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

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

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

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

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

ABSTRACT

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

*To whom correspondence should be addressed. Tel: +34 934134074; Email: david.torrents@bsc.es
Correspondence may also be addressed to Rafael de Cid. Tel: +34 930330542; Email: rdecid@igtp.cat
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.
‡Lead contact for data access.
Present address: Iván Galván-Femenía, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08028, Barcelona, Spain.

INTRODUCTION

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

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

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

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

MATERIALS AND METHODS

Benchmarking samples

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

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

Variant calling

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

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

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

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

Logistic regression model

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

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

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

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

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

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

Quality control

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

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

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

Variant calling, filtering and merging

Each of the 12 selected variant callers was first executed independently on all samples (Supplementary information section 7, Supplementary Figure S8), then merged by call and individual according to our benchmarking strategy to produce the VCF.

SNVs and indels calls were merged by (i) the chromosome, (ii) position and (iii) REF/ALT allele. SVs were merged by (i) variant type, (ii) chromosome, (iii) position, considering the breakpoint error estimated for each variant caller (Supplementary Table S4) and finally (iv) reciprocal overlap ≥80% between callers (Supplementary information section 8.2) and individuals (Supplementary information section 8.3). Given the consistently high accuracy in detecting SNVs for most callers, we considered one of these variants as a true positive if it was detected by at least two callers. For indels and SV, we applied LRM considering a variant as true positive if the prediction probability was ≥0.5.

We calculated the true positive proportion for each variant determined by the LRM prediction in all GCAT samples. We referred to this proportion as the quality score of the merged variant. Then, we considered a variant as PASS if the quality score was ≥0.5. We reported the length and position of each SV as the median length and median position of all the samples that have that SV (Supplementary methods). Finally, monomorphic variants, variants out of Hardy-Weinberg equilibrium (Bonferroni correction *P*-value $< 5 \times 10^{-8}$), and variants with ≥10% of missingness were excluded from subsequent analysis. Data and Code availability is described below. Summarized later at the resource availability section.

Variant validation

Comparison with public datasets. SNVs and indels from the GCAT dataset were compared with the NCBI dbSNP database (46) (Build version 153) (https://ftp.ncbi.nlm.nih.gov) to determine the number of unique/shared variants between them. GCAT SVs were compared with the following public databases: (i) The Genome Aggregation Database (gnomAD.v.2) (10) (https://gnomad.broadinstitute.org/downloads), (ii) the Database of Genomic Variants (DGV) (http://dgv.tcag.ca/dgv/app/downloads?ref=GRCh37/hg19) (47), (iii) the Human Genome Structural Variation Consortium set (HGSVC) (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/) (13), (iv) the Ira M Hall dataset (https://github.com/hall-lab/sv_paper_042020) (48), (v) the 1000G project (Phase3) (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/) (11) and (vi) GoNL (release 6.2) (on request) (12). Finally, we determine the number of shared variants between the GCAT and at least one other public dataset and the number of unique variants in the GCAT derived (Supplementary Information section 9.1.2).

We also carried out a comparison with the emerging long-read sequencing technologies. We analysed with our pipeline 30× short-read sequencing information from a 1000G sample (id: NA12878) that had been also independently sequenced and analysed using long-read technologies. We ran our variant calling and filtering strategies in this sample and matched the results obtained with those reported in Audano's study (15) (long-read sequencing) and

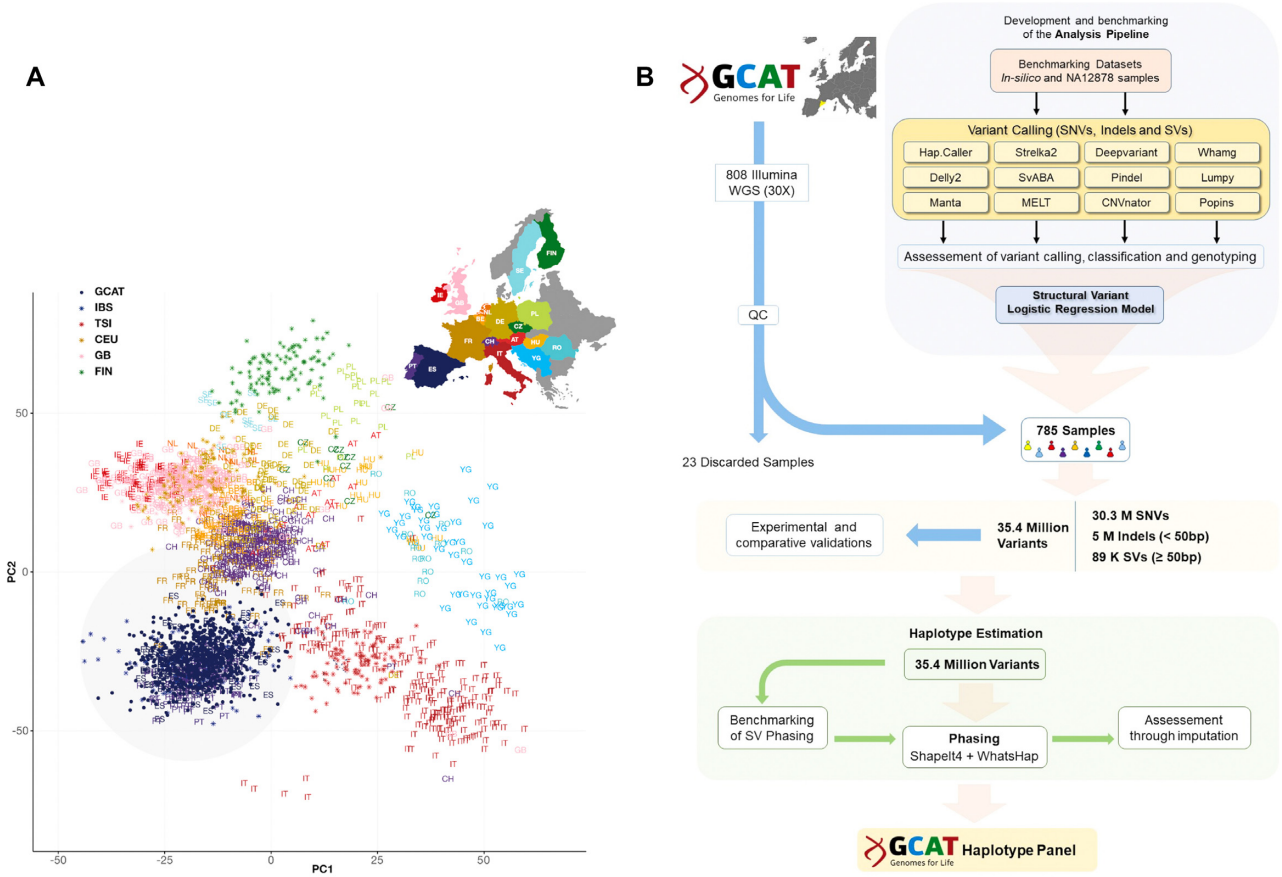


Figure 1. Overview of data and overall strategy. (**A**) Distribution of genetic data (SNVs) based on principal component analysis (PCA) (adapted from Novembre *et al.* (45)). The PC grouped by geographic localization (coloured in grey) the individuals of the GCAT cohort (blue dots) with Iberian samples from 1000G (asterisk) and POPRES (letters) projects in the context of other European samples. (**B**) Flowchart of the overall strategy followed in this study, covering from the quality control of the initial data, to the final generation of the GCAT haplotype panel, with particular focus on SVs. Overall, the complete strategy consumed ∼3.5 million CPU/hour, which highlights part of the computational challenges associated with this type of analysis (Supplementary Table S11) (See also Supplementary Figure S7).

1000G Phase 3 (3–7× coverage), obtaining the number of variants shared between projects.

Experimental validation. The validation of SNV and indel calling was performed using the SNP-array data available from 570 of the 785 individuals analysed in this study. We include QCed genotypes generated in the GCAT cohort with the Infinium Expanded Multi-Ethnic Genotyping Array (MEGAEx) (ILLUMINA, San Diego, CA, USA) as described elsewhere (18) (i.e. 732 978 SNPs and 1168 indels). Genotypes from both strategies were compared by (i) chromosome and position at base-pair resolution and (ii) REF/ALT alleles; the recall and genotype concordance for each individual sample was calculated.

Inversions were validated using a recent benchmark dataset, consisting of 59 validated human polymorphic inversions from the InvFEST project (49). Allele frequency (using CEU and TSI European populations) and length concordance was determined using an overlapping window of ±1 kb around the inversion breakpoints. Accuracy of inversion genotyping was assessed for the 785 WGS samples, using the available reference panel of experimentally-resolved genotypes (49). GCAT genotypes were imputed with IMPUTE2 (50) with a genotype posterior probability ≥0.8 and classified as missing otherwise. Missing genotypes were recovered if they had a perfect tag SNP in the reference panel ($r^2 = 1$).

Comparative genomic hybridization (CGH) method was used to validate deletions and duplications using the NA12878 sample from 1000 Genomes project as reference, for which the lists of variants had been previously described (51). For each sample, we determined gains and losses and compared them with those reported from our variant calling analysis.

Phasing and imputation performance

In order to analyse the performance of the phasing and imputation processes, all 785 GCAT samples were divided into two subsets, (i) a subset including 690 samples were first used to construct a pilot reference panel and (ii) the remaining 95 samples, with WGS and SNP-genotyping array data available, were then used as a test sample in the different analyses.

The evaluation of phasing strategies was carried out by determining the imputation accuracy of SVs, using the genotypes independently generated by WGS and imputation techniques across the 95 test GCAT samples, and with the pilot reference panel of 690 individuals (Supplementary Information section 10.1). Accuracy was determined for chromosome 22, and the quality score of imputed variants was considered as a validation proxy of the best phasing strategy. Each phasing strategy was evaluated by counting the number of variants with an info score ≥0.7, and by calculating the genotype concordance between imputed data and the calling. The phasing algorithms evaluated were ShapeIt2 (52) (version v2.r904), MVNcall (53) (version 1.0), ShapeIt4 (54) (version 4.1.3) and WhatsHap (55) (version 0.18). We used IMPUTE2 (50) (version 2.3.2) for imputation analysis (Supplementary methods).

In order to evaluate the imputation performance of the GCAT|Panel for distant ethnicities, we used the 1000G SNP-genotyping array data covering 2318 samples from 19 populations (56) (Supplementary Table S13). First, quality control was applied to the 1000G SNP-genotyping array per population by removing variants the met the following criteria: (i) ≥10% of missingness; (ii) matching A–T, C–G sites; (iii) in Hardy–Weinberg disequilibrium (P-value < 0.05); and by discarding samples with (i) ≥10% of missing, (ii) Kinship coefficient ≥0.05 and (iii) an excess of heterozygosity ±2SD, obtaining finally 1880 individuals covering 19 populations. Each population group was pre-phased with ShapeIt4 and imputed separately using IMPUTE2. Then, we compared the allele frequency, type of variant distribution, and the quality of the imputed SVs across populations.

To evaluate the imputation of SVs, we used as reference the Audano *et al.* (15) study that includes SVs identified using long-read sequencing. Imputed SVs with an info score ≥0.7 were compared considering a window of ±50 bp around the breakpoint. Furthermore, we evaluated the concordance of SV type and SV length error reported by WGS calling. On the other hand, we also evaluated the concordance of the genotype of our imputed SVs, using the SV list generated on the same samples, by Hickey *et al.* (57).

Benchmarking different panels of genetic variability

QCed genotypes generated in the GCAT cohort with the Infinium Expanded Multi-Ethnic Genotyping Array (MEGAEx) (i.e. 756 773 SNVs) were used to impute 4448 individuals (e.g. excluding those 785 with WGS) using the GCAT|Panel and the publicly available 1000G phase3 (11), GoNL-SV (12), UK10K (58) and HRC (2) reference panels. Multiple reference panel imputation was conducted using GUIDANCE (59). For comparative purposes, we considered imputed variants with info imputation score ≥0.7 and MAF >0.001. For SNVs and indels, variants were considered coincident when the position and change matched. For SVs, matching variants were considered if their positions were within a ±1 kb window, and the variant type was the same. Since allele frequency impacts imputation, we calculated the average of the info imputation score (r^2) by frequency categories: rare (MAF < 0.01), low frequency (0.01 ≤ MAF < 0.05), and common (MAF ≥ 0.05).

Functional impact of structural variants

Variant annotation. Functional, regulatory, and clinical annotations of SVs were predicted using AnnotSV (60). The functional impact of SVs was evaluated by considering (i) the level of overlap with known genes, (ii) the level of overlap with regulatory regions (61), (iii) the predicted loss of function intolerance (pLI) effect and (iv) the reported disease association studies. In addition, we used SVFX (62), a mechanism-agnostic machine learning-based workflow, to evaluate the potential pathogenicity of large deletions and duplications (>50 bp), in four major cardiometabolic conditions from the GCAT cohort; diabetes, obesity, cardiovascular diseases, and hypertension. SVs were classified using the annotations of the SVFX tool into pathogenic (SV pathogenic score ≥ 0.9) or benign (SV pathogenic score ≤ 0.2). Finally, SNVs and indels (up to 50 bp) were annotated using SnpEff (63) and SnpSif (64) (v5.0e) tools, covering LoF and pathogenicity descriptors from ClinVar (65) and CADD (66) resources.

Comparison with the GWAS catalog. GWAS catalog version 1.0.2 (r2021-05-05) was downloaded from https://www.ebi.ac.uk/gwas/docs/file-downloads. First, we selected 106 906 variant-phenotype associations of 72 849 unique autosomal entries identified in European ancestry. Second, we intersected with PLINK2.0 (67) 68 323 unique variant-phenotype associations (MAF > 0.01) with the GCAT dataset (∼30M variants) by breakpoint coordinates. Finally, we identified 1374 unique SVs (MAF > 0.01) in strong linkage disequilibrium (r^2 > 0.80) with variant-phenotype associations in 1Mb window (Supplementary Figure S27). From these 1374 SVs, we evaluated the SV type, as well as the overlap with genes and regulatory regions.

Genome-wide association analysis. Association analysis was performed by 70 independent GWAS of chronic conditions. Phenotype selection was derived from the Electronic Health Records registry from the cohort (2012–2017) and chronicity was defined using public guidelines for chronic condition definitions (68), and the Chronic Condition Indicator (CCI) (http://www.hcup-us.ahrq.gov/toolssoftware/chronic/chronic.jsp) (69,70), then grouped considering ICD-9 codes and chapter descriptions. Conditions with more than 50 cases were retained for the GWAS analysis (i.e. 70). Each association test was performed as independent logistic regression for each cohort, under the assumption of an additive model for allelic effects, with adjustments made for age, sex and the first five principal components. Gender-specific conditions were analysed only for a specific gender. The analysis was performed using PLINK2.0 (67) for autosomal chromosomes. A Bonferroni correction accounting for the 10 ICD-9 categories used (i.e. body systems) was applied. Locus Zoom was derived for specific regions, and suggestive tower profiles were analysed, based on LD patterns and gene-centered impact.

Experimental validation of the Alu element. PCR amplicon analysis was designed using Primer 3.0 software using the hg19_dna range = chr3:49 492 813–49 496 062 sequence, including the Alu element. Sequence primers are for F-primer (5′CATTGACTCATTCAGCAAGCA 3′) and

for R-primer (5′AAATTAAGCCCCACCCTAG3′). Using standard conditions (35×, $T_m = 60^oC$) in a Veriti™ 96-Well Thermal Cycler (Thermo Fisher Scientific), we obtain a 515 bp fragment corresponding to the control-allele and an 848 bp one for the Alu-allele. Fragments were resolved by e-agarose gel, in a TapeStation (Agilent). Further, the amplicon of a non-ALU allele carrier was analysed by Sanger Sequence Method to verify the insertion point (i.e. at hg19 Chr3:49 494 280) and the ALU sequence insertion (324 bp).

Statistical analyses

R software was used for data visualization and statistical analyses. 95% confidence intervals (CI) for recall, precision, and genotype error metrics were assessed as point estimation ±1.96SD. Risk ratios with 95% CI and two-tailed *P*-values from the functional enrichment of common and rare SVs were calculated using the risk ratio function from the epitools R package. Pearson correlation coefficient with 95% CI and two-tailed *P*-value were estimated using the cor.test() function implemented in R.

RESULTS

Evaluation of cohort data quality and consistency

From the GCAT cohort (17) we randomly selected 808 individuals (gender-balanced) for new Illumina whole-genome sequencing at 30× coverage. Twenty three samples were excluded based on sequence quality, ethnicity, and relatedness parameters (see Methods, Supplementary Table S10). Principal component analysis (PCA) on the remaining 785 individuals identified a unique and separated cluster compared with neighbouring populations (Figure 1A, Supplementary Figure S7), in agreement with their geographic origin (45).

Generation of a comprehensive variant identification strategy

We designed, benchmarked, and implemented a comprehensive strategy for capturing, classifying, and phasing a wide range of germline variants from short-read Illumina sequences, with particular efforts devoted to the identification and subclassification of larger structural variants (Figure 1B). Using sequencing data from an *in-silico* genome (Supplementary information 1, Supplementary Table S2), and a real sample (NA12878, from the Genome In A Bottle (GIAB) project (24)), we assessed the performance (i.e. recall, precision and *F*-score metrics) of 17 variant callers covering SNVs, small indels (<50 bp), and large SVs (≥50 bp) (see Materials and Methods), and retained the best twelve (Supplementary Table S3). SNVs were first filtered based on a minimum constraint of having the support from at least two callers, which provided high recall (>95%) and precision (>96%) values. On the other hand, for the filtering of small indels and SVs, which show high levels of discrepancy across individual callers and their combinations (Figure 2A), we built a Logistic Regression Model (LRM), to prioritize caller results through a reliability score from the weighted combination of different calling parameters (Figure 2B, Supplementary Figure S2) (see Materials and Methods), accordingly higher *F*-scores correlated with larger AUC values (Supplementary Figure S30). This approach outperformed other typical curation strategies over the entire spectrum of SV sizes (Figure 2C, Supplementary Figure S5). Furthermore, because accurate genotype calling is also key for downstream analyses, on top of this LRM, we prioritized those callers that best resolved the heterozygosity (i.e. genotypes) (see Materials and Methods), resulting in a lower rate of genotype error (<6%) across all variant types when compared to the *in-silico* sample (Figure 2D, Supplementary Figure S3).

Genome-wide variation analysis of the GCAT cohort

The application of this strategy to the selected 785 whole-genome Illumina sequences (30×), let us identify 35 431 441 unique variants across the cohort. Of these, 85.6% correspond to SNV, 14.1% to indels (<50 bp) and 0.3% ($n = 89\,178$) to SVs (≥50 bp) (Figure 3A). Median values of variants per individual were 3.52M SNVs (SD = 24 983), 606 336 indels (SD = 8060) and 6393 SVs (SD = 222), showing good consistency across the cohort (Figure 3B). SV sizes ranged from 50 bp to 197MB (duplication), with median values of 291 bp and a different distribution for each type of variation (Figure 3C), affecting globally a median of 7% of the entire genome per individual. Frequency ranges across all SVs were in agreement with other public WGS-based studies (Figure 3D), with 31% of them being common or low-frequency (MAF ≥ 0.01), and 69% being rare (MAF < 0.01), including a large fraction (50%) present only in one or two individuals (i.e. MAF ≤ 0.0025).

The robustness of these results was evaluated using comparative and experimental approaches. A large fraction of SNVs and indels (i.e. >79% and >93% respectively) matched with dbSNP (Build 153.v) (46) entries (Supplementary Figure S9a, b). Regarding SVs, the comparison against different public databases (i.e. gnomAD-SV (10), 1000G (11), GoNL (12), HGSVC (13), DGV (47), dbVar (47), Ira M. Hall Lab dataset (48); see Materials and Methods) highlighted 49,333 novel SVs (i.e. 61% of all SVs), of which 27% were present in more than two individuals (Supplementary Figure S9). As to the type, 26% of these novel variants correspond to deletions, 8% to duplications, 20% to insertions, 20% to inversions, 4% to LINEs, 1% to SVAs, and 21% to Alu elements. The comparison of our results with array-based genotypes in a fraction of our cohort ($n = 570$ individuals) validated 96% and 87% of SNVs and indels, respectively, with a genotype concordance of 97% and 96% (Supplementary Figure S10). Furthermore, we also used a benchmarking set of 59 manually-curated and experimentally-genotyped inversions with MAF >0.01 from the InvFEST project (49) to evaluate this type of variants within our catalogue. Of these 59 inversions, we detected 51 (86%), with concordant size and allele frequency values (Supplementary Figure S11a, b; see Materials and Methods). This validates ∼38 000 of ∼40 000 independent inversion calls across the entire cohort, with an average genotype concordance of 95% (Supplementary Figure S11c). In addition, we have applied CGH, which best targets duplications, as well as large deletions (>20 kb). Using this

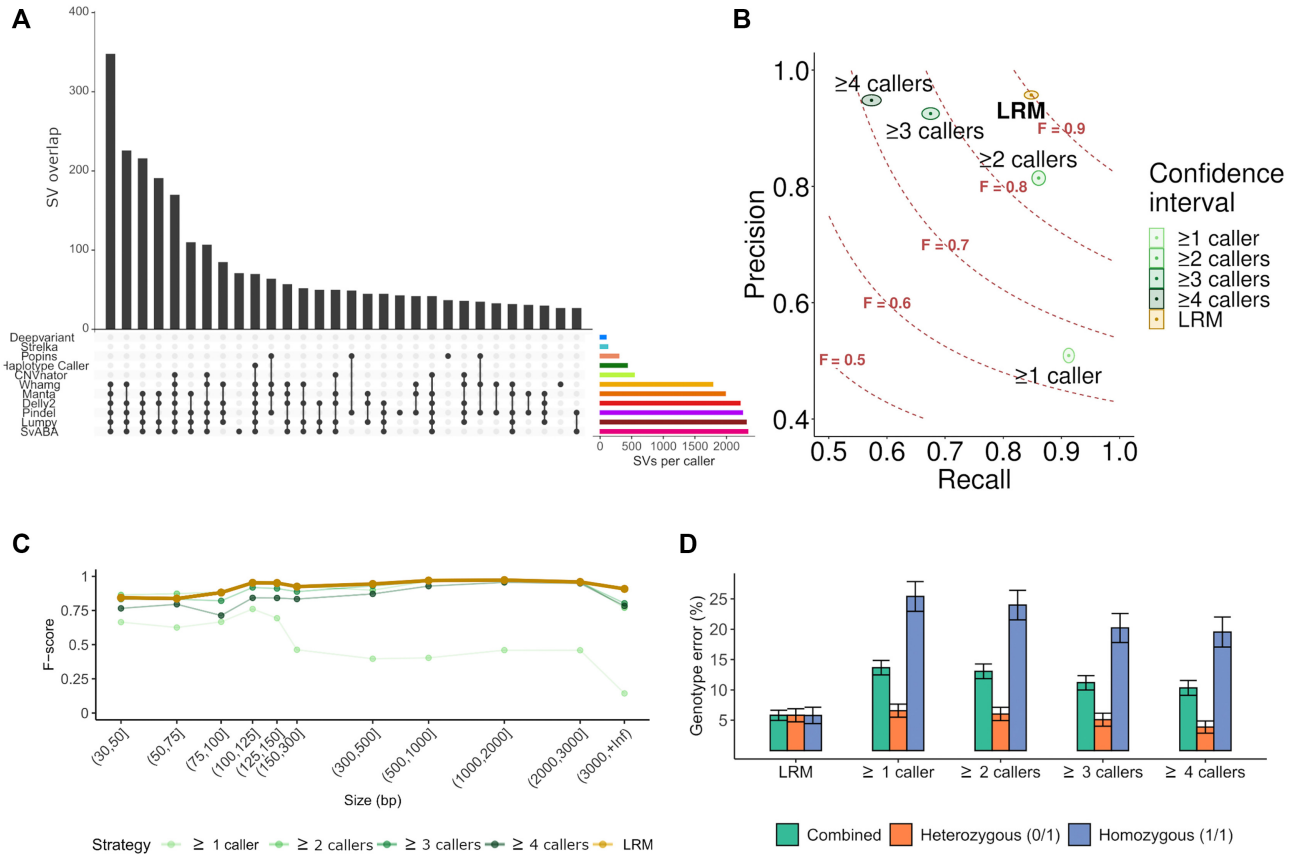


Figure 2. Benchmarking of the structural variant identification and classification pipeline. (**A**) Structural variant (SV) detection patterns according to the programs used. Lines and dots indicate the programs used and bars the number of overlapping calls resulting from that combination. The first 30 patterns with more coincident SV calling are shown. Right coloured horizontal bars indicate the total number of SVs detected by each caller. Variant callers that detect all SV types and sizes tend to recover more SVs than those that detect specific SV types (i.e. CNVnator) and smaller SVs (i.e. Strelka2). (**B**) Overview of the detection performance of different strategies and filtering results from multiple variant callers. Each strategy is plotted according to the recall and precision ratios ($F = F$-score) using the benchmarking dataset. The logistic regression model (LRM), with a F-score of 0.9, outperformed other commonly used strategies that are based on the number of coincident callers (logical rules). The confidence interval for each case is represented by coloured area of each strategy. (**C**) Comparison of performances (F-score) of different merging and filtering strategies according to the size of the structural variant. (**D**) Comparative overview of the genotype error, associated to each strategy for each allelic state. Error values and their intervals were inferred from the benchmarking dataset (see supplementary Figures S2, S3 and S5 for the information across the different SV types).

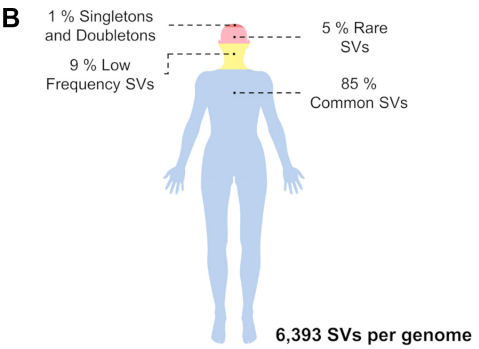
technique, we could validate 76% of our deletions, as well as 20% of the duplications (Supplementary Table S16). Finally, we contextualized our results in the frame of other SV identification efforts, through the analysis of the NA12878 sample from the 1000G project that has been sequenced and analysed using long and short read technologies at different coverages. From all SVs identified with long-read technology (15), our strategy was able to identify 24% of them when applied to NA12878 at 30× short-read sequence. This overlap is different across different SV types, as we detected 14% of the insertions and duplications, but up to 48 and 57% of the inversions and deletions, respectively. The same comparison using the 1000G annotation of NA12878 at 3–7× coverage showed a coincidence with long-read results of 4, 2 and 0.1% for deletions, inversions, and duplications respectively (Supplementary Table S17), showing a significant detection improvement when using higher coverages, identifying between a 2- and 7-fold the number of variants with 30× coverage, compared with 15× and 5× coverages, respectively (Supplementary Figure S31).

Predicted functional impact of SVs

A first assessment of the potential functional impact and pathogenicity of our SVs was obtained using AnnotSV (60). 46% of all SVs overlapped with genes, affecting a median of 2868 per individual, whereas 18% overlapped with gene regulatory regions (see Data and Code Availability at the resource availability section for the corresponding gene lists). While the majority (88%) of gene-overlapping SVs mapped within intronic regions (Supplementary Figure S24a), 9% of them affected coding sequencing regions (CDS). In agreement with known variant fixation patterns within populations, we observed that rare SVs (MAF < 0.01) tend to be more disruptive, compared to common variants (MAF ≥ 0.05), as 13% of rare SVs are overlapping coding regions, compared to 5% of the common ones (RR $= 0.13/0.054 = 2.4$, 95% CI $= [2.14,2.69]$, P-value $= 2.6 \times 10^{-67}$, Supplementary Table S15a, b). Of the affected genes, 28% (10 600 SVs) are related to disease, as indicated by the predicted loss-of-function intolerance parameter (pLI) (71) (Supplementary Figures S25a, S26 and

A

Variant type	Number of identified variants	Number of accepted variants
SNVs	58,529,907	30,325,064
Indels (1-49 bp)	10,452,204	5,017,199
Deletions	1,359,594	33,244
Duplications	674,817	6,269
Insertions	352,939	12,782
Inversions	228,091	10,115
MEIs	170,735	18,779
Translocations	117,048	7,989



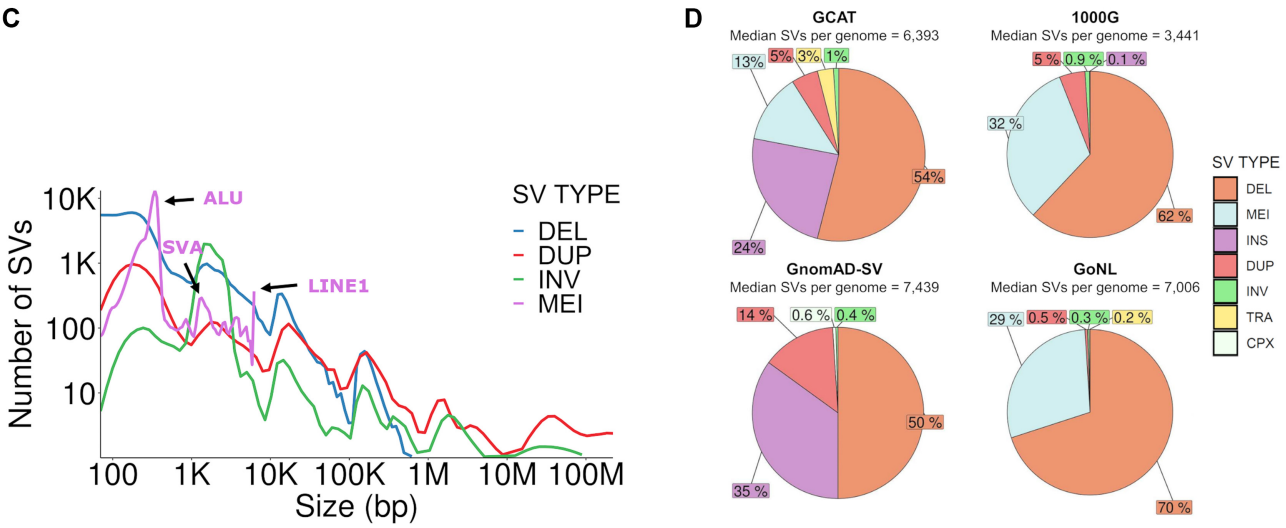


Figure 3. Overview of the GCAT variant catalogue. (**A**) Table with the numbers of identified and accepted variants after applying the filters 'at least two callers detecting the same variant' for SNVs, the LRM for indels and SVs, Hardy–Weinberg equilibrium, and discard monomorphic variants and those with >10% missingness within the GCAT cohort, according to their class. (**B**) Overview of the variant distribution within an average individual in the GCAT cohort, according to their observed minor allele frequency (MAF). (**C**) Distribution of SV type according to their genomic sizes. (**D**) Comparative overview of the SV type number and distribution across the GCAT, 1000G, GnomAD and GoNL catalogues.

Table S14). Additionally, when we analysed the putative causal role of our SVs variants across multiple phenotypes, we observed that 1374 SVs (MAF $\geq$ 0.01) are in strong linkage disequilibrium (LD) ($r^2 \geq 0.8$) with loci associated with human traits from the GWAS Catalog (version1.0.2 release 05/05/2021), tagging mainly deletions (Supplementary Figure S27), with more than half of them (799) directly overlapping genes or regulatory regions. Finally, we further refined these results with annotations from the SVFX tool (62) for four major cardiometabolic conditions; obesity, cardiovascular traits, hypertension, and diabetes. Our analysis identified 106 GWAS catalog (P-value $< 10^{-8}$) hits (i.e. 8% of total hits) that overlap with pathogenic annotated variants in the four analysed traits; 55% variants overlap with obesity and related obesity traits, 20% with diabetes, 16% with cardiovascular-related diseases and 9% with Hypertension and related traits. Of these, 95% were common and 5% were low-frequency variants. We observed a ratio of pathogenic to benign deletions of 0.95, 1.93, 1.85 and 0.40 for diabetes, hypertension, obesity, and cardiovascular traits, respectively. In the case of duplications, these ratios were 2.06, 4.42, 4.06 and 0.82 for diabetes, hypertension, obesity, and cardiovascular traits, respectively, suggesting that duplications are twice more likely to be involved in these traits.

From the annotation obtained using SnpEff (63) we extracted 2855 variants that were classified as LoF and obtained their pathogenicity using ClinVar (65) and CADD (66) data. ClinVar data was available for 243 variants 70 of which were reported as pathogenic or likely pathogenic, and CADD data was available for 2850 variants, 2330 of which were classified as deleterious (CADD PHRED score > 20).

Iberian Haplotypes estimation

As a resource for the enrichment of SVs within genome-wide association studies, we built a haplotype reference panel by phasing together all the variants identified within all GCAT samples. We first generated a cross-validation framework to identify the best available phasing strategy for SV (see Materials and Methods), using downstream imputation results as the evaluation and ranking criteria (Supplementary Figures S12 and S13 and Table S12). In our hands, the combination of ShapeIt4 (54) and WhatsHap (55), which include phase informative reads (PIRs),

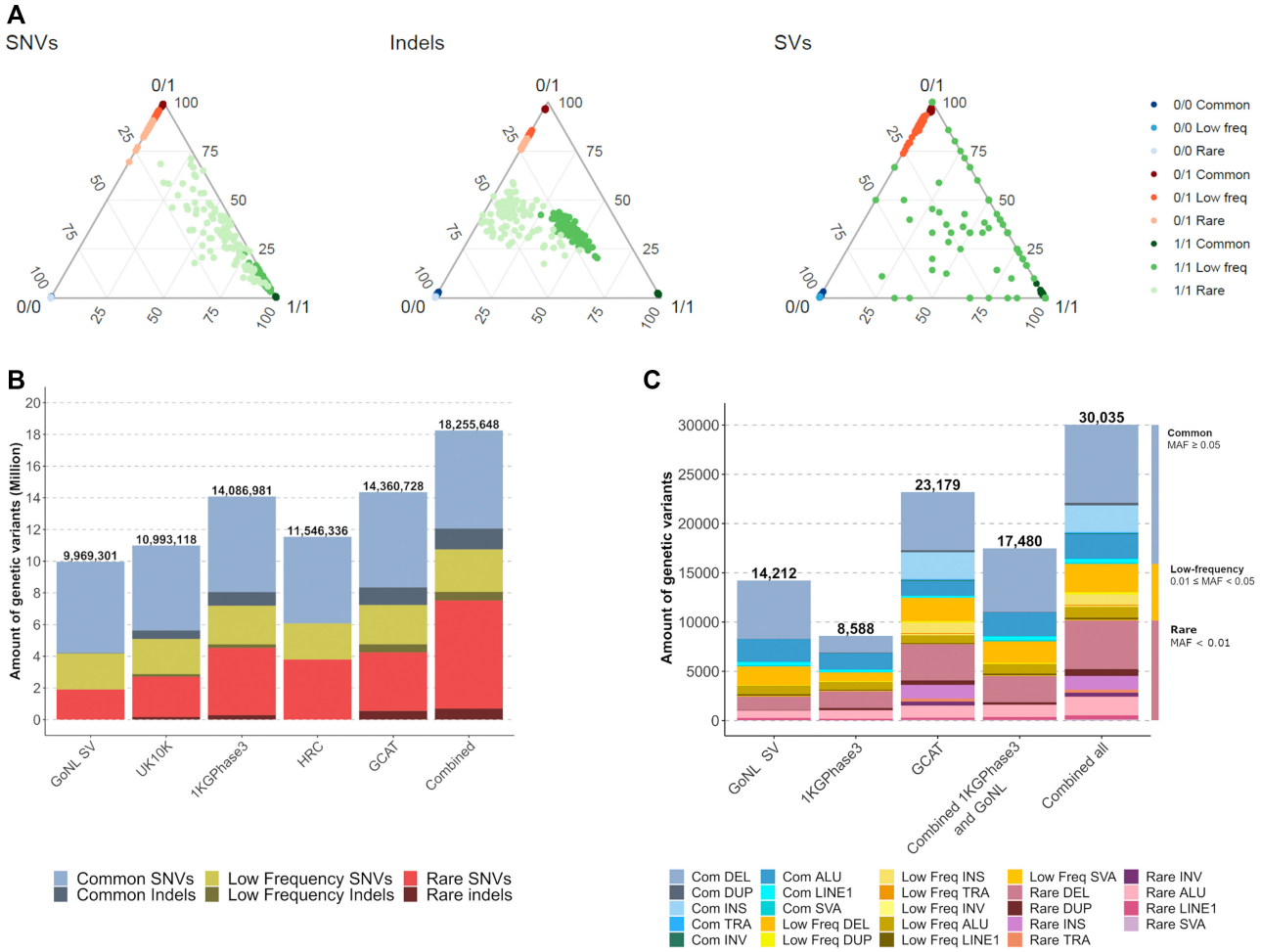


Figure 4. Phasing and Imputation performance of the GCAT|Panel. (**A**) Ternary diagram of the genotype imputation accuracy by variant type and frequency, considering the genotype calling as reference. Three dots evaluate each genotype state per sample. The samples with high concordances between genotype imputation and genotype calling were located at ternary diagram vertices. (**B**) Number of SNVs and indels imputed (info score ≥ 0.7) using different reference panels and combining their imputation results. More indels were recovered by GCAT|Panel. (**C**) Number of SVs imputed (info score ≥ 0.7) using different panels, and combining the imputation results with and without GCAT|Panel. (See also Supplementary Figure S21).

provided the best results. Using this protocol (Supplementary Figure S14), the resulting haplotype panel allowed the imputation (info scores > 0.7) of 98%, 92%, and 90% of our common SNVs, indels and SVs, respectively, recovering a median of 5120 SVs (SD = 50), from a maximum of 6393 SVs estimated per individual. While the best imputation results came from *de novo* insertions and deletions, with 96% and 95% recovery rates, respectively, duplications and translocations were imputed at lower rates, i.e. 48% and 19%, respectively (Supplementary Figure S15). Overall we imputed common SNVs, indels and SVs with a genotyping concordance of 99% (SD = 0.4), 97% (SD = 0.6) and 98% (SD = 1.2) (Figure 4A), respectively. The lowest values were observed for duplications and translocations, with genotype concordances of 84% (SD = 9.2) and 73% (SD = 27.6), respectively (Supplementary Figure S16).

As the possibilities of accurately imputing SVs are expected to correlate with the number of neighbouring SNVs and indels in LD, we next analysed the variation context of our SVs. Using one megabase window, we observed that the number of SNVs and indels in strong LD ($r^2 \geq 0.8$) with common deletions, insertions, inversions, and mobile element insertions (MEIs) was in the range of 39–42, in contrast to duplications and translocations, which showed mean values of 12 and 8 variants respectively (Supplementary Figure S17a). In fact, as expected, a positive significant correlation was observed between the number of variants in LD and the score of imputation for common SVs (Pearson's $r = 0.38$, 95% CI = [0.37, 0.40], P-value $< 2 \times 10^{-16}$) (Supplementary Figure S17b), and for all SV types (except translocations) (Supplementary Figure S18).

Imputation performance of the haplotype panel

Following this strategy, we generated a complete and operational panel of Iberian haplotypes, with all the variants of our 785 individuals. To assess the value and benefits of the resulting GCAT|Panel, as an imputation resource for enriching genetic association studies with SVs, we first imputed the genotyping array data of 4448 GCAT individuals and compared the results with those of other reference panels, such as 1000G (11), GoNL (12), HRC (2), and UK10K

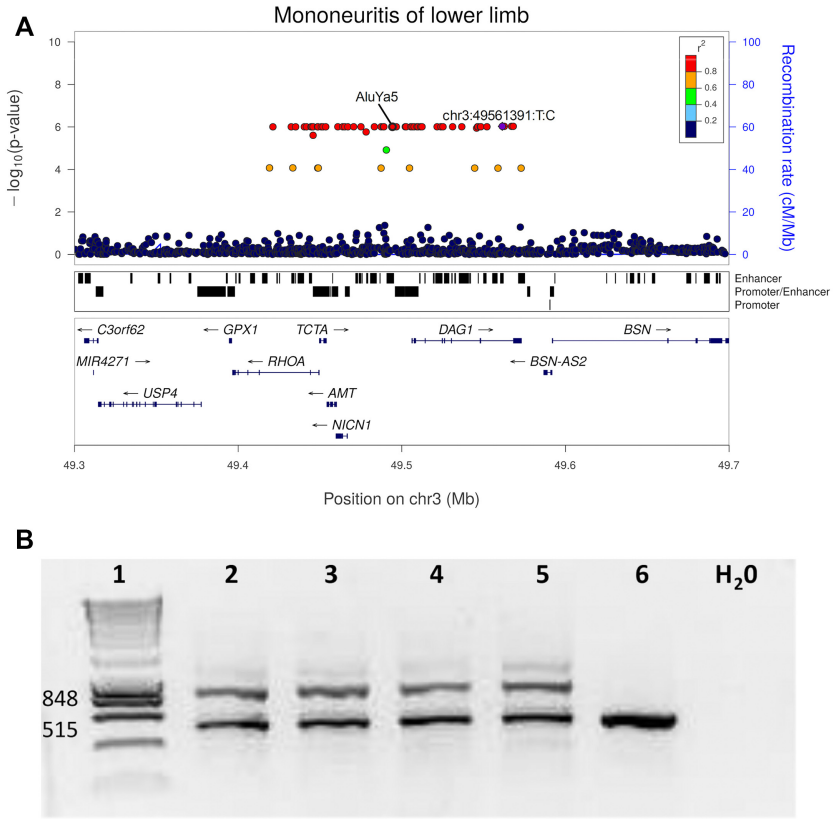


Figure 5. Genome-wide association analysis using GCAT|Panel and experimental validation of an AluYa5-element. (**A**) Locus zoom plot of the locus associated with mononeuritis of lower limb (ICD-9 355) (P-value = 9.84×10^{-7}), showing the lead variant in purple. The AluYa5-element (g.49494276_49494600ins (hs37d5) maps in an enhancer element upstream of the *DAG1*. (**B**) Experimental validation of an AluYa5-element, agarose e-gel electrophoresis of PCR products after amplification of Alu-insertion-specific DNA fragments from blood DNA Lanes: 1, 100 bp DNA ladder marker (Life Technologies), expected sizes of both states are shown to the left; 2–5 Alu carriers (EGA_04200, EGA_01901, EGA_13378, EGA_03940); six control individual (EGA_01399). The numbers to the left refer to the size (bp) of marker DNA fragments. Electrophoresis analysis of Alu carriers show two-band amplicons (515 bp and 848 bp) detected in Alu carriers (lanes 2–5) and one-band amplicon (515 bp) in control non-Alu-allele individuals (lane 6) (See also Supplementary Figure S29).

(58). With IMPUTE2 (50), the GCAT|Panel was able to impute a total of 14 383 907 variants with MAF > 0.001 and high quality (info score ≥ 0.7). Across different reference panels, the overall imputation performance for SNVs and indels (<50 bp) was generally high (Figure 4B), with slight overperformances of the GCAT|Panel on indels, and of 1000G and HRC panels on SNVs. While HRC and 1000G recovered rarer SNVs, likely because of their larger sample sizes, the GCAT|Panel was able to recover rarer indels (Figure 4B). At the structural variation level, the GCAT|Panel was able to impute a total of 23, 179 SVs with info scores ≥0.7, resulting in a 1.6-, 2.7- and 1.3-fold increase, compared with the 1000G, the GoNL, and both panels combined, respectively (Figure 4C). For common SNVs/Indels (MAF > 0.05) the GCAT|Panel showed similar performance as HRC, 1000G, GoNL and UK10K reference panels (mean $r^2 > 0.96$, Supplementary Figure S21a). For common SVs, the GCAT|Panel outperformed (mean $r^2 = 0.91$, SD = 0.15) 1000G (mean $r^2 = 0.80$, SD = 0.21) and GoNL-SV reference panels (mean $r^2 = 0.82$, SD = 0.21, Kruskal–Wallis P-value $< 2.2 \times 10^{-16}$, Supplementary Figure S21b).

In an exploratory analysis, structural variants imputed by the GCAT|Panel were also tested (together with SNV and indels) for association across 70 identified chronic conditions within the cohort. Conservatively, only structural variants with an info score >0.7 and conditions with >50 cases were included in this analysis. Forty six SV loci showed suggestive association with 26 conditions after Bonferroni correction (P-value $\leq 1 \times 10^{-6}$) (Supplementary Figure S28). Of all these associations, 63% could potentially be functionally explained through SVs, as they either lead the association (37%) or are in strong LD ($r^2 \geq 0.8$) with the lead variant (26%). A notable example is a rare AluYa5-element in chr3 (g.49494276_49494600ins (hs37d5), MAF = 0.0013), located near the dystroglycan gene (*DAG1*) and associated (P-value = 9.84×10^{-7}) with Mononeuritis of lower limb (ICD-9 355) (Figure 5A). The presence of this Alu element, imputed only with the GCAT|Panel (info score = 0.98), was experimentally confirmed in all carrier individuals (Figure 5B, Supplementary Figure S29).

Finally, we evaluated the portability of the GCAT|Panel to infer SVs across 19 different ethnic groups using 1880 individuals from the 1000G project. While the imputation

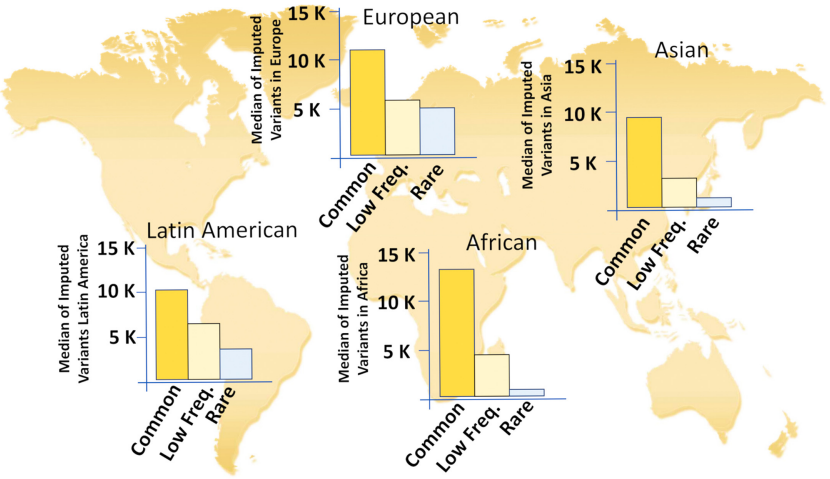


Figure 6. Structural Variant imputation performance using GCAT|Panel across all continents. European and Latin American populations recover more low frequency and rare SVs at high info scores (≥0.7) than African and Asian populations (see also Supplementary Figures S22 and S23).

quality of SVs was higher within the European populations (Supplementary Figure S22), the GCAT|Panel was also able to impute a large fraction of SVs across all other ethnicities (Figure 6, Supplementary Figure S23a). Of nearly 50K unique SVs imputed across all groups, 25%, 35% and 40% of them were detected within the Asian, African and Latin American populations, respectively (Figure 6, Supplementary Figure S23). In agreement with the mixed origin of Latin Americans, nearly half of all imputed variants within this group showed low-frequency values (MAF < 0.05), compared with other non-European groups, where the imputation covered predominantly common variants (Figure 6). In addition, 73% of all the structural variants identified and genotyped in previous studies, using long and short WGS (15,57) were also imputed by our panel on the same individuals, with 88% of matching genotypes (Supplementary Figures S19b and S20a).

DISCUSSION

Here, we present the GCAT|Panel, the first Iberian Haplotype reference panel derived from high-coverage whole-genome sequencing. The strategy developed for variant identification, classification, and phasing, has provided a comprehensive and high-quality catalogue of genetic variants, with low rates of false-positive calls and genotyping errors for all variant types, including SVs. This is due to the combination of high sequencing coverage (30×) with a comprehensive analysis strategy that integrates multiple variant callers and a Logistic Regression Model for maximzing recall and precision for each SV type and size.

Increasing the sequencing coverage to 30× allowed us to resolve a large fraction of SVs and accurately define the genotypes that cannot be properly defined with lower sequencing depths. In addition, while previous projects inferred SVs into phased haplotype scaffolds (11,12), our sequencing coverage allows us, for the first time, to phase SVs together with biallelic SNVs and indels, and to use phase informative reads (PIRs), which are expected to improve the imputation of rare variants (72). With this sequencing technology, we also expect a slight detection bias against low complexity (repeated) regions of the genome, where short-read sequencing tends to be less informative, in contrast to long-read sequencing technology (13–16). This is further highlighted by the high portion (54%) of our SVs affecting genes or regulatory regions, which also tend to be within the non-repetitive portion of the genome.

Given the increasing incorporation of whole-genome sequencing into genetic studies, it is crucial to highlight the importance of accurately identifying and resolving SVs with the correct genotype, to then obtain robust and meaningful results during the imputation in a different cohort. Here, we found a positive correlation between the number of neighbouring variants in LD with SVs and their quality of imputation, suggesting that variants with a high genotyping error show a lower number of variants in LD, which translates into a lower imputation accuracy for those variants (Supplementary Figure S17). On the other hand, software limitations (PLINK or ShapeIt4), can translate into poor estimations of haplotypes and LD, directly hampering the association test, which relies on accurate counts of variant allele frequencies and states. Improved variant calling strategies that can accurately identify and define complex structural variation events are still needed, together with new and dedicated analysis frames (e.g. phasing and LD) for SVs, where the actual size and type of the variant is considered, in contrast to the current scenario where SVs are taken as SNVs.

In our cohort, the GCAT|Panel led to the identification of potential risk SV, including those within the rare spectrum. Here, we highlight the identification of a rare polymorphic 324 bp-long AluYa5 element in chromosome 3 (g.49494276_49494600, MAF = 0.0013) associated with Mononeuritis of the lower limb (ICD-9 355). This SV is located within a multi enhancer-elite element (GeneCards) (73), proximal to *DAG1*, a gene involved in pathways responsible for neuromuscular diseases, and already causing severe limb-girdle muscular dystrophy type 2P (LGMD2P) through missense point mutations (74). Further studies are now needed to validate the resulting hypothesis, in which

this Alu element could be affecting the expression of the *DAG1* gene in this disease.

This study also provides detailed guidance for the comprehensive analysis of whole-genome sequences, including the identification, classification, and phasing of SVs. We expect that this type of analysis will soon become the standard within large genetic studies that are already incorporating whole-genome Illumina sequences and combining them with existing genotyping array information.

Taken together, the availability of a high-quality haplotype panel, including a comprehensive fraction of structural variability, will significantly impact evolutionary and biomedical studies at different levels. The possibility of enriching current genome-wide association studies (e.g. GWAS and eQTL) with SVs through imputation, directly increases the chances of variant discovery, as well as of their functional interpretations. Our analysis evidence the potential of using population-matched reference panels, for the identification of rare structural variants involved in disease, and the important contribution to the understanding of the underlying genomic architecture of genetic diseases.

RESOURCE AVAILABILITY

Below we attach the information of the data and code availability used in this study.

DATA AVAILABILITY

The data generated in this study, including the FASTQ, BAM and VCF files of the 808 individuals with their genotyping information, as well as the entire GCAT|Panel, are accessible upon request (rdecid@igtp.cat) from the European Genome-phenome Archive (EGA), under the accession number EGAS00001003018. All the GCAT catalogue variants, the SV (Figure 3A), SNVs and indels annotations files, and the *in-silico* information (i.e. FASTQ, BAM files, catalogue of variants inserted) are available at http://cg.bsc.es/GCAT_BSC_iberianpanel.

All original code has been deposited at (https://github.com/gcatbiobank/GCAT_panel) and is publicly available as of the date of publication. DOIs are listed in the key resources table.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Lluís Puig and Vanessa Plegezuelos on behalf of Blood and Tissue Bank from Catalonia (BST), who collaborated in GCAT recruitment, and all the GCAT volunteers that participated in the study. We also thank the Centro Nacional de Análisis Genómico (CNAG-CRG), who collaborated in the sequence analysis of the study; members of the Comparative and Functional Genomics Group at the UAB for helping with the inversion validation; Dr. Francesc Calafell for his comments on the manuscript; Marta Morell from qGenomics for the technical assistance in the Alu validation; the Computational Genomics Group at the BSC, specially Ignasi Morán and Lorena Alonso, for their helpful discussions and valuable comments on the manuscript; and the technical support group from the Barcelona Supercomputing Center Center. We acknowledge Red Española de Supercomputación (RES) for awarding us access to MareNostrum4 supercomputer from Barcelona Supercomputing Center (proposal numbers BCV-2018-3-0010 and BCV-2019-1-0015). Figure 1a has been Adapted/Translated by permission from Springer Nature: Nature Genes mirror geography within Europe, John Novembre *et al.*, August 31, 2008.

Author contributions: J.V.M., I.G.F., D.M.S., D.T. and R.dC. designed and planned the whole study. J.V.M., I.G.F., D.M.S., D.T. and R.dC. contributed to the writing and editing of the manuscript. A.C. prepared and QCed samples for NGS. J.V.M. created the *in-silico* and D.M.S. prepared the NA12878 sample. J.V.M., I.G.F. and D.M.S. performed the variant calling and designed the Logistic Regression Models. J.V.M., I.G.F., D.M.S. contributed to the creation of BAM files, quality control, variant merging, filtering and genotyping, with the collaboration of M.P. J.V.M., I.G.F. and D.M.S. performed a comprehensive comparative analysis of the Iberian catalogue with other repositories. J.V.M., I.G.F. and D.M.S. conducted the SNV, indel and large deletion and duplication validations, collaborating with L.A. on behalf of qGenomics conducting CGH validation analysis. The inversion validation and genotype data analysis was provided by J.L.J., M.Pu and M.C. J.V.M., I.G.F. and D.M.S. performed the SV annotation, the creation of the GCAT|Panel, and their benchmarking, in collaboration with C.S. R.A. executed and adapted GUIDANCE with ShapeIt4 and GCAT|Panel. I.G.F., N.B., X.F., B.C. conducted and analysed the Phenome analysis, from phenotype extraction to GWAS analysis. N.B., R.dC., X.F. and L.A. conducted the AluY5a validation. L.S., J.F.S.H. conducted SNP-Array validation analysis of SV. V.M. and M.Pe contributed to the editing of the manuscript. D.T. and R.dC. supervised the study. All authors reviewed and approved the manuscript. J.V.M., I.G.F. and D.M.S. contributed equally to this study.

FUNDING

GCAT|Genomes for Life, a cohort study of the Genomes of Catalonia, Fundació Institut Germans Trias i Pujol (IGTP); IGTP is part of the CERCA Program/Generalitat de Catalunya; GCAT is supported by Acción de Dinamización del ISCIII-MINECO; Ministry of Health of the Generalitat of Catalunya [ADE 10/00026]; Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) [2017-SGR 529]; B.C. is supported by national grants [PI18/01512]; X.F. is supported by VEIS project [001-P-001647] (co-funded by European Regional Development Fund (ERDF), 'A way to build Europe'); a full list of the investigators who contributed to the generation of the GCAT data is available from www.genomesforlife.com/; Severo Ochoa Program, awarded by the Spanish Government [SEV-2011-00067 and SEV2015-0493]; Spanish Ministry of Science [TIN2015-65316-P]; Innovation and by the Generalitat de Catalunya [2014-SGR-1051 to D.T.]; Agencia Estatal de Investigación (AEI, Spain) [BFU2016-77244-

R and PID2019-107836RB-I00]; European Regional Development Fund (FEDER, EU) (to M.C.); Spanish Ministry of Science and Innovation [FPI BES-2016-0077344 to J.V.M.]; C.S. received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement [H2020-MSCA-COFUND-2016-754433]; this study made use of data generated by the UK10K Consortium from UK10K COHORT IMPUTATION [EGAS00001000713]; formal agreement with the Barcelona Supercomputing Center (BSC); this study made use of data generated by the Genome of the Netherlands' project, which is funded by the Netherlands Organization for Scientific Research [184021007], allowing us to use the GoNL reference panel containing SVs, upon request (GoNL Data Access request 2019203); this study also used data generated by the Haplotype Reference Consortium (HRC) accessed through the European Genome-phenome Archive with the accession numbers EGAD00001002729; formal agreement of the Barcelona Supercomputing Center (BSC) with WTSI; this study made use of data generated by the 1000 Genomes (1000G), accessed through the FTP portal (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/); this study used the GeneHancer-for-AnnotSV dump for GeneCards Suite Version 4.14, through a formal agreement between the BSC and the Weizmann Institute of Science. Funding for open access charge: GCAT|Genomes for Life, a cohort study of the Genomes of Catalonia, Fundació Institut Germans Trias i Pujol (IGTP); IGTP is part of the CERCA Program/Generalitat de Catalunya; GCAT is supported by Acción de Dinamización del ISCIII-MINECO; Ministry of Health of the Generalitat of Catalunya [ADE 10/00026]; Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) [2017-SGR 529]; B.C. is supported by national grants [PI18/01512]; X.F. is supported by VEIS project [001-P-001647] (co-funded by European Regional Development Fund (ERDF), 'A way to build Europe'); a full list of the investigators who contributed to the generation of the GCAT data is available from www.genomesforlife.com/; Severo Ochoa Program, awarded by the Spanish Government [SEV-2011-00067 and SEV2015-0493]; Spanish Ministry of Science [TIN2015-65316-P]; Innovation and by the Generalitat de Catalunya [2014-SGR-1051 to D.T.]; [Agencia Estatal de Investigación (AEI, Spain) [BFU2016-77244-R and PID2019-107836RB-I00]; European Regional Development Fund (FEDER, EU) (to M.C.); Spanish Ministry of Science and Innovation [FPI BES-2016-0077344 to J.V.M.]; C.S. received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement [H2020-MSCA-COFUND-2016-754433]; this study made use of data generated by the UK10K Consortium from UK10K COHORT IMPUTATION [EGAS00001000713]; formal agreement with the Barcelona Supercomputing Center (BSC); this study made use of data generated by the Genome of the Netherlands' project, which is funded by the Netherlands Organization for Scientific Research [184021007], allowing us to use the GoNL reference panel containing SVs, upon request (GoNL Data Access request 2019203); this study also used data generated by the Haplotype Reference Consortium (HRC) accessed through the European Genome-phenome Archive with the accession numbers EGAD00001002729; formal agreement of the Barcelona Supercomputing Center (BSC) with WTSI; this study made use of data generated by the 1000 Genomes (1000G), accessed through the FTP portal (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/); this study used the GeneHancer-for-AnnotSV dump for GeneCards Suite Version 4.14, through a formal agreement between the BSC and The Weizmann Institute of Science.
Conflict of interest statement. None declared.

REFERENCES

1. Belmont,J.W., Boudreau,A., Leal,S.M., Hardenbol,P., Pasternak,S., Wheeler,D.A., Willis,T.D., Yu,F., Yang,H., Gao,Y. *et al.* (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
2. Loh,P., Danecek,P., Palamara,P.F., Fuchsberger,C., Reshef,A., Finucane,H.K., Schoenherr,S., Forer,L., Mccarthy,S., Abecasis,G.R. *et al.* (2016) Reference-based phasing using the haplotype reference consortium panel. *Nat. Genet.*, **48**, 1443–1448.
3. Taliun,D., Harris,D.N., Kessler,M.D., Carlson,J., Szpiech,Z.A., Torres,R., Taliun,S.A.G., Corvelo,A., Gogarten,S.M., Kang,H.M. *et al.* (2021) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature*, **590**, 290–299.
4. Weischenfeldt,J., Symmons,O., Spitz,F. and Korbel,J.O. (2013) Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.*, **14**, 125–138.
5. Brandler,W.M., Antaki,D., Gujral,M., Kleiber,M.L., Whitney,J., Maile,M.S., Hong,O., Chapman,T.R., Tan,S., Tandon,P. *et al.* (2018) Paternally inherited cis-regulatory structural variants are associated with autism. *Science*, **20**, 327–331.
6. González,J.R., Ruiz-Arenas,C., Cáceres,A., Morán,I., López-Sánchez,M., Alonso,L., Tolosana,I., Guindo-Martínez,M., Mercader,J.M., Esko,T. *et al.* (2020) Polymorphic inversions underlie the shared genetic susceptibility of obesity-related diseases. *Am. J. Hum. Genet.*, **106**, 846–858.
7. Thibodeau,M.L., O'Neill,K., Dixon,K., Reisle,C., Mungall,K.L., Krzywinski,M., Shen,Y., Lim,H.J., Cheng,D., Tse,K. *et al.* (2020) Improved structural variant interpretation for hereditary cancer susceptibility using long-read sequencing. *Genet. Med.*, **22**, 1892–1897.
8. Manolio,T.A., Collins,F.S., Cox,N.J., Goldstein,D.B., Hindorff,L.A., Hunter,D.J., McCarthy,M.I., Ramos,E.M., Cardon,L.R., Chakravarti,A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
9. Becker,T., Lee,W.P., Leone,J., Zhu,Q., Zhang,C., Liu,S., Sargent,J., Shanker,K., Mil-homens,A., Cerveira,E. *et al.* (2018) FusorSV: an algorithm for optimally combining data from multiple structural variation detection methods. *Genome Biol.*, **19**, 38.
10. Collins,R.L., Brand,H., Karczewski,K.J., Zhao,X., Alföldi,J., Francioli,L.C., Khera,A. V., Lowther,C., Gauthier,L.D., Wang,H. *et al.* (2020) A structural variation reference for medical and population genetics. *Nature*, **581**, 444–451.
11. Sudmant,P.H., Rausch,T., Gardner,E.J., Handsaker,R.E., Abyzov,A., Huddleston,J., Zhang,Y., Ye,K., Jun,G., Hsi-Yang Fritz,M. *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**, 75–81.
12. Hehir-Kwa,J.Y., Marschall,T., Kloosterman,W.P., Francioli,L.C., Baaijens,J.A., Dijkstra,L.J., Abdellaoui,A., Koval,V., Thung,D.T., Wardenaar,R. *et al.* (2016) A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat. Commun.*, **7**, 12989.
13. Chaisson,M.J.P., Sanders,A.D., Zhao,X., Malhotra,A., Porubsky,D., Rausch,T., Gardner,E.J., Rodriguez,O.L., Guo,L., Collins,R.L. *et al.* (2019) Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.*, **10**, 1784.
14. Ho,S.S., Urban,A.E. and Mills,R.E. (2020) Structural variation in the sequencing era. *Nat. Rev. Genet.*, **21**, 171–189.
15. Audano,P.A., Sulovari,A., Graves-Lindsay,T.A., Cantsilieris,S., Sorensen,M., Welch,A.E., Dougherty,M.L., Nelson,B.J., Shah,A., Dutcher,S.K. *et al.* (2019) Characterizing the major structural variant alleles of the human genome. *Cell*, **176**, 663–675.

16. Ebert,P., Audano,P.A., Zhu,Q., Rodriguez-Martin,B., Porubsky,D., Bonder,M.J., Sulovari,A., Ebler,J., Zhou,W., Mari,R.S. *et al.* (2021) Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, **372**, eabf7117.
17. Obón-Santacana,M., Vilardell,M., Carreras,A., Duran,X., Velasco,J., Galván-Femenía,I., Alonso,T., Puig,L., Sumoy,L., Duell,E.J. *et al.* (2018) GCAT|Genomes for life: a prospective cohort study of the genomes of catalonia. *BMJ Open*, **8**, e018324.
18. Galván-Femenía,I., Obón-Santacana,M., Piñeyro,D., Guindo-Martinez,M., Duran,X., Carreras,A., Pluvinet,R., Velasco,J., Ramos,L., Aussó,S. *et al.* (2018) Multitrait genome association analysis identifies new susceptibility genes for human anthropometric variation in the GCAT cohort. *J. Med. Genet.*, **55**, 765–778.
19. Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.M., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C., Stuart,J.M. and Cancer Genome Atlas Research Network. (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
20. Huang,W., Li,L., Myers,J.R. and Marth,G.T. (2012) ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–594.
21. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**, 1754–1760.
22. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
23. Van der Auwera,G.A., Carneiro,M.O., Hartl,C., Poplin,R., Del Angel,G., Levy-Moonshine,A., Jordan,T., Shakir,K., Roazen,D., Thibault,J. *et al.* (2013) From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics*,**43**:11.10.1–11.10.33.
24. Zook,J.M., McDaniel,J., Olson,N.D., Wagner,J., Parikh,H., Heaton,H., Irvine,S.A., Trigg,L., Truty,R., McLean,C.Y. *et al.* (2019) An open resource for accurately benchmarking small variant and reference calls. *Nat. Biotechnol.*, **37**, 561–566.
25. Poplin,R., Ruano-Rubio,V., DePristo,M.A., Fennell,T.J., Carneiro,M.O., Auwera,G.A. Van der, Kling,D.E., Gauthier,L.D., Levy-Moonshine,A., Roazen,D. *et al.* (2017) Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv doi: https://doi.org/10.1101/201178, 24 July 2018, preprint: not peer reviewed.
26. Poplin,R., Chang,P.C., Alexander,D., Schwartz,S., Colthurst,T., Ku,A., Newburger,D., Dijamco,J., Nguyen,N., Afshar,P.T. *et al.* (2018) A universal snp and small-indel variant caller using deep neural networks. *Nat. Biotechnol.*, **36**, 983.
27. Kim,S., Scheffer,K., Halpern,A., Bekritsky,M., Enhuo,N., Källberg,M., Chen,X., Yeobin,K., Beyter,D., Krusche,P. *et al.* (2018) Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods*, **15**, 591–594.
28. Rimmer,A., Phan,H., Mathieson,I., Iqbal,Z. and Twigg,S.R.F. (2014) Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet*, **46**, 912–918.
29. Koboldt,D.C., Larson,D.E. and Wilson,R.K. (2013) Using varscan 2 for germline variant calling and somatic mutation detection. *Curr Protoc Bioinforma.*, **44**, 15.4.1–15.4.17.
30. Rausch,T., Zichner,T., Schlattl,A., Stütz,A.M., Benes,V. and Korbel,J.O. (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**, 333–339.
31. Chen,X., Schulz-Trieglaff,O., Shaw,R., Barnes,B., Schlesinger,F., Källberg,M., Cox,A.J., Kruglyak,S. and Saunders,C.T. (2016) Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, **32**, 1220–1222.
32. Ye,K., Guo,L., Yang,X., Lamijer,E.W., Raine,K. and Ning,Z. (2018) Split-read indel and structural variant calling using PINDEL. *Methods Mol. Biol.*, **1833**, 95–105.
33. Layer,R.M., Chiang,C., Quinlan,A.R. and Hall,I.M. (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.*, **15**, R84.
34. Kronenberg,Z.N., Osborne,E.J., Cone,K.R., Kennedy,B.J., Domyan,E.T., Shapiro,M.D., Elde,N.C. and Yandell,M. (2015) Wham: identifying structural variants of biological consequence. *PLoS Comput. Biol.*, **11**, e1004572.
35. Wala,J.A., Bandopadhayay,P., Greenwald,N.F., O'Rourke,R., Sharpe,T., Stewart,C., Schumacher,S., Li,Y., Weischenfeldt,J., Yao,X. *et al.* (2018) SvABA: Genome-wide detection of structural variants and indels by local assembly. *Genome Res.*, **28**, 581–591.
36. Abyzov,A., Urban,A.E., Snyder,M. and Gerstein,M. (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.*, **21**, 974–984.
37. Kehr,B., Melsted,P. and Halldórsson,B.V. (2016) PopIns: Population-scale detection of novel sequence insertions. *Bioinformatics*, **32**, 961–967.
38. Handsaker,R.E., Van Doren,V., Berman,J.R., Genovese,G., Kashin,S., Boettger,L.M. and Mccarroll,S.A. (2015) Large multiallelic copy number variations in humans. *Nat. Genet.*, **47**, 296–303.
39. Kavak,P., Lin,Y.Y., Numanagić,I., Asghari,H., Güngör,T., Alkan,C. and Hach,F. (2017) Discovery and genotyping of novel sequence insertions in many sequenced individuals. *Bioinformatics*, **33**, i161–i169.
40. Liu,S., Huang,S., Rao,J., Ye,W., Krogh,A. and Wang,J. (2015) Discovery, genotyping and characterization of structural variation and novel sequence at single nucleotide resolution from de novo genome assemblies on a population scale. *Gigascience*, **4**, 64.
41. Gardner,E.J., Lam,V.K., Harris,D.N., Chuang,N.T., Scott,E.C., Stephen Pittard,W., Mills,R.E. and Devine,S.E. (2017) The mobile element locator tool (MELT): Population-scale mobile element discovery and biology. *Genome Res.*, **27**, 1916–1929.
42. Tischler,G. and Leonard,S. (2014) Biobambam: tools for read pair collation based algorithms on BAM files. *Source Code Biol. Med.*, **9**, 13.
43. Rausch,T., Fritz,Hsi-Yang, Korbel,M. and Benes,V. (2019) Alfred: interactive multi-sample BAM alignment statistics, feature counting and feature annotation for long- and short-read sequencing. *Bioinformatics*, **35**, 2489–2491.
44. Jun,G., Flickinger,M., Hetrick,K.N., Romm,J.M., Doheny,K.F., Abecasis,G.R., Boehnke,M. and Kang,H.M. (2012) Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.*, **91**, 839–848.
45. Novembre,J., Johnson,T., Bryc,K., Kutalik,Z., Boyko,A.R., Auton,A., Indap,A., King,K.S., Bergmann,S., Nelson,M.R. *et al.* (2008) Genes mirror geography within europe. *Nature*, **456**, 98–101.
46. Sherry,S.T., Ward,M. and Sirotkin,K. (2001) dbSNP-Database for Single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.*, **1**, 1–8.
47. Lappalainen,I., Lopez,J., Skipper,L., Hefferon,T., Spalding,J.D., Garner,J., Chen,C., Maguire,M., Corbett,M., Zhou,G. *et al.* (2013) DbVar and DGVa: public archives for genomic structural variation. *Nucleic. Acids. Res.*, **41**, 936–941.
48. Abel,H.J., Larson,D.E., Regier,A.A., Chiang,C., Das,I., Kanchi,K.L., Layer,R.M., Neale,B.M., Salerno,W.J., Reeves,C. *et al.* (2020) Mapping and characterization of structural variation in 17,795 human genomes. *Nature*, **583**, 83–89.
49. Lerga-Jaso,J. (2019) In: *Integrative Analysis of the Functional Consequences of Inversions in the Human Genome*. Univ. Autònoma Barcelona.
50. Howie,B.N., Donnelly,P. and Marchini,J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLos Genet.*, **5**, e1000529.
51. Haraksingh,R.R., Abyzov,A. and Urban,A.E. (2017) Comprehensive performance comparison of high-resolution array platforms for genome-wide copy number variation (CNV) analysis in humans. *BMC Genomics*, **18**, 321.
52. Delaneau,O., Howie,B., Cox,A.J., Zagury,J.F. and Marchini,J. (2013) Haplotype estimation using sequencing reads. *Am. J. Hum. Genet.*, **93**, 687–696.
53. Menelaou,A. and Marchini,J. (2013) Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics*, **29**, 84–91.
54. Delaneau,O., Zagury,J.F., Robinson,M.R., Marchini,J.L. and Dermitzakis,E.T. (2019) Accurate, scalable and integrative haplotype estimation. *Nat. Commun.*, **10**, 24–29.
55. Patterson,M.D., Marschall,T., Pisanti,N., Van Iersel,L., Stougie,L., Klau,G.W. and Schönhuth,A. (2015) WhatsHap: weighted haplotype

assembly for future-generation sequencing reads. *J. Comput. Biol.*, **22**, 498–509.
56. Via,M., Gignoux,C. and Burchard,E.G. (2010) The 1000 genomes project: new opportunities for research and social challenges. *Genome Med.*, **2**, 8–10.
57. Hickey,G., Heller,D., Monlong,J., Sibbesen,J.A., Sirén,J., Eizenga,J., Dawson,E.T., Garrison,E., Novak,A.M. and Paten,B. (2020) Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.*, **21**, 35.
58. Walter,K., Min,J.L., Huang,J., Crooks,L., Memari,Y., McCarthy,S., Perry,J.R.B., Xu,C., Futema,M., Lawson,D. *et al.* (2015) The UK10K project identifies rare variants in health and disease. *Nature*, **526**, 82–89.
59. Guindo-martínez,M., Amela,R., Bonàs-guarch,S., Salvoro,C., Miguel-escalada,I., Carey,C.E., Cole,J.B., Rüeger,S., Atkinson,E., Leong,A. *et al.* (2021) The impact of non-additive genetic associations on age-related complex diseases. *Nat. Commun.*, **12**, 2436.
60. Geoffroy,V., Herenger,Y., Kress,A., Stoetzel,C., Piton,A., Dollfus,H. and Muller,J. (2018) AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics*, **34**, 3572–3574.
61. Fishilevich,S., Nudel,R., Rappaport,N., Hadar,R., Plaschkes,I., Iny Stein,T., Rosen,N., Kohn,A., Twik,M., Safran,M. *et al.* (2017) GeneHancer: genome-wide integration of enhancers and target genes in genecards. *Database (Oxford)*, **2017**, bax028.
62. Kumar,S., Harmanci,A., Vytheeswaran,J. and Gerstein,M.B. (2019) SVFX: a machine-learning framework to quantify the pathogenicity of structural variants. *Genome Biol.*, **21** , 274.
63. Cingolani,P., Platts,A., Lely Wang,L., Coon,M., Nguyen,T., Wang,L., Land,S., Lu,X. and Ruden,D. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80–92.
64. Cingolani,P., Patel,V.M., Coon,M., Nguyen,T., Land,S.J., Ruden,D.M. and Lu,X. (2012) Using drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, snpsift. *Front. Genet.*, **3**, 35.
65. Landrum,M.J., Lee,J.M., Benson,M., Brown,G.R., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Jang,W. *et al.* (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, **46**, D1062–D1067.
66. Rentzsch,P., Witten,D., Cooper,G.M., Shendure,J. and Kircher,M. (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic. Acids. Res.*, **47**, D886–D894.
67. Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A.R., Bender,D., Maller,J., Sklar,P., De Bakker,P.I.W., Daly,M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
68. Sutton,A., Crew,A. and Wysong,A. (2016) Redefinition of skin cancer as a chronic disease. *JAMA Dermatol.*, **152**, 255–256.
69. Chi,M. ju, Lee,C. yi and Wu,S. chong (2011) The prevalence of chronic conditions and medical expenditures of the elderly by chronic condition indicator (CCI). *Arch. Gerontol. Geriatr.*, **52**, 284–289.
70. Friedman,B., Jiang,H.J., Elixhauser,A. and Segal,A. (2006) Hospital inpatient costs for adults with multiple chronic conditions. *Med. Care Res. Rev.*, **63**, 327–346.
71. Fuller,Z.L., Berg,J.J., Mostafavi,H., Sella,G. and Przeworski,M. (2019) Measuring intolerance to mutation in human genetics. *Nat. Genet.*, **51**, 772–776.
72. Marchini,J. (2019) Haplotype estimation and genotype imputation. In: David,B., Ida,M. and John,M. (eds). *Handbook of Statistical Genomics*. John Wiley & Sons Ltd, Vol. **1**, 87–114.
73. Stelzer,G., Rosen,N., Plaschkes,I., Zimmerman,S., Twik,M., Fishilevich,S., Iny Stein,T., Nudel,R., Lieder,I., Mazor,Y. *et al.* (2016) The genecards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinforma.*, **2016**, 1.30.1–1.30.33.
74. Hara,Y., Balci-Hayta,B., Yoshida-Moriguchi,T., Kanagawa,M., Beltrán-Valero de Bernabé,D., Gündeşli,H., Willer,T., Satz,J.S., Crawford,R.W., Burden,S.J. *et al.* (2011) A dystroglycan mutation associated with limb-girdle muscular dystrophy. *N. Engl. J. Med.*, **364**, 939–946.