. Genet.* **6**, 847–859 (2005).
 118. Maruki, T. & Lynch, M. Genotype calling from population-genomic sequencing data. *G3 Genes, Genomes, Genet.* **7**, 1393–1404 (2017).
 119. Chiang, C. *et al.* SpeedSeq: Ultra-fast personal genome analysis and interpretation. *Nat Methods* **12**, 966–968 (2015).
 120. Sibbesen, J. A., Maretty, L. & Krogh, A. Accurate genotyping across variant classes and lengths using variant graphs. *Nat. Genet.* **50**, 1054–1059 (2018).
 121. Hickey, G. *et al.* Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* **21**, 1–17 (2020).
 122. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).
 123. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 1–26 (2016).
 124. Zook, J. M. *et al.* An open resource for accurately benchmarking small variant and reference calls. *Nat. Biotechnol.* **37**, 561–566 (2019).
 125. Zook, J. M. *et al.* A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38**, (2020).
 126. Escalona, M., Rocha, S. & Posada, D. A comparison of tools for the simulation of genomic next-generation sequencing data. *Nat. Rev. Genet.* **17**, 459–469 (2017).
 127. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: A next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).
 128. Chen, J., Li, X., Zhong, H., Meng, Y. & Du, H. Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. *Sci. Rep.* **9**, 1–13 (2019).
 129. Supernat, A., Vidarsson, O. V., Steen, V. M. & Stokowy, T. Comparison of three variant callers for human whole genome sequencing. *Sci. Rep.* **8**, 1–6 (2018).

130. Cameron, D. L., Di Stefano, L. & Papenfuss, A. T. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat. Commun.* **10**, 1–11 (2019).
131. Abyzov, A. & Gerstein, M. AGE: Defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics* **27**, 595–603 (2011).
132. Becker, T. *et al.* FusorSV: An algorithm for optimally combining data from multiple structural variation detection methods. *Genome Biol.* **19**, 1–14 (2018).
133. Wong, K., Keane, T. M., Stalker, J. & Adams, D. J. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol.* **11**, (2010).
134. Mohiyuddin, M. *et al.* MetaSV: An accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics* **31**, 2741–2744 (2015).
135. Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 1–11 (2017).
136. Zarate, S. *et al.* Parliament2: Fast Structural Variant Calling Using Optimized Combinations of Callers. *bioRxiv* 424267 (2018). doi:10.1101/424267
137. Sperandei, S. Understanding logistic regression analysis. *Biochem. Medica* **24**, 12–18 (2014).
138. Slatkin, M. Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9**, 477–485 (2008).
139. Bush, W. S. & Moore, J. H. Chapter 11: Genome-Wide Association Studies. *PLoS Comput. Biol.* **8**, (2012).
140. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
141. McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40**, 1166–1174 (2008).
142. Havrilla, J. M., Pedersen, B. S., Layer, R. M. & Quinlan, A. R. A map of constrained coding regions in the human genome. *Nat. Genet.* **51**, 88–95 (2019).
143. Lin, Y. C., Tseng, J. T., Jeng, S. L. & Sun, H. S. Comprehensive analysis of common coding sequence variants in Taiwanese Han population. *Biomarkers Genomic Med.* **6**, 133–143 (2014).
144. Shameer, K., Tripathi, L. P., Kalari, K. R., Dudley, J. T. & Sowdhamini, R. Interpreting functional effects of coding variants: Challenges in proteome-scale prediction, annotation and assessment. *Brief. Bioinform.* **17**, 841–862 (2016).
145. Fuller, Z. L., Berg, J. J., Mostafavi, H., Sella, G. & Przeworski, M. Measuring intolerance to mutation in human genetics. *Nat. Genet.* **51**, 772–776 (2019).
146. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
147. Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet.* **6**, 1–11 (2010).
148. Chen, Y. *et al.* Association of Clinical Phenotypes in Haploinsufficiency A20

- (HA20) With Disrupted Domains of A20. *Front. Immunol.* **11**, 1–8 (2020).
149. James R, L., John W, B., Eric, B. & Richard A, G. Clan Genomics and the Complex Architecture of Human Disease. *Cell* **147**, 32–43 (2011).
 150. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789–D798 (2015).
 151. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
 152. Cingolani, P. *et al.* 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 (2012).
 153. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, 1–7 (2010).
 154. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).
 155. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
 156. Das, S., Abecasis, G. R. & Browning, B. L. Genotype imputation from large reference panels. *Annu. Rev. Genomics Hum. Genet.* **19**, 73–96 (2018).
 157. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
 158. Génin, E. Missing heritability of complex diseases: case solved? *Hum. Genet.* **139**, 103–113 (2020).
 159. Christine, L. Algorithms for Haplotype Phasing. (2014).
 160. Choi, Y., Chan, A. P., Kirkness, E., Telenti, A. & Schork, N. J. Comparison of phasing strategies for whole human genomes. *PLoS Genet.* **14**, 1–26 (2018).
 161. 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).
 162. Marchini, J. Haplotype Estimation and Genotype Imputation. in *Handbook of Statistical Genomics* (eds. David, B., Ida, M. & John, M.) **1**, 87–114 (John Wiley & Sons Ltd, 2019).
 163. Delaneau, O., Howie, B., Cox, A. J., Zagury, J. F. & Marchini, J. Haplotype estimation using sequencing reads. *Am. J. Hum. Genet.* **93**, 687–696 (2013).
 164. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
 165. Loh, P. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
 166. Delaneau, O., Zagury, J. F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 24–

- 29 (2019).
167. Patterson, M. D. *et al.* WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *J. Comput. Biol.* **22**, 498–509 (2015).
 168. Rodriguez, O. L., Ritz, A., Sharp, A. J. & Bashir, A. MsPAC: A tool for haplotype-phased structural variant detection. *Bioinformatics* **36**, 922–924 (2020).
 169. Menelaou, A. & Marchini, J. Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics* **29**, 84–91 (2013).
 170. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
 171. 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**, (2009).
 172. Guindo-martínez, M., Amela, R. & Bonàs-guarch, S. The impact of non-additive genetic associations on age-related complex diseases . Corresponding authors : Josep M Mercader Programs in Metabolism and Medical and Population Genetics Broad Institute of Harvard and MIT 75 Ames St 02142 , Cambridge , MA Unit. *bioRxiv* (2020).
 173. Obón-Santacana, M. *et al.* GCAT|Genomes for life: A prospective cohort study of the genomes of Catalonia. *BMJ Open* **8**, (2018).
 174. Galván-Femenía, I. *et al.* Multitrait genome association analysis identifies new susceptibility genes for human anthropometric variation in the GCAT cohort. *J. Med. Genet.* 765–778 (2018). doi:10.1136/jmedgenet-2018-105437
 175. Bycroft, C. *et al.* Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. *Nat. Commun.* **1**, 551 (2019).
 176. Homburger, J. R. *et al.* Genomic Insights into the Ancestry and Demographic History of South America. *PLoS Genet.* **11**, 1–26 (2015).
 177. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
 178. Pàmols, T. *et al.* A view on clinical genetics and genomics in Spain: Of challenges and opportunities. *Mol. Genet. Genomic Med.* **4**, 376–391 (2016).
 179. Weinstein, John N; Collisson, Eric A; Mills, Gordon B; Shaw, K. M., Ozenberger, Brad A; Ellrott, Kyle; Shmulevich, Ilya; Sander, Chris; Stuart, J. M, and C. & Network, G. A. R. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.* **45**, 1113–1120 (2013).
 180. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 4–9 (2015).
 181. K. Uguen, C. Jubin, Y. Duffourd, C. Bardel, V. Malan, J. Dupont, L. El Khattabi, N. Charton, A. Vitobello, P. Rollat-Farnier, C. Baulard, M. Lelorch, A. Leduc, E. Tisserant, F. Mau-Them, V. Danjean, M. Delepine, M. Till, V. Meyer, S. Lyonnet, A. M.-B. Genome sequencing in cytogenetics: Comparison of short-read and linked-read approaches for germline structural variant detection and characterization. *Mol. Genet. Genomic Med.* **e11114.**, (2020).
 182. Kavak, P. *et al.* Discovery and genotyping of novel sequence insertions in many sequenced individuals. *Bioinformatics* **33**, i161–i169 (2017).

183. Liu, S. *et al.* 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**, (2015).
184. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91**, 839–848 (2012).
185. Wang, C. *et al.* Ancestry estimation and control of population stratification for sequence-based association studies. *Nat. Genet.* **46**, 409–415 (2014).
186. Galván-Femenía, I., Graffelman, J. & Barceló-i-Vidal, C. Graphics for relatedness research. *Mol. Ecol. Resour.* **17**, 1271–1282 (2017).
187. 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**, 1–14 (2017).
188. Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
189. 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.* **23**, 955–959 (2012).
190. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
191. Zook, J. M. *et al.* A robust benchmark for germline structural variant detection. *bioRxiv* 664623 (2019). doi:10.1101/664623
192. Pei, S. *et al.* Benchmarking variant callers in next-generation and third-generation sequencing analysis. *Brief. Bioinform.* **00**, 1–11 (2020).
193. Zhang, Y., Imoto, S., Miyano, S. & Yamaguchi, R. Enhancing breakpoint resolution with deep segmentation model: a general refinement method for read-depth based structural variant callers. *bioRxiv* 503649 (2020). doi:10.1101/503649
194. Yan, Q. *et al.* The impact of genotype calling errors on family-based studies. *Sci. Rep.* **6**, 4–9 (2016).
195. Huddleston, J. *et al.* Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–685 (2017).
196. Wong, L. P. *et al.* Deep whole-genome sequencing of 100 southeast Asian Malays. *Am. J. Hum. Genet.* **92**, 52–66 (2013).
197. Vicente-Salvador, D. *et al.* Detailed analysis of inversions predicted between two human genomes: Errors, real polymorphisms, and their origin and population distribution. *Hum. Mol. Genet.* **26**, 567–581 (2017).
198. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
199. González, J. R. *et al.* Polymorphic Inversions Underlie the Shared Genetic Susceptibility of Obesity-Related Diseases. *Am. J. Hum. Genet.* **106**, 846–858 (2020).

200. Zheng, G. X. Y. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* **34**, 303–311 (2016).
201. Jónsson, H. *et al.* Data Descriptor : Whole genome characterization of sequence diversity of 15 , 220 Icelanders. 1–9 (2017).
202. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
203. 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).
204. Han, L. *et al.* Functional annotation of rare structural variation in the human brain. *Nat. Commun.* **11**, (2020).
205. William M, B. *et al.* Paternally inherited cis-regulatory structural variants are associated with autism. *Science (80-.).* **20**, 327–331 (2018).
206. Shanta, O. *et al.* The effects of common structural variants on 3D chromatin structure. *BMC Genomics* **21**, 1–10 (2020).
207. Valton, A.-L. & Dekker, J. TAD disruption as oncogenic driver. *Curr. Opin. Genet. Dev.* 34–40 (2016). doi:10.1016/j.gde.2016.03.008.TAD
208. Michieletto, D., Lusic, M., Marenduzzo, D. & Orlandini, E. Physical principles of retroviral integration in the human genome. *Nat. Commun.* **10**, (2019).
209. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).
210. Pistis, G. *et al.* Rare variant genotype imputation with thousands of study-specific whole-genome sequences: Implications for cost-effective study designs. *Eur. J. Hum. Genet.* **23**, 975–983 (2015).
211. 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**, 1–14 (2018).
212. Bai, W. Y. *et al.* Genotype imputation and reference panel: A systematic evaluation on haplotype size and diversity. *Brief. Bioinform.* **21**, 1806–1817 (2020).
213. Gazal, S. *et al.* Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nat. Genet.* **50**, 1600–1607 (2018).
214. Manrai, A. K. *et al.* Genetic Misdiagnoses and the Potential for Health Disparities. *N. Engl. J. Med.* **375**, 655–665 (2016).
215. Byrska-Bishop, M. *et al.* High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *bioRxiv* (2021).

