Article

Mapping and characterization of structural variation in 17,795 human genomes

https://doi.org/10.1038/s41586-020-2371-0

Received: 29 December 2018

Accepted: 18 May 2020

Published online: 27 May 2020

Check for updates

Haley J. Abel[1,2,68], David E. Larson[1,2,68], Allison A. Regier[1,3], Colby Chiang[1], Indraniel Das[1], Krishna L. Kanchi[1], Ryan M. Layer[4,5], Benjamin M. Neale[6,7,8], William J. Salerno[9], Catherine Reeves[10], Steven Buyske[11], NHGRI Centers for Common Disease Genomics*, Tara C. Matise[12], Donna M. Muzny[9], Michael C. Zody[10], Eric S. Lander[6,13,14], Susan K. Dutcher[1,2], Nathan O. Stitziel[1,2,3] & Ira M. Hall[1,2,3]✉

A key goal of whole-genome sequencing for studies of human genetics is to interrogate all forms of variation, including single-nucleotide variants, small insertion or deletion (indel) variants and structural variants. However, tools and resources for the study of structural variants have lagged behind those for smaller variants. Here we used a scalable pipeline[1] to map and characterize structural variants in 17,795 deeply sequenced human genomes. We publicly release site-frequency data to create the largest, to our knowledge, whole-genome-sequencing-based structural variant resource so far. On average, individuals carry 2.9 rare structural variants that alter coding regions; these variants affect the dosage or structure of 4.2 genes and account for 4.0–11.2% of rare high-impact coding alleles. Using a computational model, we estimate that structural variants account for 17.2% of rare alleles genome-wide, with predicted deleterious effects that are equivalent to loss-of-function coding alleles; approximately 90% of such structural variants are noncoding deletions (mean 19.1 per genome). We report 158,991 ultra-rare structural variants and show that 2% of individuals carry ultra-rare megabase-scale structural variants, nearly half of which are balanced or complex rearrangements. Finally, we infer the dosage sensitivity of genes and noncoding elements, and reveal trends that relate to element class and conservation. This work will help to guide the analysis and interpretation of structural variants in the era of whole-genome sequencing.

Human genetics studies use whole-genome sequencing (WGS) to enable comprehensive trait-mapping analyses across the full diversity of genome variation, including structural variants (SVs) of 50 base pairs (bp) or greater, such as deletions, duplications, insertions, inversions and other rearrangements. Previous work suggests that SVs have a disproportionately large role (relative to their abundance) in the biology of rare diseases[2] and in shaping heritable differences in gene expression in the human population[3–5]. Rare and de novo SVs have been implicated in the genetics of autism[6–10] and schizophrenia[11–14], but few other complex trait association studies have directly assessed SVs[15,16].

One challenge for the interpretation of SVs in WGS-based studies is the lack of high-quality publicly available variant maps from large populations. Our current knowledge is based primarily on three sources: (1) a large and disparate collection of array-based studies[17–19], with limited allele-frequency data and low resolution; (2) the 1000 Genomes Project callset[5], which has been invaluable but is limited by the modest sample size and low-coverage design; and (3) an assortment of smaller WGS-based studies with varied coverage, technologies, methods of analysis and levels of data accessibility[8,9,20–22].

There is an opportunity to improve our knowledge of SVs in human populations through the systematic analysis of large-scale WGS data resources that are generated by initiatives such as the National Human Genome Research Institute (NHGRI) Centers for Common Disease Genomics (CCDG). A key barrier to the creation of larger and more-informative catalogues of SVs is the lack of computational tools that can scale to the size of ever-growing datasets. To this aim, we have developed an SV analysis pipeline that is open source and highly scalable[1], and used it to map and characterize SVs in 17,795 deeply sequenced human genomes.

A population-scale map of SVs

The samples analysed here are derived from case–control studies and quantitative trait-mapping collections of common diseases that

[1]McDonnell Genome Institute, Washington University School of Medicine, St Louis, MO, USA. [2]Department of Genetics, Washington University School of Medicine, St Louis, MO, USA. [3]Department of Medicine, Washington University School of Medicine, St Louis, MO, USA. [4]BioFrontiers Institute, University of Colorado, Boulder, CO, USA. [5]Department of Computer Science, University of Colorado, Boulder, CO, USA. [6]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [7]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [8]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [9]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. [10]New York Genome Center, New York, NY, USA. [11]Department of Statistics, Rutgers University, Piscataway, NJ, USA. [12]Department of Genetics, Rutgers University, Piscataway, NJ, USA. [13]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [14]Department of Systems Biology, Harvard Medical School, Boston, MA, USA. [68]These authors contributed equally: Haley J. Abel, David E. Larson. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: ihall@genome.wustl.edu

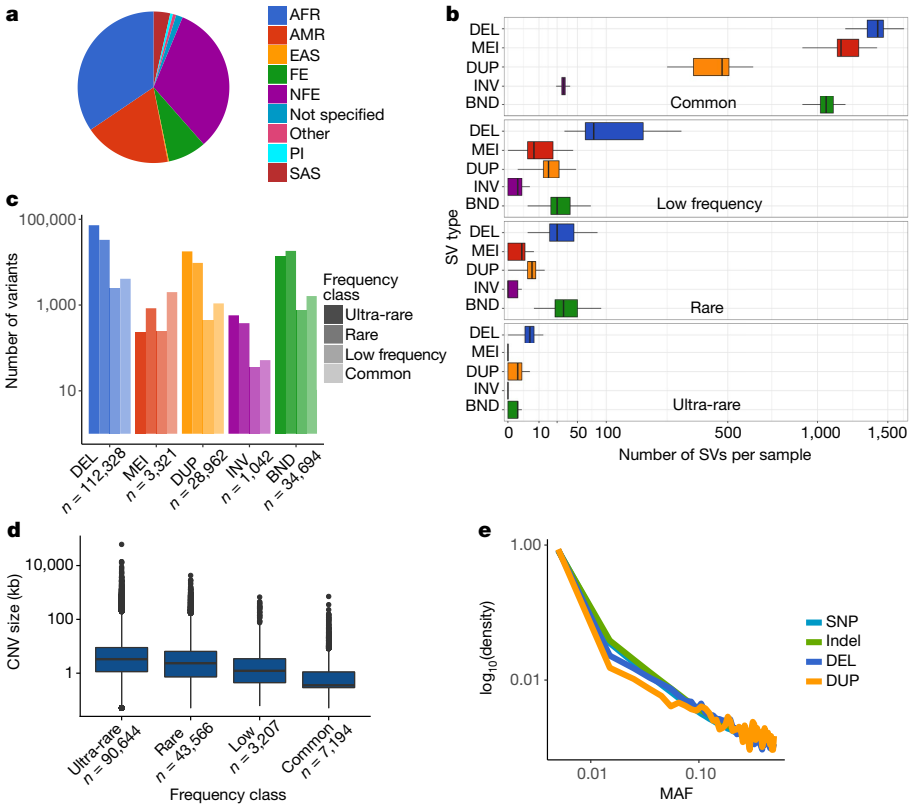


Fig. 1 | The public version of the B38 callset derived from 14,623 samples. **a**, Self-reported ancestry. AFR, African; AMR, admixed American; EAS, East Asian; FE, Finnish European; NFE, non-Finnish European; PI, Pacific Islander; SAS, South Asian. **b**, Number of SVs per sample (*x* axis, square-root-scaled) by SV type (*y* axis) and frequency class. SV types are: deletion (DEL), mobile-element insertion (MEI), duplication (DUP), inversion (INV) and breakend (BND). MAF bins are defined as ultra-rare (unique to an individual or family), rare (MAF < 1%), low frequency (1% < MAF < 5%) or common (MAF > 5%). **c**, Number of high-confidence SVs by type and frequency bin. **d**, CNV length distributions for each frequency class. **e**, MAF distribution for SNV (n = 85,687,916), indel (n = 9,477,540), DEL (n = 43,872) and DUP (n = 10,805) variants for a subset of 4,298 samples for which Genome Analysis Toolkit (GATK)-based SNV and indel calls were also available. All box plots in this figure indicate the median (centre line) and the first and third quartiles (box limits). The upper whiskers extend to the lesser extreme of the maximum and the third quartile plus 1.5 times the interquartile range (IQR); the lower whiskers extend to the lesser extreme of the minimum and the first quartile minus 1.5 times the IQR.

were sequenced under the CCDG programme, supplemented with ancestrally diverse samples from the Population Architecture Using Genomics and Epidemiology (PAGE) consortium and the Simons Genome Diversity Panel. The final ancestry composition includes 24% African, 16% Latino, 11% Finnish, 39% non-Finnish European and 9% other diverse samples from around the world (Extended Data Table 1).

The tools and pipelines used for this work are described elsewhere[1]. In brief, we developed a highly scalable software toolkit (svtools) and workflow for the generation of SV callsets on a large scale, which combines per-sample variant discovery[23], resolution-aware cross-sample merging, breakpoint genotyping[24], copy-number annotation and variant classification (Extended Data Fig. 1). We created two distinct SV callsets using different reference genome and pipeline versions. The 'B37' callset includes 118,973 high-confidence SVs from 8,426 samples that were sequenced at the McDonnell Genome Institute and aligned to the GRCh37 reference genome. The 'B38' callset includes 241,031 high-confidence SVs from 23,175 samples that were sequenced at four CCDG sites and aligned to GRCh38 using the 'functional equivalence' pipeline[25] (Methods). Of the 26,347 distinct samples in the union of the two callsets, aggregate-level sharing is permitted for 17,795; these make up the official public release (Supplementary Files 1, 2). For simplicity of presentation, most analyses below focus on the larger B38 callset (Supplementary Table 1).

We observed a mean of 4,442 high-confidence SVs per genome—predominantly deletions (35%), mobile-element insertions (MEIs) (27%) and tandem duplications (11%) (Fig. 1b, Extended Data Figs. 2, 3). Variant counts and linkage disequilibrium patterns are consistent with previous studies that used similar methods[4,5], and most SVs are mapped to base-pair resolution (Extended Data Figs. 2, 3). As expected, the site-frequency spectrum approximates that of single-nucleotide variants (SNVs) and indels, the size distribution shows increasing length with decreasing frequency, and principal component analysis (PCA) reveals a population structure that is consistent with self-reported ancestry (Fig. 1, Extended Data Figs. 2–4). Per-genome SV counts are broadly consistent and vary as expected on the basis of ancestry, with more genetic variation in individuals of African ancestry and fewer singletons in Finnish individuals (Extended Data Figs. 2, 3). Although we observe some technical variability owing to cohort and sequencing centre, these effects are mainly limited to small (less than 1 kb) copy-number variants (CNVs) that are detected solely by read-pair signals, which are sensitive to methods of library preparation and alignment filtering (Methods, Extended Data Fig. 3).

We further characterized callset quality using independent data and analyses (Supplementary Note) including (1) validation by deep-coverage (greater than 50×) long-read data from nine genomes; (2) sensitivity relative to a comprehensive long-read callset[26]; (3) inheritance patterns within a set of three-generation pedigrees; and (4) comparison to well-characterized short-read callsets[5,26] (Supplementary Tables 2–4, Extended Data Figs. 5–7). We achieve a validation rate of 84% by long-read data, with higher validation rates for the variant

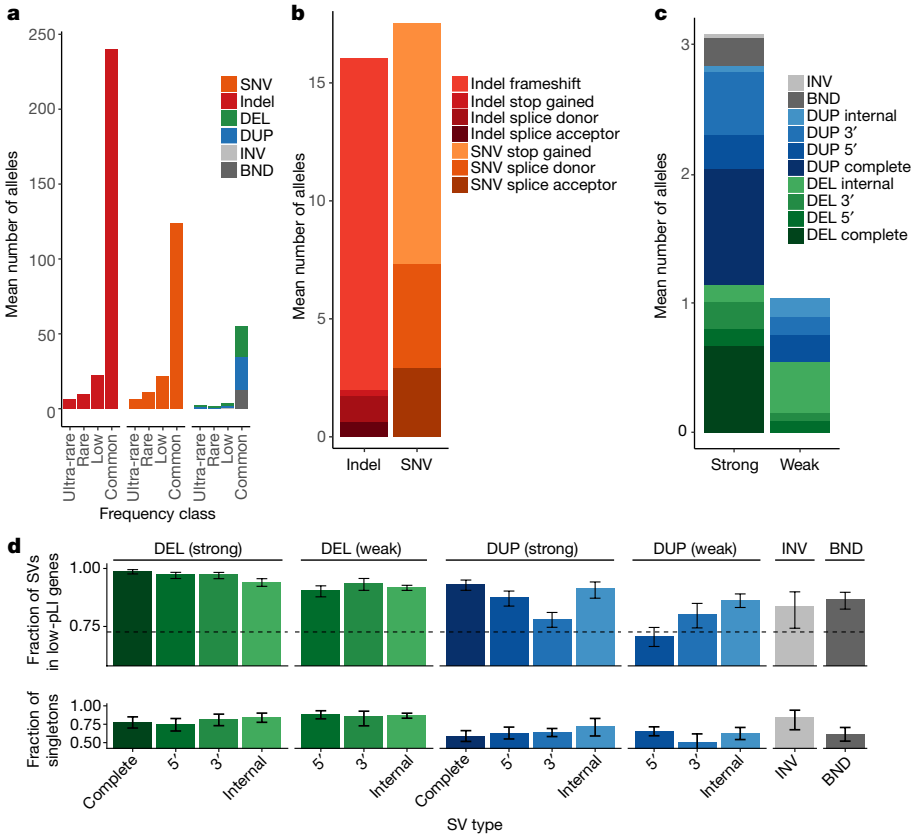


Fig. 2 | Burden of rare gene-altering SVs. a, Mean number of gene alterations per sample by type and frequency class ($n = 4,298$ samples). **b**, Mean number of rare (MAF < 1%) high-confidence protein-truncating variants per sample by type and VEP consequence. **c**, Mean number of rare (MAF < 1%) SV-derived gene alterations per sample by type. DEL and DUP are classified into strong (affecting more than 20% of exons of the principal transcript) and weak (affecting less than 20% of exons of the principal transcript) and sub-classified as internal (variant overlaps at least one coding exon, but neither the 3′ nor the 5′ end of the principal transcript), 3′ (variant overlaps the 3′ end of the transcript), 5′ (variant overlaps the 5′ end of the transcript) and complete (variant overlaps all coding exons in the principal transcript). **d**, Top, fraction of rare (MAF < 1%) gene-altering variants occurring in genes with a low pLI score (pLI < 0.9) by SV type and size class, stratified by affected gene region in the B38 callset ($n = 14,623$). The dotted line indicates the expected fraction, assuming a uniform distribution of SVs in coding exons. Bottom, fraction of singletons for gene-altering variants by type in the B38 callset ($n = 14,623$), restricted to genes with pLI > 0.1. Error bars (**d**, **e**) indicate 95% confidence intervals (Wilson score method). See Supplementary Table 5 for the number of variants in each category.

classes that are most relevant to the findings below: deletions (87%), rare SVs (90%) and singleton SVs (95%). On the basis of the validation rates of SV frequency classes and their relative abundance in the full dataset, we estimate a false discovery rate of 7.0%. Although the overall sensitivity is low (49%) compared to long-read SV maps—owing to the inherent difficulty of detecting repetitive variants from short reads—it is comparable to published short-read callsets[4,5,26] and is substantially higher for functionally relevant subtypes, such as SVs larger than 1 kb (63%) and predicted high-impact variants (82%).

Burden of deleterious rare SVs

The contribution of rare SVs to human disease remains unclear. Well-powered WGS-based trait-mapping studies will ultimately be required to address this; however, the overall burden of predicted pathogenic mutations in the human population is informative and can be estimated from our data. Our analysis of 14,623 individuals identified 42,765 rare SV alleles (minor allele frequency (MAF) of less than 1%) that are predicted to decrease gene dosage ($n = 9,416$), alter gene function (for example, single exon deletion; $n = 26,337$) or increase gene dosage ($n = 7,012$). The majority of rare gene-altering SVs are deletions (54.5%), with fewer duplications (42.2%) and a small fraction of other variant types, primarily inversions and complex rearrangements that interrupt or rearrange exons. Of these, 23.4% affect multiple genes and 10.4% affect three or more genes, resulting in a mean of 4.2 SV-altered genes per individual. On the basis of a strict definition of loss-of-function SVs—gene disruptions and gene deletions that affect more than 20% of exons—we identified a mean of 1.39 rare SV-based loss-of-function alleles per person. An analysis of 4,298 samples with SV calls and SNV or indel calls revealed that individuals carry a mean of 33.6 rare high-confidence loss-of-function SNVs and small indels (Fig. 2), consistent with previous studies[27]. Thus, SVs account for 4–11.2% of rare, predicted high-impact gene alterations in a population sample, depending on whether we consider all coding SVs or a strictly defined set of loss-of-function variants (Fig. 2c). These are likely to be underestimates, considering that the false-negative rate of SV detection is typically higher than that of SNVs and small indels[24,26].

To characterize the relative effect of different coding SV classes we calculated two measures of purifying selection (Fig. 2d): (1) the fraction of variants that affect dosage-tolerant genes with a loss-of-function intolerance (pLI)[27,28] score of less than 0.9; and (2) the fraction of variants that are present as singletons found in only one individual or family. By these measures, deletions are more deleterious than duplications, and complete gene deletions are the most deleterious class. Notably, on the basis of the fraction of variants in dosage-intolerant genes, complete gene duplications and sub-genic deletions that affect fewer than 20% of exons are relatively depleted; this suggests that many gene-altering

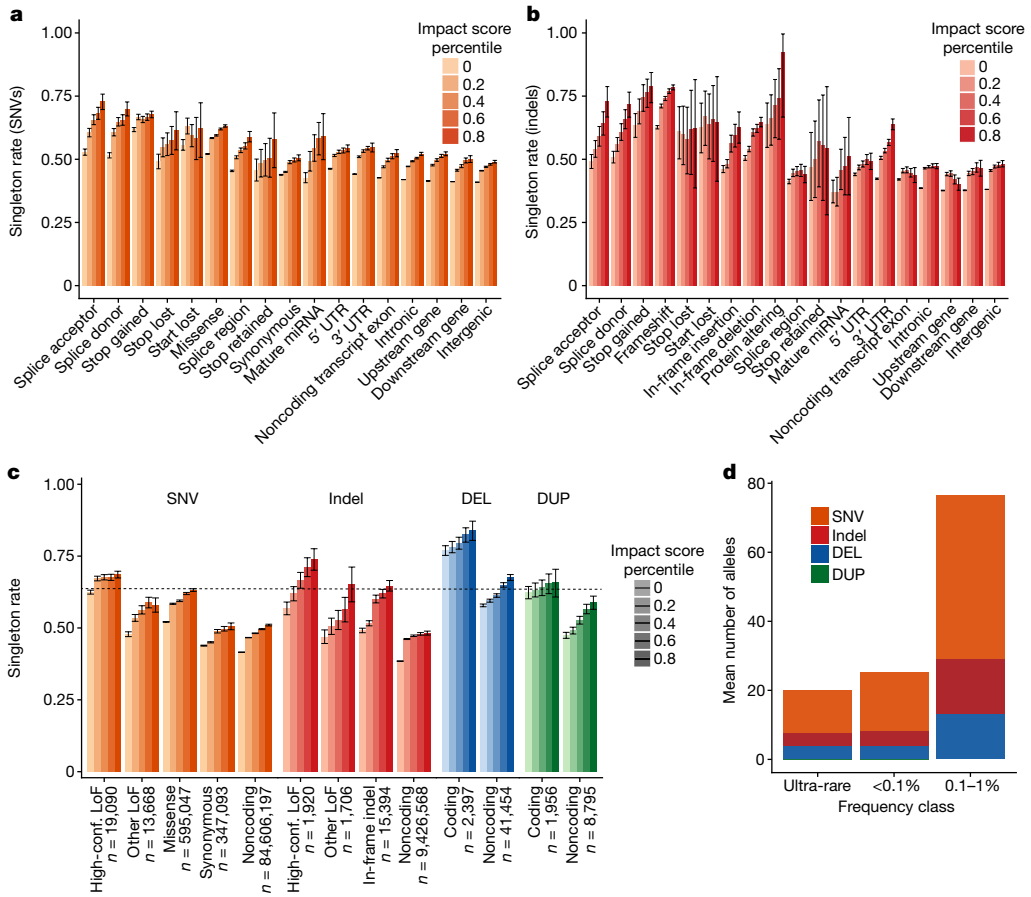


Fig. 3 | Estimation of genome-wide burden of high-impact functional alleles. a, Singleton rates for SNVs, by VEP consequence and percentile of the impact score (derived from combined LINSIGHT and CADD impact scores). miRNA, microRNA; UTR, untranslated region. **b**, Singleton rates for indels. **c**, Singleton rates by variant type and percentile of combined CADD–LINSIGHT impact score. The horizontal dotted line shows the singleton rate for all high-confidence (high-conf.) SNV or indel loss-of-function (LoF) mutations. 'Other LoF' indicates VEP-annotated protein-truncating variants that are not classified as high confidence by LOFTEE. DELs and DUPs that intersect with any coding exon of the principal transcript are classified as coding; otherwise, they are noncoding. **d**, Mean number of strongly deleterious alleles genome-wide per sample, by type and frequency class. Error bars (**a**–**c**) indicate 95% confidence intervals (Wilson score method). See Supplementary Table 6 for counts of variants in each category.

SVs are strongly deleterious, even those not predicted to completely obliterate gene function.

The above calculations ignore missense and noncoding variants, which are expected to make up a large fraction of rare functional variation. Predicting the effect of these variant types is challenging, but we can approximate their relative contribution to the deleterious variant burden under two simplifying assumptions: (1) impact-prediction algorithms such as CADD[29] and LINSIGHT[30] are capable of ranking variants within a given class (SNV, indel, SV) by their degree of deleteriousness; and (2) the mean deleterious impact of a given set of variants is reflected by its singleton rate. The first assumption is somewhat tenuous, but should be valid here given that impact-prediction inaccuracies are likely to affect all variant classes similarly; the second should hold under an infinite sites model of mutation, which is reasonable for the sample size (n = 4,298 samples) used in this analysis. We note that other evolutionary forces such as positive selection, background selection and biased gene conversion can also shape the site-frequency spectrum; however, we expect that these forces would act similarly on the variant classes examined here, given that this is a genome-wide analysis of a very large number of sites.

We used CADD and LINSIGHT to generate impact scores for SNVs, indels, deletions and duplications (Methods). As expected, these are highly correlated with singleton rate and variant effect predictions from the Ensembl Variant Effect Predictor (VEP)[31] and LOFTEE[27] (Fig. 3). We sought to identify 'strongly deleterious' variants from each class by choosing impact-score thresholds to match the singleton rate of the entire set of high-confidence loss-of-function mutations. Individuals carried a mean of 121.9 strongly deleterious rare variants, comprising 63% SNVs, 19.8% indels and 17.2% SVs (Fig. 3d). Given the relative numerical abundance of different rare variant classes, this suggests that a given rare SV is 841-fold more likely to be strongly deleterious than a rare SNV, and 341-fold more likely than a rare indel. Predicted deleterious SVs are slightly larger than rare SVs on the whole (median 4.5 versus 2.8 kb). Whereas only a minority (13.1%) of predicted strongly deleterious SNVs and indels are noncoding, 90.1% of predicted strongly deleterious rare SVs are noncoding. In particular, the top 50% of noncoding deletions show similar levels of purifying selection (as measured by singleton rate) as high-confidence loss-of-function variants that are caused by SNVs or indels (Fig. 3c), suggesting that a typical individual carries 19.1 alleles for strongly deleterious rare noncoding deletions. This suggests that noncoding deletions may have strongly deleterious effects, and may have a larger than expected role in human disease.

Landscape of ultra-rare SVs

Most ultra-rare SVs represent recent or de novo structural mutations, and thus the relative abundance of different classes of ultra-rare SV sheds light on the underlying mutational processes. We identified 158,991 ultra-rare SVs (105,175 high-confidence) that were

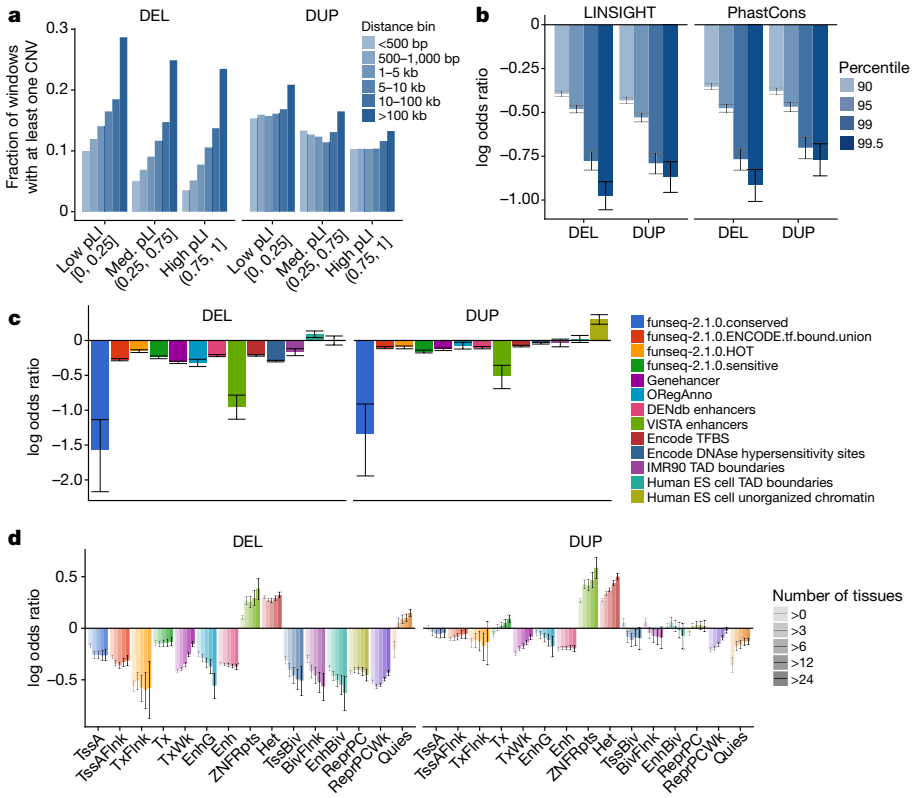


Fig. 4 | Dosage sensitivity of functional annotations. a, Fraction of 1-kb genomic windows that contain at least one CNV, as a function of the distance to the nearest coding exon and the pLI of that gene. **b**, Depletion of CNVs in conserved genomic regions. Odds ratios (log-transformed) for the occurrence of CNVs in highly conserved (based on LINSIGHT or PhastCons percentile) versus less-conserved regions. Odds ratios are Cochran–Mantel–Haenszel estimates, stratified by the distance to the nearest coding exon and the pLI of that gene. **c**, Odds ratios (log-transformed; estimated as in **b**) for the occurrence of CNVs in 1-kb windows that intersect various functional annotation tracks. Human ES cell, human embryonic stem cell; TAD, topologically associated domain; TFBS, transcription-factor-binding site. **d**, Odds ratios (log-transformed; estimated as in **b**) for the occurrence of CNVs in 1-kb windows that overlap Roadmap segmentations, stratified by the number of Roadmap tissues in which the region is observed. Error bars (**b**–**d**) indicate 95% confidence intervals estimated by block bootstrap.

present in only one of 14,623 individuals or were unique to a family. This corresponds to a mean of around 11.4 per individual (Extended Data Fig. 8a). Ultra-rare SVs are mainly composed of deletions (5.2 per person) and duplications (1.3), with a smaller number of inversions (0.17).

It is notable that around 40% of ultra-rare SV breakpoints in our dataset cannot be readily classified into the canonical forms of SV. This is a known limitation of short-read WGS, and such variants are often ignored. Formally, these SVs are of the 'breakend' (BND) class, which is a generic term in the VCF specification for SV breakpoints that cannot be unequivocally classified[32]. We examined the 63,559 ultra-rare BNDs for insights into their composition and origin. Many (17.0%) appear to be deletions that are too small (less than 100 bp) to exhibit convincing read-depth support, and that our pipeline conservatively classifies as BNDs (for example, complex SVs can masquerade as deletions). Some (2.4%) of the ultra-rare BNDs stem from 1,542 'retrogene insertions', which are caused by retroelement machinery acting on mRNAs. This set of retrogene insertions is around 10-fold larger than those of previous maps[33–35] and will be valuable for future studies. Another 5.5% of ultra-rare BNDs are complex genomic rearrangements with multiple breakpoints in close proximity (less than 100 kb). The remainder are variants that are difficult to classify, which involve local (49.9%, within 1 Mb) or distant (5.7%, more than 1 Mb apart) intra-chromosomal alterations or inter-chromosomal alterations (27.2%), and of which many (78.0%) are classified as low-confidence SV calls. This final class is probably caused primarily by variation in repetitive elements, but is also expected to be enriched for false positives.

A variety of sporadic disorders are caused by extremely large and/or complex SVs, but—owing to the limitations of the array-based methods that have been used in previous large-scale studies[36], which fail to detect balanced events or resolve complex variant architectures—our knowledge of the frequency and architecture of these marked alterations in the general population is incomplete. We observed 138 megabase-scale CNVs, which corresponds to a frequency of around 0.01 per individual; these include 47 deletions and 91 duplications, and affect a mean of 12.1 genes (Extended Data Fig. 8b). Three individuals carried two megabase-scale CNVs, apparently owing to independent mutations. We observed 19 reciprocal translocations (0.001 per individual), consistent with previous cytogenetic-based estimates[37,38]. Of these translocations, 14 affect one gene and two affect two genes, producing one predicted in-frame gene fusion (PI4KA:MGLL). We applied breakpoint clustering (as in a previous study[39]) to identify ultra-rare complex rearrangements and discovered 33 complex SVs that span more than 1 Mb (0.003 per individual). Most of these (20 out of 33, 60.6%) involve three breakpoints; however, we observed five large-scale rearrangements with five or more breakpoints. Notably, when the entire SV size distribution is considered, 3.3% of ultra-rare SVs are complex variants, which is consistent with previous smaller-scale studies[40–44].

Dosage sensitivity

A motivation for creating population-scale SV maps is to annotate genomic regions on the basis of their tolerance to dosage changes and structural rearrangements, thus revealing the genes and noncoding

elements that are most important (or dispensable) for human development and viability. The pLI score from the Exome Aggregation Consortium (ExAC) and the Genome Aggregation Database (gnomAD)[27,28] has proven invaluable for this purpose, but does not predict the effects of increased dosage or include noncoding elements.

We first generated deletion (DEL) and duplication (DUP) sensitivity scores for each gene on the basis of the observed frequency of CNVs in the combined dataset of 17,795 samples (as in a previous study[45]; see Methods). The resulting scores correlate with the CNV scores from ExAC[45], and with the DECIPHER haploinsufficiency score[46] (Extended Data Fig. 9). Despite their relatively modest correlations with one another, all three measures are informative compared with pLI, which was generated using an independent set of variants (SNVs and indels). A combined score from multiple datasets performs better than any single score, and may be useful for interpreting rare SVs (Supplementary File 4).

We next performed a genome-wide analysis based on the frequency of dosage alterations in 1-kb genomic windows (Methods). Our current dataset is not large enough to predict dosage-sensitive noncoding elements on the basis of the absence of variation; however, we can investigate the relative sensitivity of genomic features in aggregate. As expected, we observed a strong depletion of CNVs near coding exons, which varied according to the proximity to the nearest exon as well as the pLI of the corresponding gene (Fig. 4a). We therefore estimated odds ratios for depletion of CNVs in each functionally annotated region, stratified by distance to and pLI of the nearest exon. The resulting dosage-sensitivity scores mirror independent measures of selective constraint including LINSIGHT and PhastCons (Fig. 4b).

We also examined the relative dosage sensitivity of regulatory and epigenomic annotations from various projects[47–52] (Fig. 4). Regulatory elements such as enhancers, polycomb repressors, DNase hypersensitivity sites and transcription-factor-binding sites show strong sensitivity to dosage loss through deletion, whereas regions of inert noncoding annotations do not. The patterns of sensitivity to dosage gains through duplication are broadly similar, albeit weaker, with no obviously distinct patterns at (for example) enhancers, repressors or insulators. The dosage sensitivity of regulatory elements at 'bivalent' genomic regions from the NIH Roadmap Epigenomics project is greater than their counterparts (for example, enhancers versus bivalent enhancers), suggesting that such elements may be under especially strong selection. Furthermore, dosage sensitivity increases with the number of cell types that share a given annotation, suggesting that sensitivity is higher for constitutive regulatory elements compared to those that act in a more cell-type-specific manner.

Discussion

Here, we have conducted the largest—to our knowledge—WGS-based study of SVs in the human population so far. The sample size and use of deep (greater than 20×) WGS allowed us to map rare SVs at high genomic resolution and estimate the relative burden of deleterious SVs. Our data suggest that rare SVs account for 4–11.2% of deleterious coding alleles and 17.2% of deleterious alleles genome-wide—a disproportionate contribution considering that SVs comprise roughly 0.1% of variants. The burden of rare, strongly deleterious noncoding deletions that is apparent in our dataset is notable: we estimate that a typical individual carries 19.1 rare noncoding deletions that exhibit levels of purifying selection similar to loss-of-function SNVs and indels (of which there are 33.6 per individual). These results indicate that comprehensive assessment of SVs will improve power in rare-variant association studies.

The public site-frequency maps reported here will also aid the interpretation of variants in smaller-scale WGS-based studies (for example, through look-ups of allele frequency), in particular as they were generated by a systematic joint analysis of large datasets from diverse populations (similar to ExAC and gnomAD[27]). One limitation is the high false-negative rate for repetitive SVs, including MEIs, short tandem repeats (STRs) and multi-allelic CNVs, owing to the limitations of algorithms that rely on unique short-read alignments. Whereas we have reported a mean of 4,442 SVs per genome, recent long-read analyses predict up to around 27,662 SVs per genome, including STRs and other highly repetitive elements[26]. Although the inherent limitations of short-read WGS cannot be overcome, this resource could be made more comprehensive in future work with specialized algorithms tailored to MEIs, STRs and multi-allelic CNVs.

Finally, we have mined this resource to assess the dosage sensitivity of genes and noncoding elements. For genes, our results complement existing estimates from exome-sequencing and microarray data; for noncoding elements, we observe strong correlations with measures of nucleotide conservation, purifying selection, activity of regulatory elements and cell-type specificity. Although our current sample size is insufficient to assess the dosage sensitivity of individual noncoding elements, this will become feasible as large-scale WGS resources from ongoing international programs become available.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2371-0.

1. Larson, D. E. et al. svtools: population-scale analysis of structural variation. *Bioinformatics* **35**, 4782–4787 (2019).
2. 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).
3. Stranger, B. E. et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007).
4. Chiang, C. et al. The impact of structural variation on human gene expression. *Nat. Genet*. **49**, 692–699 (2017).
5. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
6. Sebat, J. et al. Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
7. Weiss, L. A. et al. Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med*. **358**, 667–675 (2008).
8. Turner, T. N. et al. Genomic patterns of de novo mutation in simplex autism. *Cell* **171**, 710–722 (2017).
9. Werling, D. M. et al. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat. Genet*. **50**, 727–736 (2018).
10. Brandler, W. M. et al. Paternally inherited cis-regulatory structural variants are associated with autism. *Science* **360**, 327–331 (2018).
11. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237–241 (2008).
12. Walsh, T. et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539–543 (2008).
13. McCarthy, S. E. et al. Microduplications of 16p11.2 are associated with schizophrenia. *Nat. Genet*. **41**, 1223–1227 (2009).
14. Marshall, C. R. et al. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat. Genet*. **49**, 27–35 (2017).
15. Craddock, N. et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**, 713–720 (2010).
16. Kathiresan, S. et al. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat. Genet*. **41**, 334–341 (2009).
17. MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res*. **42**, D986–D992 (2014).
18. Bragin, E. et al. DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic Acids Res*. **42**, D993–D1000 (2014).
19. Lappalainen, I. et al. dbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res*. **41**, D936–D941 (2013).
20. 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).
21. Maretty, L. et al. Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature* **548**, 87–91 (2017).

22. Sudmant, P. H. et al. Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**, aab3761 (2015).
23. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*. **15**, R84 (2014).
24. Chiang, C. et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods* **12**, 966–968 (2015).
25. Regier, A. A. et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat. Commun*. **9**, 4038 (2018).
26. Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun*. **10**, 1784 (2019).
27. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
28. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
29. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet*. **46**, 310–315 (2014).
30. Huang, Y. F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet*. **49**, 618–624 (2017).
31. McLaren, W. et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).
32. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
33. Ewing, A. D. et al. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome Biol*. **14**, R22 (2013).
34. Schrider, D. R. et al. Gene copy-number polymorphism caused by retrotransposition in humans. *PLoS Genet*. **9**, e1003242 (2013).
35. Abyzov, A. et al. Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. *Genome Res*. **23**, 2042–2052 (2013).
36. Cooper, G. M. et al. A copy number variation morbidity map of developmental delay. *Nat. Genet*. **43**, 838–846 (2011).
37. Hook, E. B. & Hamerton, J. L. in *Population Cytogenetics: Studies in Humans* (eds Hook, E. B. & Porter, I. H.) 63–79 (Academic Press, 1977).
38. Forabosco, A., Percesepe, A. & Santucci, S. Incidence of non-age-dependent chromosomal abnormalities: a population-based study on 88965 amniocenteses. *Eur. J. Hum. Genet*. **17**, 897–903 (2009).
39. Malhotra, A. et al. Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. *Genome Res*. **23**, 762–776 (2013).
40. Conrad, D. F. et al. Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nat. Genet*. **42**, 385–391 (2010).
41. Quinlan, A. R. et al. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res*. **20**, 623–635 (2010).
42. Mills, R. E. et al. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
43. Kidd, J. M. et al. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* **143**, 837–847 (2010).
44. Quinlan, A. R. & Hall, I. M. Characterizing complex structural variation in germline and somatic genomes. *Trends Genet*. **28**, 43–53 (2012).
45. Ruderfer, D. M. et al. Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nat. Genet*. **48**, 1107–1111 (2016).
46. Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet*. **6**, e1001154 (2010).
47. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
48. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
49. Khurana, E. et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013).
50. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Res*. **35**, D88–D92 (2007).
51. Fishilevich, S. et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* **2017**, bax028 (2017).
52. Lesurf, R. et al. ORegAnno 3.0: a community-driven resource for curated regulatory annotation. *Nucleic Acids Res*. **44**, D126–D132 (2016).

NHGRI Centers for Common Disease Genomics

Goncalo R. Abecasis[15], Elizabeth Appelbaum[1], Julie Baker[16], Eric Banks[6], Raphael A. Bernier[17], Toby Bloom[10], Michael Boehnke[15], Eric Boerwinkle[9,18], Erwin P. Bottinger[19], Steven R. Brant[20], Esteban G. Burchard[21], Carlos D. Bustamante[16], Lei Chen[1], Judy H. Cho[19,22,23], Rajiv Chowdhury[24], Ryan Christ[1], Lisa Cook[1], Matthew Cordes[1], Laura Courtney[1], Michael J. Cutler[25], Mark J. Daly[6,26,27], Scott M. Damrauer[28], Robert B. Darnell[10,29,30], Tracie Deluca[1], Huyen Dinh[9], Harsha Doddapaneni[9], Evan E. Eichler[31,32], Patrick T. Ellinor[6,33], Andres M. Estrada[34], Yossi Farjoun[6], Adam Felsenfeld[35], Tatiana Foroud[36], Nelson B. Freimer[37], Catrina Fronick[1], Lucinda Fulton[1], Robert Fulton[1], Stacy Gabriel[6], Liron Ganel[1], Shailu Gargeya[10], Goren Germer[10], Daniel H. Geschwind[38,39,40], Richard A. Gibbs[9], David B. Goldstein[41,42], Megan L. Grove[9], Namrata Gupta[6], Christopher A. Haiman[43], Yi Han[9], Daniel Howrigan[6,27], Jianhong Hu[9], Carolyn Hutter[35], Ivan Iossifov[44], Bo Ji[1], Lynn B. Jorde[45], Goo Jun[18], John Kane[46], Chul Joo Kang[1], Hyun Min Kang[15], Sek Kathiresan[6,33,47], Eimear E. Kenny[19,22,48,49], Lily Khaira[10], Ziad Khan[9], Amit Khera[6,33,47], Charles Kooperberg[50], Olga Krasheninina[9], William E. Kraus[51], Subra Kugathasan[52], Markku Laakso[53], Tuuli Lappalainen[10,54], Adam E. Locke[1,3], Ruth J. F. Loos[19], Amy Ly[1], Robert Maier[6,27], Tom Maniatis[10,55], Loic Le Marchand[56], Gregory M. Marcus[57], Richard P. Mayeux[58], Dermot P. B. McGovern[59], Karla S. Mendoza[34], Vipin Menon[9], Ginger A. Metcalf[9], Zeineen Momin[9], Guiseppe Narzisi[10], Joanne Nelson[1], Caitlin Nessner[9], Rodney D. Newberry[3], Kari E. North[60], Aarno Palotie[6,26,27], Ulrike Peters[50], Jennifer Ponce[1], Clive Pullinger[46], Aaron Quinlan[45], Daniel J. Rader[61], Stephen S. Rich[62], Samuli Ripatti[6,26,27], Dan M. Roden[63], Veikko Salomaa[64], Jireh Santibanez[9], Svati H. Shah[51], M. Benjamin Shoemaker[63], Heidi Sofia[35], Taylorlyn Stephan[35], Christine Stevens[6], Stephan R. Targan[59], Marja-Riitta Taskinen[65,66], Kathleen Tibbetts[6], Charlotte Tolonen[6], Tychele Turner[31], Paul De Vries[18], Jason Waligorski[1], Kimberly Walker[9], Vivian Ota Wang[35], Michael Wigler[10,44], Richard K. Wilson[1,67], Lara Winterkorn[10], Genevieve Wojcik[16], Jinchuan Xing[12], Erica Young[1,3], Bing Yu[18] & Yeting Zhang[12]

[15]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, School of Public Health, Ann Arbor, MI, USA. [16]Department of Genetics, Stanford University, Stanford, CA, USA. [17]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA. [18]Human Genetics Center and Department of Epidemiology, University of Texas Health Science Center, Houston, TX, USA. [19]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [20]Department of Medicine, Rutgers Robert Wood Johnson Medical School, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA. [21]Department of Bioengineering, University of California, San Francisco, San Francisco, CA, USA. [22]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [23]Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [24]MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [25]Intermountain Heart Institute, Intermountain Medical Center, Murray, UT, USA. [26]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. [27]Analytical and Translational Genetics Unit, Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [28]Department of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [29]Laboratory of Molecular Neuro-Oncology, The Rockefeller University, New York, NY, USA. [30]Howard Hughes Medical Institute, The Rockefeller University, New York, NY, USA. [31]Department of Genome Science, University of Washington, Seattle, WA, USA. [32]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [33]Division of Cardiology, Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [34]National Laboratory of Genomics for Biodiversity (LANGEBIO), CINVESTAV, Irapuato, Mexico. [35]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [36]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA. [37]Center for Neurobehavioral Genetics, Jane and Terry Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, Los Angeles, CA, USA. [38]Program in Neurogenetics, Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [39]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [40]Institute of Precision Health, University of California, Los Angeles, Los Angeles, CA, USA. [41]Institute for Genomic Medicine, Columbia University Medical Center, New York, NY, USA. [42]Department of Genetics and Development, Columbia University Medical Center, New York, NY, USA. [43]Department of Preventative Medicine, University of Southern California, Los Angeles, CA, USA. [44]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. [45]Department of Human Genetics, University of Utah, Salt Lake City, UT, USA. [46]Cardiovascular Research Institute, University of California, San Francisco, San Francisco, CA, USA. [47]Center for Genomic Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [48]The Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [49]Center for Statistical Genetics, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [50]Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [51]Department of Medicine, Duke University, Durham, NC, USA. [52]Department of Pediatrics, Emory University School of Medicine, Atlanta, GA, USA. [53]Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland, Kuopio, Finland. [54]Department of Systems Biology, Columbia University, New York, NY, USA. [55]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA. [56]Cancer Center, University of Hawaii, Honolulu, HI, USA. [57]Department of Medicine, University of California, San Francisco, San Francisco, CA, USA. [58]Department of Neurology, Columbia University, New York, NY, USA. [59]F. Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [60]Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA. [61]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [62]Center for Public Health Genomics, University of Virginia School of Medicine, Charlottesville, VA, USA. [63]Department of Medicine, Vanderbilt University, Nashville, TN, USA. [64]National Institute for Health and Welfare, Helsinki, Finland. [65]Research Programs Unit, Diabetes and Obesity, University of Helsinki, Helsinki, Finland. [66]Heart and Lung Centre, Helsinki University Hospital, Helsinki, Finland. [67]Present address: Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, USA. *A full list of authors and their affiliations appears in the Supplementary Information.

Methods

Generation of the B38 callset

Per-sample processing. This callset is derived from 23,559 individuals who were part of the CCDG programme as well as 950 Latino samples from the PAGE consortium. All data were produced at one of the four CCDG-funded sequencing centres and aligned to genome build GRCh38 using each individual centre's functionally equivalent pipeline implementation[25]. Per-sample calling was performed on 23,547 samples using LUMPY[23] (v.0.2.13), CNVnator[53] (v.0.3.3) and SVTyper[24] (v.0.1.4). We excluded human leukocyte antigen (HLA) sequences, decoy or alternate contigs and regions with copy number much higher than that expected (mean of 12 or more copies per genome across 409 samples) from SV calling with LUMPY (https://github.com/hall-lab/speedseq/blob/master/annotations/exclude.cnvnator_100bp.GRCh38.20170403.bed).

Per-sample quality control. We observed an excess of small (400–1,000-bp) singleton deletions (that is, present in only a single sample), suggesting a large number of false positives. On further investigation, this excess arose from differences between centres in library insert-size distribution. To reduce the number of false-positive small deletions, deletions of ≤1,000 bp were eliminated unless they had split read support in at least one sample. Subsequently, per-sample quality control was performed to eliminate outlier samples. We removed 213 samples in which variant counts (for any SV type) were >6 median absolute deviations from the median count for that type.

Merging and cohort-level re-genotyping. The remaining samples were processed into a single, joint callset using svtools[1] (https://github.com/hall-lab/svtools) (v.0.3.2), modified to allow for multi-stage merging. The code for this merging is available in a container hosted on DockerHub (https://hub.docker.com/r/ernfrid/svtools_merge_beta) (ernfrid/svtools_merge_beta:292bd3). Samples were merged using svtools lsort followed by svtools lmerge in batches of 1,000 samples (or fewer) within each cohort. The resulting per-cohort batches were then merged again using svtools lsort and svtools lmerge to create a single set of variants for the entire set of 23,331 remaining samples. This site list was then used to genotype each candidate site in each sample across the entire cohort using SVTyper (v.0.1.4). Genotypes for all samples were annotated with copy-number information from CNVnator. Subsequently, the per-sample VCFs were combined together using svtools vcfpaste. The resulting VCF was annotated with allele frequencies using svtools afreq, duplicate SVs were pruned using svtools prune, variants were reclassified using svtools classify (large sample mode) and any identical lines were removed. For reclassification of chromosomes X and Y, we used a container hosted on DockerHub (https://hub.docker.com/r/ernfrid/svtools_classifier_fix) (ernfrid/svtools_classifier_fix:v1). All other steps to assemble the cohort above used the same container that was used for merging.

Callset tuning. Using the variant calling control trios, we chose a mean sample quality (MSQ) cut-off for INV and BND variant calls that yielded a Mendelian error rate of approximately 5%. INVs passed if: MSQ ≥ 150; neither split-read nor paired-end LUMPY evidence made up more than 10% of total evidence; each strand provided at least 10% of read support. BNDs passed if MSQ ≥ 250.

Genotype refinement. MEI and DEL genotypes were set to missing on a per-sample basis (https://github.com/hall-lab/svtools/blob/develop/scripts/filter_del.py, commit 5c32862) if the site was poorly captured by split reads. Genotypes were set to missing if the size of the DEL or MEI was smaller than the minimum size discriminated at 95% confidence by SVTyper (https://github.com/hall-lab/svtools/blob/develop/scripts/del_pe_resolution.py, commit 3fc7275). DEL and MEI genotypes for sites with allele frequency ≥ 0.01 were refined based on clustering of allele balance and copy-number values within the datasets produced by each sequencing centre (https://github.com/hall-lab/svtools/blob/develop/scripts/geno_refine_12.py, commit 41fdd60). In addition, duplications were re-genotyped with more-sensitive parameters to better reflect the expected allele balance for simple tandem duplications (https://github.com/ernfrid/regenotype/blob/master/resvtyper.py, commit 4fadcc4).

Filtering for size. The remaining variants were filtered to meet the size definition of a SV (≥50 bp). The length of intra-chromosomal generic BNDs was calculated using vawk (https://github.com/cc2qe/vawk) as the difference between the reported positions of each breakpoint.

Large callset sample quality control. Of the remaining samples, we evaluated per-sample counts of deletions, duplications and generic BNDs within the low-allele-frequency (0.1%–1%) class. Samples with variant counts exceeding 10 median absolute deviations from the mean for any of the 3 separate variant classes were removed. In addition, we removed samples with genotype missingness >2%. These quality control filters removed a total of 120 additional samples. Finally, we removed 64 samples that were identified as duplicates or twins in a larger set of data.

Breakpoint resolution

Breakpoint resolution was calculated using BCFtools (v.1.3.1) query to create a table of confidence intervals for each variant in the callset, but excluding secondary BNDs. Each breakpoint contains two 95% confidence intervals, one each around the start location and end location. Summary statistics were calculated in RStudio (v.1.0.143; R v.3.3.3).

Self-reported ethnicity

Self-reported ethnicity was provided for each sample via the sequencing centre and aggregated by the NHGRI Genome Sequencing Program (GSP) coordinating centre. For each combination of reported ethnicity and ancestry, we assigned a super-population, continent (based on the cohort) and ethnicity. Samples in which ancestry was unknown, but the sample was Hispanic, were assigned to the Americas (AMR) super-population. Summarized data are presented in Extended Data Table 1.

Sample relatedness

As SNV calls were not yet available for all samples at the time of the analysis, relatedness was estimated using large (>1 kb), high-quality autosomal deletions and MEIs with allele frequency >1%. These were converted to plink format using PLINK (v.1.90b3.38) and then subjected to kinship calculation using KING[54] (v.2.0). The resulting output was parsed to build groups of samples connected through first-degree relationships (kinship coefficient > 0.177). Correctness was verified by the successful recapitulation of the 36 complete Coriell trios included as variant calling controls. We note that, in analyses of the full B38 callset (which contains cohorts of families), 'ultra-rare' or 'singleton' variants were defined as those unique to a family. For analyses of the of 4,298 sample subset of unrelated individuals with both SV and SNP/indel calls, 'singleton' variants were defined as those present as a single allele.

Callset summary metrics

Callset summary metrics were calculated by parsing the VCF files with BCFtools (v.1.3.1) query to create tables containing information for each variant–sample pairing or variant alone, depending on the metric. Breakdowns of the BND class of variation were performed using vawk to calculate orientation classes and sizes. These were summarized using Perl and then transformed and plotted using RStudio (v.1.0.143; R v.3.3.3).

Ultra-rare variant analysis

We defined an ultra-rare variant as any variant unique to one individual or one family of first-degree relatives. We expect the false-positive rate of ultra-rare variants to be low because systematic false positives owing to alignment issues are likely to be observed in multiple unrelated individuals. Therefore, we considered both high- and low-confidence variants in all ultra-rare analyses.

Constructing variant chains. Complex variants were identified as described previously[4] by converting each ultra-rare SV to bed format and, within a given family, clustering breakpoints occurring within 100,000 bp of each other using BEDTools[55] (v.2.23.0) cluster. Any clusters linked together by BND variants were merged together. The subsequent collection of variant clusters and linked variant clusters (hereafter referred to as chains) were used for both retrogene and complex variant analyses.

Manual review. Manual review of variants was performed using the Integrative Genomics Viewer (IGV) (v.2.4.0). Variants were converted to BED12 using svtools (v.0.3.2) for display within IGV. For each sample, we generated copy-number profiles using CNVnator (v.0.3.3) in 100-bp windows across all regions contained in the variant chains.

Retrogene insertions. Retrogene insertions were identified by examining the ultra-rare variant chains constructed as described above. For each chain, we identified any constituent SV with a reciprocal overlap of 90% to an intron using BEDTools (v.2.23.0). For each variant chain, the chain was deemed a retrogene insertion if it contained one or more BND variants with +/− strand orientation that overlapped an intron. In addition, we flagged any chains that contained non-BND SV calls, as their presence was indicative of a potential misclassification, and manually inspected them to determine whether they represented a true retrogene insertion.

Complex variants. We retained any cluster(s) incorporating three or more SV breakpoint calls, but removed SVs identified as retrogene insertions either during manual review or algorithmically. In addition, we excluded one call deemed to be a large, simple variant after manual review.

Large variants. Ultra-rare variants >1 Mb in length were selected and any overlap with identified complex variants identified and manually reviewed. Of five potential complex variants, one was judged to be a simple variant and included as a simple variant, whereas the rest were clearly complex variants and excluded. Gene overlap was determined as an overlap ≥ 1 bp with any exon occurring within protein-coding transcripts from Gencode v.27 marked as a principal isoform according to APPRIS[56].

Balanced translocations. Ultra-rare generic BND variants, of any confidence class, connecting two chromosomes and with support (>10%) from both strand orientations were initially considered as candidate translocations. We further filtered these candidates to require exactly two reported strand orientations indicating reciprocal breakpoints (that is, +−/−+, −+/+−, −−/++, ++/−−), no read support from any sample with a homozygous reference genotype, at least one split read supporting the translocation from samples containing the variant, and <25% overlap of either breakpoint with any simple repeat (downloaded from ftp://hgdownload.cse.ucsc.edu/goldenPath/hg38/database/simpleRepeat.txt.gz).

Comprehensive annotations from the Gencode v.27 GTF (ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_27/gencode.v27.annotation.gtf.gz) were used to determine the number of affected genes. A BED file of all introns was created by converting transcripts and exons to BED entries and subtracting all exons from their respective transcripts using BEDTools (v.2.23.0). To identify translocations affecting genes, the translocations were converted to BEDPE using svtools (v.0.3.1), padded by 1 bp and intersected with introns using BEDTools (v.2.23.0). The number of unique chromosome–gene name pairs for each translocation was used to determine the number of affected genes affected by each breakpoint.

To determine whether a translocation resulted in an in-frame fusion, we converted to BEDPE, padded by 1 bp and intersected the breakpoints with all introns using BEDTools (v.2.23.0). Each intron entry was then padded by 1 bp and intersected with the Gencode GTF file using BEDTools (v.2.23.0) and restricting to coding exons of the same transcript as the intron. Then, for each set of exons intersected by a given translocation, all combinations of transcripts were compared, taking into account their orientation and the orientation of the breakpoint, to determine whether the frame was maintained across the potentially fused exons. The resulting two candidate translocations were manually reviewed by reconstructing the transcript sequence of the fusion and translating the resulting DNA sequence using https://web.expasy.org/translate/ to confirm a single open-reading frame was maintained.

Generation of the B37 callset

Per-sample processing. This callset was constructed starting from a set of 8,455 individuals: 8,181 samples from 8 cohorts sequenced at the McDonnell Genome Institute, as well as 274 samples from the Simons Genome Diversity Project downloaded from EMBL-EBI (https://www.ebi.ac.uk/ena/data/view/PRJEB9586). All samples passed standard production quality control metrics and had a mean depth of coverage >20×. Data were aligned to GRCh37 using the SpeedSeq (v0.1.2) realignment pipeline. Per-sample SV calling was performed with SpeedSeq sv (v.0.1.2) using LUMPY (v.0.2.11), CNVnator-multi and SVTyper (v.0.1.4) on our local compute cluster. For LUMPY SV calling, we excluded high-copy-number outlier regions derived from >3,000 Finnish samples as described previously[1] (https://github.com/hall-lab/speedseq/blob/master/annotations/exclude.cnvnator_100bp.112015.bed).

Per-sample quality control. Following a summary of per-sample counts, samples with counts of any variant class (DEL, DUP, INV or BND) exceeding the median plus 10 times the median absolute deviation for that class were excluded from further analysis; 17 such samples were removed.

Merging. The remaining samples were processed into a single, joint callset using svtools (v.0.3.2) and the two-stage merging workflow (as described above): each of the nine cohorts was sorted and merged separately in the first stage, and the merged calls from each cohort sorted and merged together in the second stage.

Cohort-level re-genotyping. The resulting SV loci were then re-genotyped with SVTyper (v.0.1.4) and copy-number annotated using svtools (v.0.3.2) in parallel, followed by a combination of single-sample VCFs, frequency annotation and pruning using the standard workflow for svtools (v.0.3.2). A second round of re-genotyping with more-sensitive parameters to better reflect the expected allele balance for simple tandem duplications (https://github.com/ernfrid/regenotype/blob/master/resvtyper.py, commit 4fadcc4) was then performed, followed by another round of frequency annotation, pruning and finally reclassification using svtools (v.0.3.2) and the standard workflow.

Callset tuning and site-level filtering. Genotype calls for samples in 452 self-reported trios were extracted, and Mendelian error rates calculated using a custom R script; we counted as a Mendelian error any child genotype inconsistent with inheritance of exactly one allele from the mother and exactly one allele from the father. Filtering was performed as described for the B38 callset: INVs passed if: MSQ ≥ 150;

neither split-read nor paired-end LUMPY evidence made up <10% of total evidence; each strand provided at least >10% of read support. Generic BNDs passed if MSQ ≥ 250. SVs of length <50 bp were removed, according to our working definition of 'structural variation'.

Final sample-level filtering. Nine samples with retracted consents, and two hydatidiform mole samples were removed from the callset. Subsequently, the numbers of quality-control-passing, very rare (<0.1% MAF) DELs, DUPs and BNDs per sample were determined. Excluding the samples in the Simons Genome Diversity cohort (which were expected, in general, to have unusually high counts of rare variants), we determined the median and median absolute deviation (MAD) of the per-sample counts of each type, and excluded outlier samples with a count exceeding the median + 10 × MAD of any type. Nine samples were removed in this way. Finally, kinship was estimated using KING (v.2.0) based on high-quality, autosomal deletion and MEI calls with population allele frequency >1%. Each SV was annotated in the VCF according to the number of distinct, first-degree family clusters in which it was observed, as for the B38 callset.

PCA. A set of unrelated individuals (containing no first- or second-degree relatives) was extracted using KING (v.2.0). PCA was performed using smartpca (v.13050) on a VCF of all high-quality DEL and MEI variant calls with population allele frequency >1%. Eigenvectors were estimated based on the set of unrelated samples, and then all samples projected onto the eigenvectors.

Generation of the B38 SNV and indel callset and quality control

Per-sample calling was performed at the Broad Institute as part of CCDG joint-calling of 22,609 samples using GATK[57,58]. HaplotypeCaller v.3.5-0-g36282e4. All samples were joint-called at the Broad Institute using GATK v.4.beta.6, filtered for sites with an excess heterozygosity value of more than 54.69 and recalibrated using VariantRecalibrator with the following features: QD, MQRankSum, ReadPosRankSum, FS, MQ, SOR and DP. Individual cohorts were subset out of the whole CCDG callset using Hail v.0.2 (https://github.com/hail-is/hail). After SNV and indel variant recalibration, multi-allelic variants were decomposed and normalized with vt (v.0.5)[59]. Duplicate variants and variants with symbolic alleles were then removed. Afterwards, variants were annotated with custom computed allele balance statistics, 1000 Genomes Project allele frequencies[28], gnomAD-based population data[27], VEP (v.88)[60], CADD[29] (v.1.2) and LINSIGHT[30]. Variants having greater than 2% missingness were soft-filtered. Samples with high rates of missingness (>2%) or with mismatches between reported and genetically estimated sex (determined using PLINK v.1.90b3.45 sex-check) were excluded. The LOFTEE plug-in (v.0.2.2-beta; https://github.com/konradjk/loftee) was used to classify putative loss-of-function SNVs and indels as high or low confidence.

Annotation of gene-altering SV calls

The VCF was converted to BEDPE format using svtools vcftobedpe The resulting BEDPE file was intersected (using BEDTools (v.2.23.0) intersect and pairtobed) with a BED file of coding exons from Gencode v.27 with principal transcripts marked according to APPRIS[56]. The following classes of SV were considered as putative gene-altering events: (1) DEL, DUP, or MEI intersecting any coding exon; (2) INV intersecting any coding exon and with either breakpoint located within the gene body; and (3) BND with either breakpoint occurring within a coding exon.

Gene-based estimation of dosage sensitivity

We followed a previously described method[45], to estimate genic dosage sensitivity scores using counts of exon-altering deletions and duplications in a combined callset comprising the 14,623 sample pan-CCDG callset plus 3,172 non-redundant samples from the B37 callset. B37 CNV calls were lifted over to B38 as BED intervals using CrossMap (v.0.2.1)[61]. We determined the counts of deletions and duplications that intersect coding exons of principal transcripts of any autosomal gene. In the previous study[45], the expected number of CNVs per gene was modelled as a function of several genomic features (GC content, mean read depth and so on), some of which were relevant to their exome read-depth CNV callset but not to our WGS-based breakpoint mapping lumpy/svtools callset. To select the relevant features for prediction, using the same set of gene-level annotations as described previously[45], we restricted to the set of genes in which fewer than 1% of samples carried an exon-altering CNV, and used l^1-regularized logistic regression (from the R glmnet package[62], v.2.0-13), with the penalty λ chosen by tenfold cross-validation. The selected parameters (gene length, number of targets and segmental duplications) were then used as covariates in a logistic regression-based calculation of per-gene intolerance to DEL and DUP, similar to that described previously[45]. For deletions (or duplications, respectively), we restricted to the set of genes with <1% of samples carrying a DEL, to estimate the parameters of the logistic model. We then applied the fitted model to the full set of genes to calculate genic CNV intolerance scores as the residuals of the logistic regression of CNV frequency on the genomic features, standardized as z-scores and with winsorization of the lower 5th percentile.

Genome-wide estimation of deleterious variants

To estimate the relative numbers of deleterious SNVs, indels, DELs and DUPs genome-wide in the normal population, we relied on a subset of 4,298 samples from the B38 callset for which we had joint variant callsets for both SNVs/indels (GATK) and SVs (lumpy/svtools). Each SNV and indel was annotated with CADD[29] and LINSIGHT[30] scores as described above. CADD and LINSIGHT scores were converted to percentiles and singleton rates (where 'singleton' was defined as a variant present as a single allele) calculated for variants above each score threshold. CADD and LINSIGHT scores were then calibrated to a standard scale by matching singleton rates. Each DEL and DUP was annotated with CADD and LINSIGHT scores, calculated as the mean of the top 10 single-base CADD or LINSIGHT scores, respectively, for the span of the CNV (similar to SVScore[63]). The CNV-level CADD and LINSIGHT scores were then standardized using the above calibration curves. Finally, each variant (SNV, indel or CNV) was assigned a combined CADD–LINSIGHT score, calculated as the maximum of the two distinct scores.

The combined scores provided a means to rank, within each variant class, variants in order of deleteriousness. We calculated the singleton rate for the set of all LOFTEE high confidence protein-truncating SNVs and indels in autosomal genes. We then estimated the number of deleterious variants of each type genome-wide by choosing the combined CADD–LINSIGHT score threshold as the minimum value, such that the singleton rate for the set of higher-scoring variants was greater than or equal to the singleton-rate for LOFTEE high-confidence protein-truncating variants.

Annotation of noncoding elements

We divided the genome into 1-kb non-overlapping windows to investigate the rates of CNV occurrence relative to various classes of coding and noncoding elements, genome-wide. Windows intersecting assembly gaps or high-copy-number outlier regions (as described above) and windows with fewer than 50% of bases uniquely mappable as determined using GEM-mappability (build 1.315)[64] were excluded from analysis. BED tracks of genomic annotations for the noncoding dosage sensitivity analysis were created as described below.

The phastcons-20way[65] conservation track was downloaded from the UCSC genome browser (rsync://hgdownload.cse.ucsc.edu/goldenPath/hg38/phastCons20way/hg38.phastCons20way.wigFix.gz) and converted to bed format. The mean PhastCons score for each 1-kb window was calculated using BEDTools map. Quantiles of mean

window-level PhastCons scores were calculated and used as thresholds for the sensitivity analysis.

The LINSIGHT[30] score track was downloaded from CSHL (http://compgen.cshl.edu/LINSIGHT/LINSIGHT.bw). The 1-kb genomic windows were lifted over to hg19 using CrossMap (v.0.2.1), annotated with mean per-window LINSIGHT scores using BEDTools map and lifted back to GRChb38. Quantiles of mean window-level LINSIGHT scores were calculated and used as thresholds for the sensitivity analysis.

Genehancer[51] enhancers were downloaded from GeneCards (https://genecards.weizmann.ac.il/geneloc/index.shtml) and converted to bed format.

Vista[50] enhancers were downloaded from LBL (https://enhancer.lbl.gov/cgi-bin/imagedb3.pl?page_size=20000;show=1;search.result=yes;page=1;form=search;search.form=no;action=search;search.sequence=1), restricted to human enhancers, converted to bed format and lifted over to GRChb38 using CrossMap.

Encode[47] DNase hypersensitivity sites and transcription-factor-binding sites were downloaded from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegDnaseClustered/wgEncodeRegDnaseClusteredV3.bed.gz, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegTfbsClustered/wgEncodeRegTfbsClusteredV3.bed.gz) and lifted over to GRCHb38 using CrossMap.

Oreganno[66] literature-curated enhancers were downloaded from UCSC (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/oreganno.txt.gz) converted to bed format and lifted over to GRChb38 using CrossMap.

Sensitive[49], transcription-factor-bound, ultra-conserved[67] and HOT[68] regions were downloaded from the funseq2[69] resources (http://archive.gersteinlab.org/funseq2.1.0_data).

Dragon enhancers were downloaded from DENdb[70] (http://www.cbrc.kaust.edu.sa/dendb/src/enhancers.csv.zip), converted to bed format, lifted over to GRChb37 and filtered for score >2.

Chromatin interaction domains derived from Hi-C on human ES cell and IMR90 cells[71] were downloaded from http://compbio.med.harvard.edu/modencode/webpage/hic/, and distances between adjacent topological domains were calculated with BEDTools. When the physical distance between adjacent topological domains was <400 kb, these were classified as TAD boundaries; otherwise, they were classified as unorganized chromatin. The TAD boundaries and unorganized chromatin data were converted to bed format and lifted over to GRCh38 using CrossMap.

Roadmap chromatin state segmentations for 127 epigenomes were downloaded from Roadmap[48] (https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/) and lifted over to GRCh38. BEDTools multiinter was used to determine the number of epigenomes in which each segment was present.

Dosage sensitivity of noncoding elements

To maximize power, DEL and DUP calls from the non-redundant combination of the B37 and B38 callsets (as described above) were used for this analysis. Each window was further characterized by its distance to the nearest exon (the minimum distance between any point in the window and any point in the exon) and the pLI score of the gene corresponding to the nearest exon. The pLI score was set to zero for genes with pLI undefined. In the event that exons of two genes were equidistant to the window, the max of the two pLI scores was selected.

For a given SV type (DUP or DEL) and a given functional annotation (for example, VISTA enhancers), each window was characterized by the presence or absence of one or more SV and the presence or absence of one of more genomic features. We observed a depletion of CNVs in windows near exons, and in particular near exons of loss-of-function-intolerant genes (Fig. 4a). As such, we used a Cochran–Mantel–Haenszel estimate of the odds ratios for each SV type or functional annotation, while stratifying for the proximity to the nearest exon as well as that exon's loss-of-function intolerance score (pLI). Because adjacent windows are not strictly independent observations—that is, CNV or features may overlap adjacent windows, inducing some spatial correlations—we used a block bootstrap method (resampling was performed on blocks of 10 windows) to estimate robust confidence intervals.

Long-read validation

PacBio long-read sequences from nine 1000 Genomes Project (1KG) samples sequenced to deep coverage (>68-87x) at the McDonnell Genome Institute were used as an orthogonal means of validating SV calls.These PacBio data are available in SRA (see accessions in Supplementary Table 2) and were generated independently from the long-read data used by the Human Genome Structural Variation Consortium (HGSVC) to create the long-read SV callset used for sensitivity analyses described below[26]. The long-read sequences were aligned to GRCh38 using minimap2 (ref. [72]) (v.2.16-r922; parameters -ax map-pb). Split-read alignments indicating putative SVs were converted to BEDPE format[55] as described previously[23,24,73]. Similarly, deletions or insertions longer than 50 bp contained within PacBio reads (as determined based on the cigar strings) were converted to BEDPE format. We used BEDTools to judge the overlap between short-read SV calls and the long-read alignments. We judged an SV call to be validated when ≥ 2 long-reads exhibited split-read mappings in support of the SV call. For a long-read mapping to support an SV call, we required that it must predict a consistent SV type (for example, deletion) and exhibit substantial physical overlap with the SV call, where overlap can be met by either of the following criteria: (1) the two breakpoint intervals predicted by the SV call and the two breakpoint intervals predicted by the long-read split-read mapping overlap with each other on both sides, as determined by BEDTools pairtopair using 100 bp of "slop" (-type -is both -slop 100); or (2) the SV call and the long-read split-read (or cigar-derived indel variant) exhibit 90% reciprocal overlap with one another (BEDTools intersect -r -f 0.9). The above criteria for SV validation based on long-read support were selected based on extensive manual review of SV calls in the context of supporting data including read-depth profiles and long-read mappings from all nine samples, and are the basis for the validation rates reported in the main text and in Supplementary Table 3. However, we also show the range of validation rates that are obtained when using more lenient or strict measures of physical overlap, and when requiring a varying number of supporting PacBio reads (Extended Data Fig. 5), in both carriers and non-carriers of SVs from various classes. We also note that 3 of the 6 singleton SV calls that are not validated by long reads appear to be true variants based on manual review of read-level evidence, in which it appears that long-reads failed to validate true short-read SV calls owing to subtle differences in how coordinates were reported at local repeats. Our false discovery rate estimates may be conservative owing to these effects.

To conduct a comparison to HGSVC using the three samples shared between our datasets (NA19240, HG00514, HG00733), all non-reference, autosomal SV calls for each of the three samples were extracted from the CCDG B38 and HGSVC[26] Illumina short-read callsets. For HGSVC variants detected solely by read-depth analysis, for which genotype information was not available, a variant was defined to be non-reference if its predicted copy-number differed from the mode for that site across the nine samples in that callset (which includes the parents of NA19240, HG00514 and HG00733). The short-read calls from our study and HGSVC for the three relevant samples were converted to BEDPE format using svtools vcftobedpe. The three single-sample VCFs from the HGSVC PacBio long-read SV callset were converted to BEDPE format in similar fashion. For HGSVC Illumina calls (which had been taken from a callset comprising three trios, rather than a large cohort) variants were classified as rare if seen in only one of the six trio founders and either absent from or observed at frequency <1% in the 1KG phase 3 SV callset.

Long-read SV truth set construction

To evaluate the sensitivity of our callset, we constructed a high-confidence truth set from the comprehensive HGSVC long-read SV callset created using reference-guided de novo assembly[26]. The assembly-based long-read truth set includes all autosomal SVs reported by HGSVC[26] that were also validated by split-read alignments from the PacBio data generated independently at our centre. Here, an HGSVC call was judged to be validated by long-read data when two or more long reads exhibited split-read mappings or cigar-derived SV calls that match the HGSVC call in terms of the predicted SV type and breakpoint intervals, allowing 100 bp of "slop" to account for positional uncertainty (BEDTools pairtopair -type -is both -slop 100). To account for the variant classification scheme of the HGSVC callset—which only has two variant categories, INS and DEL—we allowed INS variants to be validated by long reads suggesting either insertion or tandem duplication variants. Variants were classified as STRs if either >50% of sequence from both reported breakpoint intervals or >50% of sequence contained in the outer span of the variant overlapped a GRCh38 track of simple repeats downloaded from the UCSC Table Browser. The interval spanned by each variant was converted to bed format and lifted over to hg19 using CrossMap. A combined CADD–LINSIGHT score was calculated for each variant based on the mean of the top 10 CADD-scoring and the mean of the top 10 LINSIGHT-scoring positions, as described in the section 'Genome-wide estimation of deleterious variants'.

Lifting over of the 1KG phase3 SV callset

The 1KG phase 3 SV callset was lifted over from GRCh37 to GRCh38 by first converting to BEDPE format using svtools vcftobedpe. The outer span of each variant was then converted to bed format and lifted over using CrossMap[61]. For SVs that were not lifted over as contiguous intervals, discontiguous regions within 1 kb were merged using BEDTools merge, and the largest of the merged variants were selected. The lifted-over bed interval was then converted back to BEDPE by padding each endpoint with 100 bp.

Assessment of sensitivity using the HGSVC long-read truth set

Sensitivity of the CCDG B38 and HGSVC Illumina short-read callsets to detect variants in the HGSVC long-read truth set was determined by converting each single-sample VCF to BEDPE format using svtools vcftobedpe and calculating overlaps using BEDTools pairtopair, allowing for 100 bp of "slop". For DEL calls, a variant was considered to be detected only if both breakpoints overlapped, and the type of the overlapping call was consistent with a deletion (that is, DEL, MEI, CNV or BND). For INS calls in the long-read callset, variants were considered to be detected if either breakpoint overlapped and the overlapping call was consistent with an insertion (that is, DUP, INS, CNV, MEI or BND).

Comparison with the 1KG phase 3 SV callset[5] necessitated the use of a slightly different sensitivity metric, as 1KG analysed the parents of HG00733 and NA19249, but not the trio offspring themselves. As, with rare exception, germline variants present in the child should also be present in one of the two parents, the rate at which HGSVC long-read calls in the truth set were detected in at least one parent in each of the CCDG B38, HGSVC and 1KG callsets serves as an informative alternate measure of 'sensitivity'.

Genotype comparison to 1KG

Genotype comparisons were performed for the five parental samples (NA19238, NA19239, HG00513, HG00731 and HG00732) present in both the CCDG B38 and the 1KG phase 3 SV callsets. Each callset was subset (using BCFtools) to the set of autosomal SVs with a non-reference call in at least one of the five parental samples and converted to BEDPE format. Variants in the 1KG callset detected using read-depth methods only were excluded. Bedtools pairtopair (100 bp slop, overlap at both breakpoints) was used to determine the set of variants called in both the five-sample CCDG callset and the five-sample 1KG callset, requiring consistent SV type. For each variant site in each sample, genotypes from the two callsets were compared. Results were tallied, and concordance rates and kappa statistics ('irr' package) were calculated in R.

Pedigree analysis

Pedigree analyses were performed on three-generation pedigrees from Utah collected as part of the Centre d'Étude du Polymorphisme Humain (CEPH) consortium. The analyses used a set of 576 CEPH samples contained in the B37 callset that remained after excluding 21 samples that had been deemed low-quality and/or possibly contaminated based on analysis of a SNV–indel callset (data not shown). The remaining samples comprise 409 trios, which were used in the estimation of transmission rates. The counts of all high-quality SVs called heterozygous in one parent, homozygous reference in the other and non-missing in the offspring were used to estimate transmission rates by frequency class, with Wilson score confidence intervals calculated using R binconf.

Mendelian errors for all high-quality (filter = PASS) SVs were calculated using PLINK (v.1.90b3.45), with the output restricted to variant-trio observations in which all three genotypes (father, mother and offspring) were non-missing. For each sample in the third generation (the 'F_2'; see Extended Data Fig. 6a) of any of the CEPH kindreds, Mendelian errors were counted by frequency class. The Mendelian error rate was calculated as the total number of Mendelian errors divided by the total number of non-reference, non-missing genotypes in F_2 generation samples for variants of that frequency class. De novo variants were defined as variants private to a single family in which both parental genotypes are 0/0 and the offspring genotype either 0/1 or 1/1, and were obtained by parsing the PLINK output. (Note that these variant counts are used as callset quality metrics and do not necessarily represent true de novo mutations.)

Transmission rates for putative de novo variants were calculated by restricting to all high-quality autosomal variants heterozygous in a second generation ('F_1') sample and homozygous reference in both of his/her parents ('P_0' generation) and his/her F_1 spouse. Each such variant was classified as transmitted if carried by any F_2 offspring, with transmission rates calculated as the number of transmitted variants out of the total. 'Missed heterozygous calls' were counted as the set of all family-private variants non-reference in at least two F_2 offspring siblings, but homozygous reference in both of the F_1 parents. The rate of missed heterozygous calls was calculated by dividing this count by the total count of family-private variants carried by at least two F_2 offspring siblings.

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The sequencing data can be accessed through dbGaP (https://www.ncbi.nlm.nih.gov/gap) under the accession numbers provided in Supplementary Table 7. PacBio long-read data used for SV validation can be accessed through the Sequence Read Archive (SRA), under the accession numbers provided in Supplementary Table 2. The set of high-confidence HGSVC long-read-derived SV calls, validated by our independent PacBio data and used as a truth set, can be found in Supplementary File 3. Supplementary Files 1–4 can be found at https://github.com/hall-lab/sv_paper_042020.

Code availability

Custom code used in the long-read validation can be found here: https://github.com/abelhj/long-read-validation/tree/master.

53. 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).
54. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
55. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
56. Rodriguez, J. M. et al. APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res.* **41**, D110–D117 (2013).
57. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
58. Poplin, R. et al. Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at bioRxiv https://www.biorxiv.org/content/10.1101/201178v3 (2018).
59. Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants. *Bioinformatics* **31**, 2202–2204 (2015).
60. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
61. Zhao, H. et al. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).
62. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
63. Ganel, L., Abel, H. J. & Hall, I. M. SVScore: an impact prediction tool for structural variation. *Bioinformatics* **33**, 1083–1085 (2017).
64. Derrien, T. et al. Fast computation and applications of genome mappability. *PLoS One* **7**, e30377 (2012).
65. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
66. Griffith, O. L. et al. ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res.* **36**, D107–D113 (2008).
67. Bejerano, G. et al. Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
68. Yip, K. Y. et al. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.* **13**, R48 (2012).
69. Fu, Y. et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* **15**, 480 (2014).
70. Ashoor, H., Kleftogiannis, D., Radovanovic, A. & Bajic, V. B. DENdb: database of integrated human enhancers. *Database)* **2015**, bav085 (2015).
71. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
72. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
73. Faust, G. G. & Hall, I. M. YAHA: fast and flexible long-read alignment with optimal breakpoint detection. *Bioinformatics* **28**, 2417–2424 (2012).

Acknowledgements We thank staff at the NHGRI for supporting this effort. This study was funded by NHGRI CCDG awards to Washington University in St Louis (UM1 HG008853), the Broad Institute of MIT and Harvard (UM1 HG008895), Baylor College of Medicine (UM1 HG008898) and New York Genome Center (UM1 HG008901); an NHGRI GSP Coordinating Center grant to Rutgers (U24 HG008956); and a Burroughs Wellcome Fund Career Award to I.M.H. Additional data production at Washington University in St Louis was funded by a separate NHGRI award (5U54HG003079). We thank S. Sunyaev for comments on the manuscript; T. Teshiba for coordinating samples for FINRISK and EUFAM sequencing; and the staff and participants of the ARIC study for their contributions; and we acknowledge all individuals who were involved in the collection of samples that were analysed for this study. Data production for EUFAM was funded by 4R01HL113315-05; the Metabolic Syndrome in Men (METSIM) study was supported by grants to M. Laakso from the Academy of Finland (no. 321428), the Sigrid Juselius Foundation, the Finnish Foundation for Cardiovascular Research, Kuopio University Hospital and the Centre of Excellence of Cardiovascular and Metabolic Diseases supported by the Academy of Finland; data collection for the CEPH pedigrees was funded by the George S. and Dolores Doré Eccles Foundation and NIH grants GM118335 and GM059290; study recruitment at Washington University in St Louis was funded by the DDRCC (NIDDK P30 DK052574) and the Helmsley Charitable Trust; study recruitment at Cedars-Sinai was supported by the F. Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, NIH/NIDDK grants P01 DK046763 and U01 DK062413 and the Helmsley Charitable Trust; study recruitment at Intermountain Medical Center was funded by the Dell Loy Hansen Heart Foundation; the Late Onset Alzheimer's Disease Study (LOAD) study was funded by grants to T. Foroud (U24AG021886, U24AG056270, U24AG026395 and R01AG041797); the Atherosclerosis Risk in Communities (ARIC) study was funded by the NHLBI (HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700004I and HHSN268201700005I); and the PAGE programme is funded by the NHGRI with co-funding from the NIMHD (U01HG007416, U01HG007417, U01HG007397, U01HG007376 and U01HG007419). Samples from the BioMe Biobank were provided by The Charles Bronfman Institute for Personalized Medicine at the Icahn School of Medicine at Mount Sinai. The Hispanic Community Health Study/Study of Latinos was carried out as a collaborative study supported by the NHLBI (N01-HC65233, N01-HC65234, N01-HC65235, N01-HC65236 and N01-HC65237), with contributions from the NIMHD, NIDCD, NIDCR, NIDDK, NINDS and NIH ODS. The Multiethnic Cohort (MEC) study is funded through the NCI (R37CA54281, R01 CA63, P01CA33619, U01CA136792 and U01CA98758). For the Stanford Global Reference Panel, individuals from Puno, Peru were recruited by J. Baker and C. Bustamante, with funding from the Burroughs Welcome Fund, and individuals from Rapa Nui (Easter Island) were recruited by K. Sandoval Mendoza and A. Moreno Estrada, with funding from the Charles Rosenkranz Prize for Health Care Research in Developing Countries. The Women's Health Initiative (WHI) programme is funded by the NHLBI (HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C and HHSN271201100004C). The GALA II study and E. G. Burchard are supported by the Sandler Family Foundation, the American Asthma Foundation, the RWJF Amos Medical Faculty Development Program, the Harry Wm. and Diana V. Hind Distinguished Professor in Pharmaceutical Sciences II, the NHLBI (R01HL117004, R01HL128439, R01HL135156 and X01HL134589), the NIEHS (R01ES015794, R21ES24844), the NIMHD (P60MD006902, R01MD010443, RL5GM118984) and the Tobacco-Related Disease Research Program (24RT-0025). We acknowledge the following GALA II co-investigators for recruitment of individuals, sample processing and quality control: C. Eng, S. Salazar, S. Huntsman, D. Hu, A. C.Y. Mak, L. Caine, S. Thyne, H. J. Farber, P. C. Avila, D. Serebrisky, W. Rodriguez-Cintron, Jose R. Rodriguez-Santana, R. Kumar, L. N. Borrell, E. Brigino-Buenaventura, A. Davis, M. A. LeNoir, K. Meade, S. Sen and F. Lurmann, and we thank the staff and participants who contributed to the GALA II study.

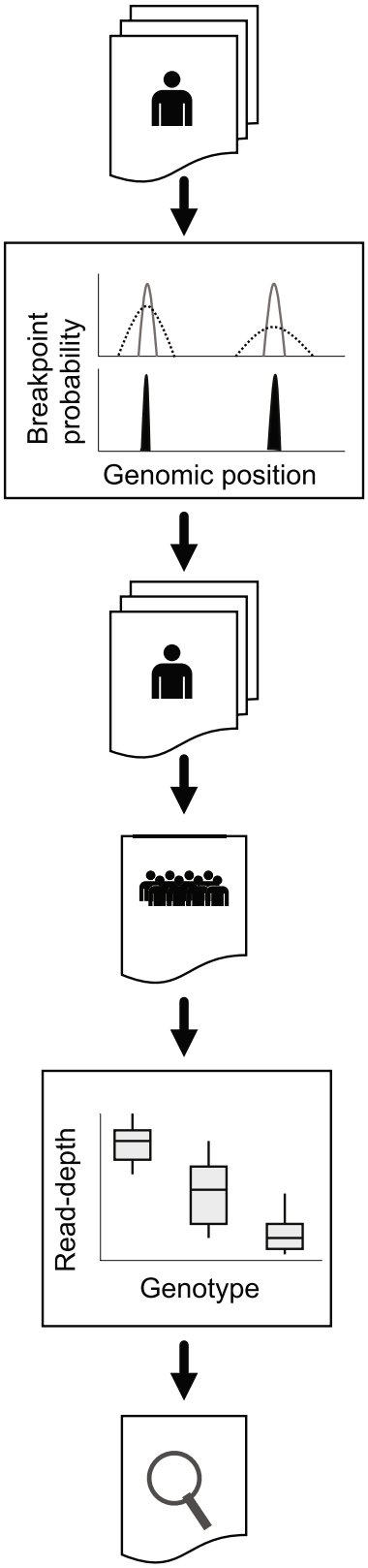
Author contributions I.M.H. conceived and directed the study. D.E.L. and H.J.A. developed the final version of the SV calling pipeline, constructed the SV callsets and performed the data analyses. C.C. and R.M.L. helped design the SV calling pipeline. A.A.R. contributed to long-read validation. I.D. was instrumental in the migration of workflows to the Google Cloud Platform. K.L.K. assisted with data management. E.S.L., B.M.N. and N.O.S. provided input on population genetic analyses. W.J.S., D.M.M., E.S.L., B.M.N., M.C.Z., C.R., T.C.M., S.B., S.K.D., I.M.H. and N.O.S. directed data production, processing and management at their respective sites, and edited the manuscript. Members of the NHGRI CCDG consortium provided samples, produced sequencing data and coordinated and administered data-sharing efforts. H.J.A., D.E.L. and I.M.H. wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2371-0.

Correspondence and requests for materials should be addressed to I.M.H.



Per-sample SV discovery

Merging and breakpoint refinement

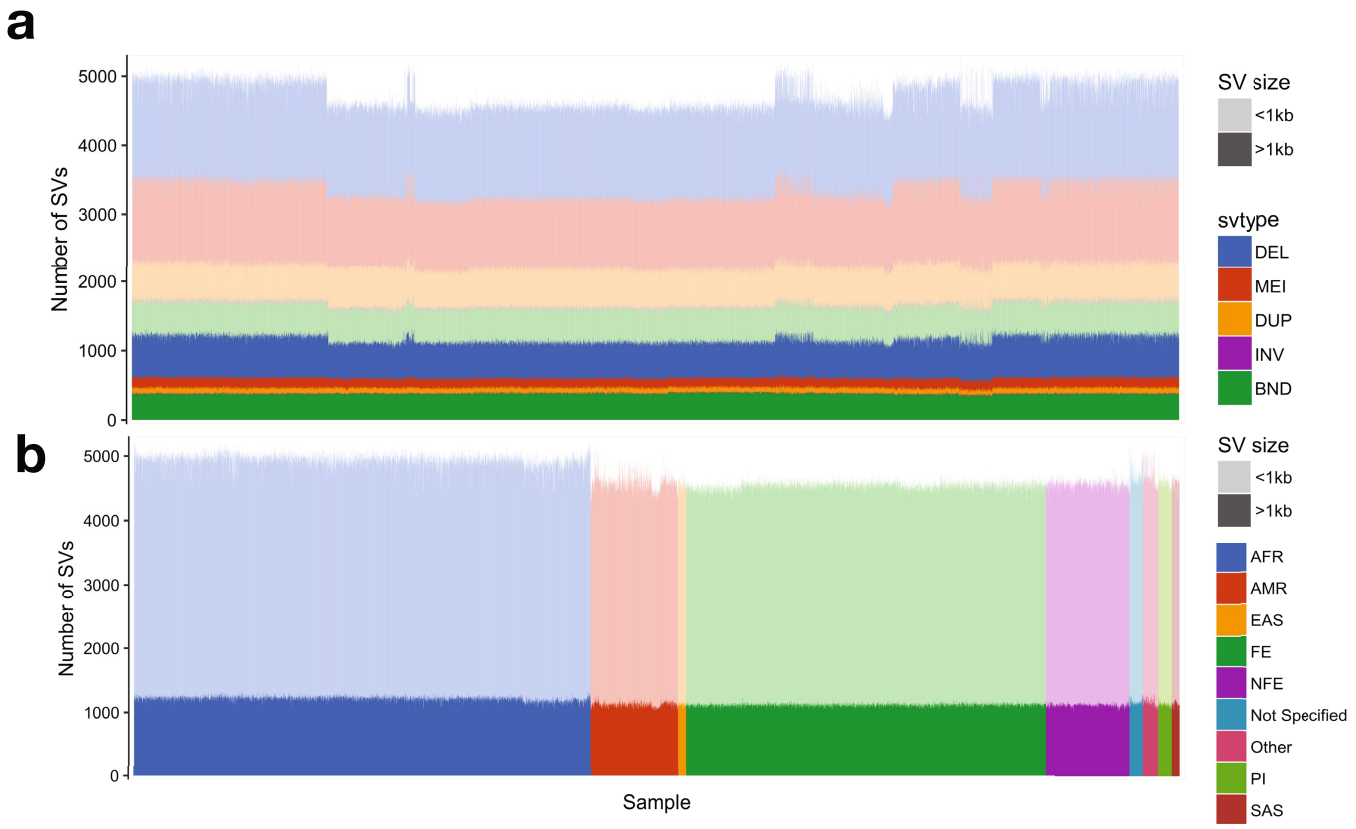
Per-sample re-genotyping and copy number

Cohort-level VCF creation

Type reclassification and genotype refinement

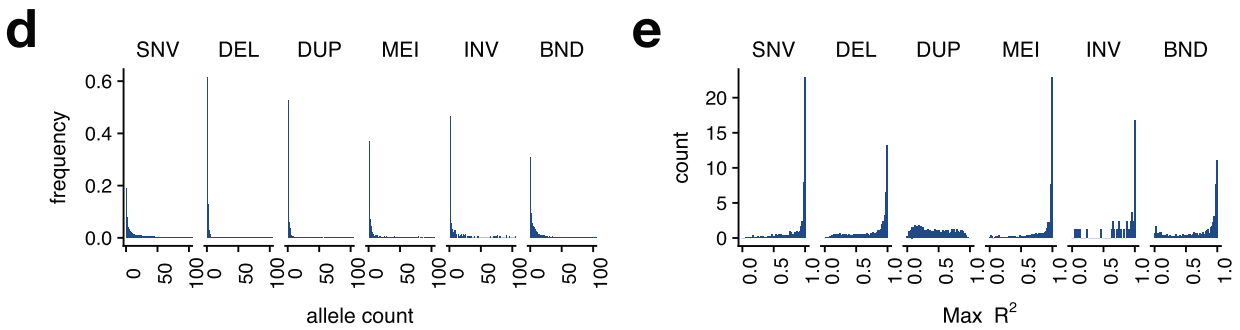
QC and Confidence Scoring

Extended Data Fig. 1 | SV mapping pipeline. SVs are detected within each sample using LUMPY. Breakpoint probability distributions are used to merge and refine the position of detected SVs within a cohort, followed by parallelized re-genotyping and copy-number annotation. Samples are merged into a single cohort-level VCF file, variant types reclassified and genotypes refined with svtools using the combined breakpoint genotype and read-depth information. Finally, sample-level quality control (QC) and variant confidence scoring is conducted to produce the final callset.

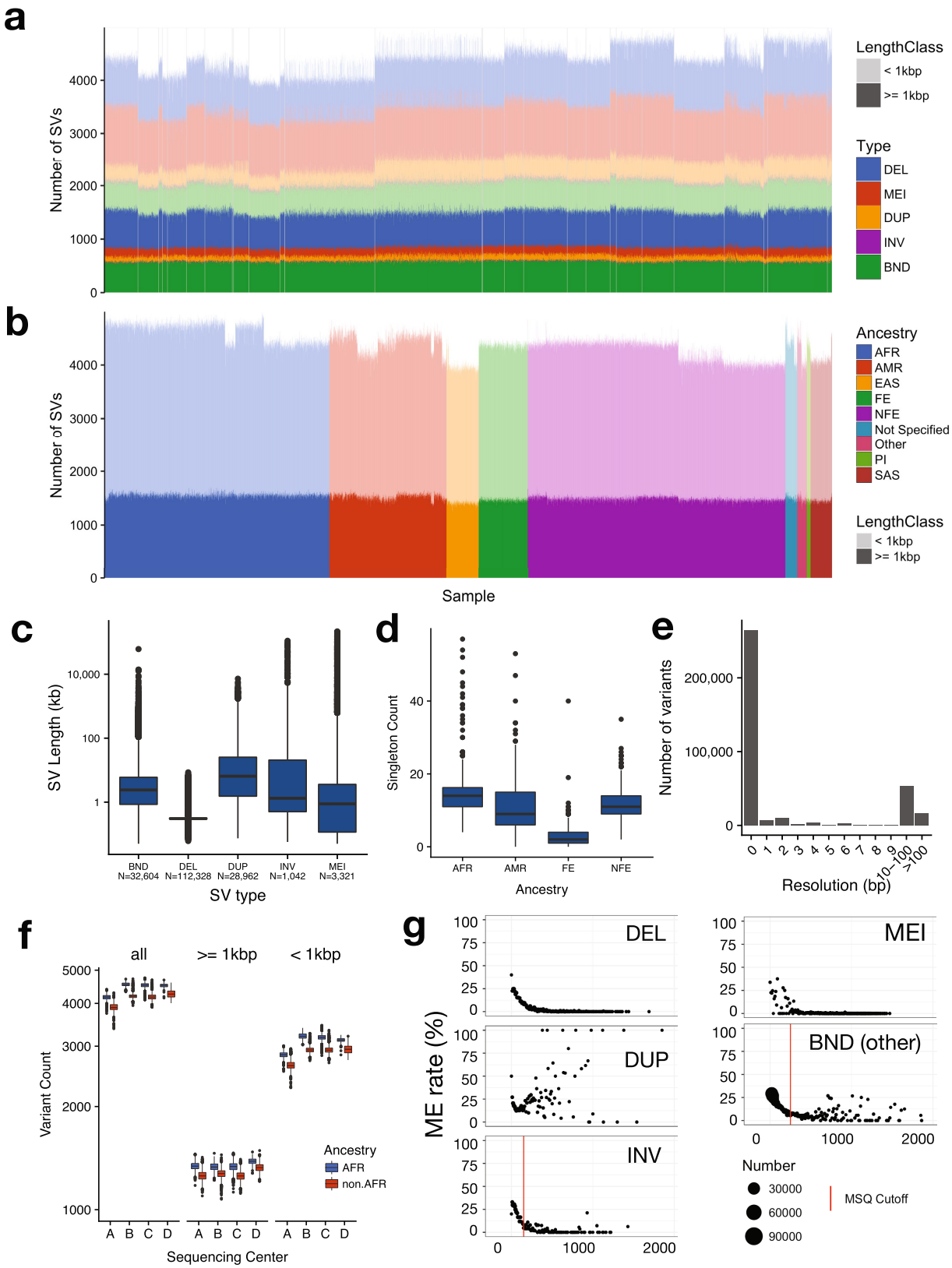


c

Type	Variant counts				Mendelian Error Rate
	Ultra-Rare	Rare	Low Freq	Common	
DEL	49322	25054	2831	4562	0.041
DUP	9056	4276	542	1344	0.113
MEI	621	1923	268	1960	0.009
INV	343	247	35	36	0.009
Low-confidence INV	828	1528	455	874	0.082
BND	7064	7782	465	1242	0.022
Low-confidence BND	29787	87056	20606	21199	0.14

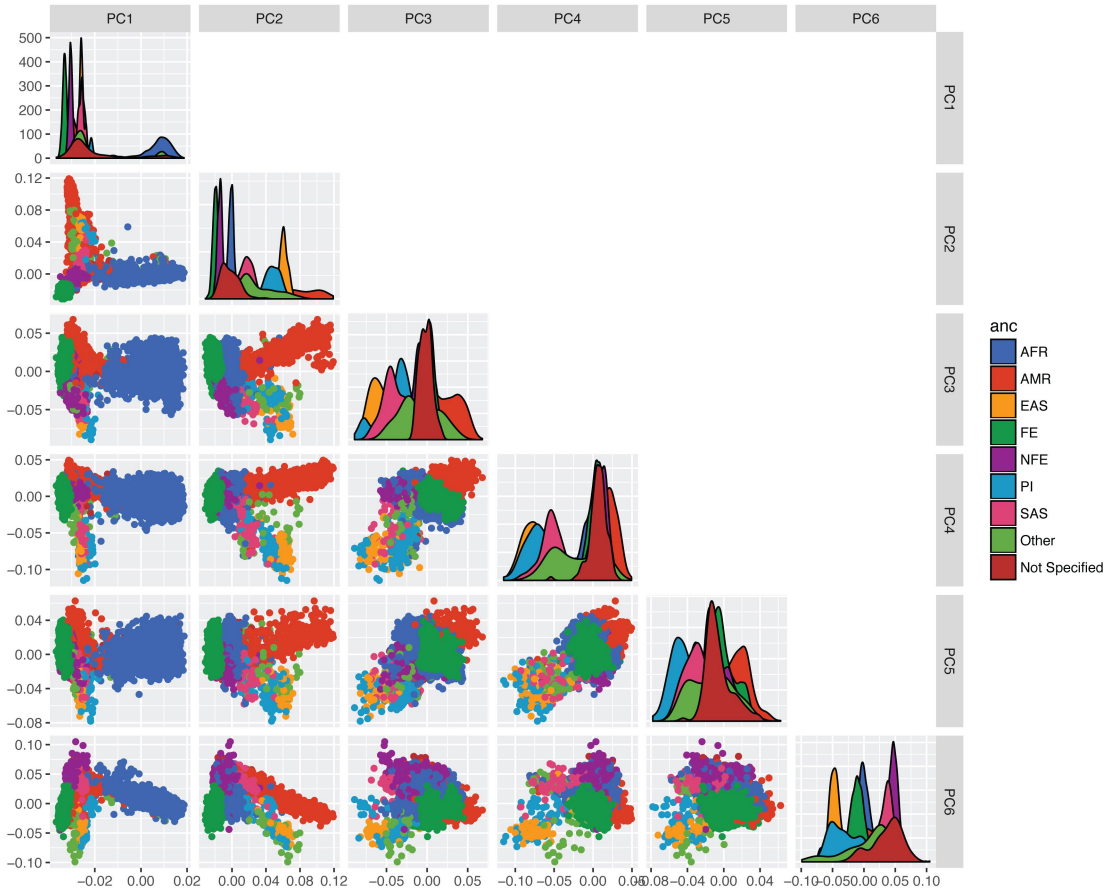


Extended Data Fig. 2 | The B37 callset. a, Variant counts (*y* axis) for each sample (*x* axis) in the callset, ordered by cohort. Large (>1 kb) variants are shown in dark shades and smaller variants in light shades. **b**, Variant counts per sample, ordered by self-reported ancestry according to the colour scheme on the right. Abbreviations as in Fig. 1a. Note that African-ancestry samples show more variant calls, as expected. **c**, Table showing the number of variant calls by variant type and frequency class, and Mendelian error rate by variant type. **d**, Histogram of allele count for each variant class, showing alleles with counts ≤ 100. **e**, Linkage disequilibrium of each variant class as represented by maximum R^2 value to nearby SNVs, for $n = 1{,}581$ samples. Note that these distributions mirror those from our previous SV callset for GTEx[4], which was characterized extensively in the context of expression quantitative trait loci.

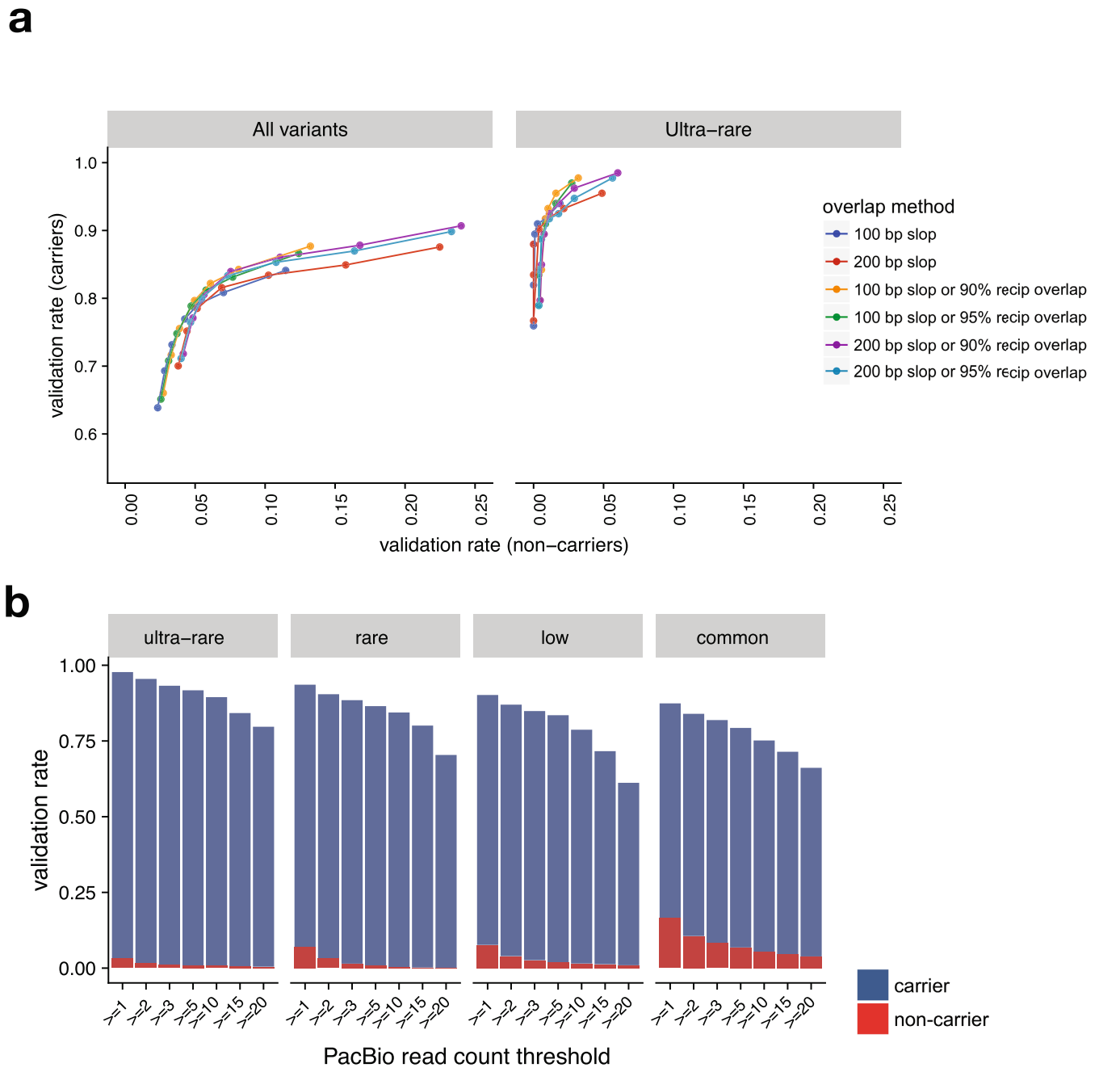


Extended Data Fig. 3 | See next page for caption.

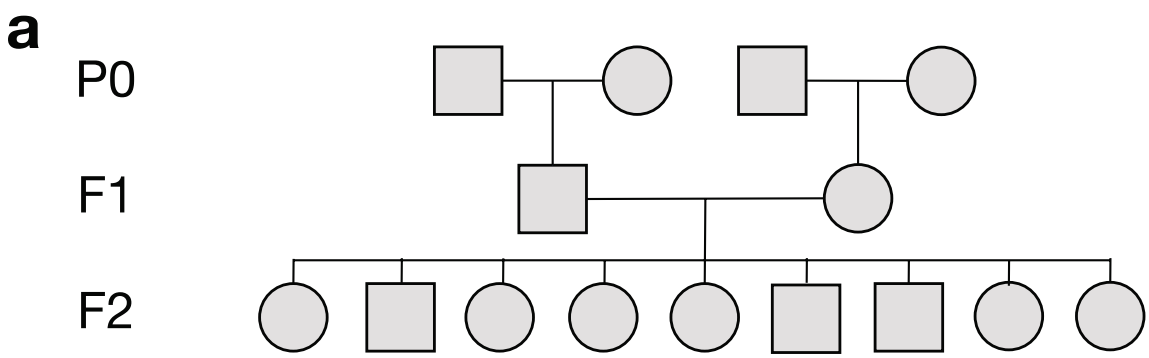
Extended Data Fig. 3 | The B38 callset. a, Variant counts (*y* axis) for each sample (*x* axis) in the callset, ordered by cohort. Large (>1 kb) variants are shown in dark shades and smaller variants in light shades. **b**, Variant counts per sample, ordered by self-reported ancestry according to the colour scheme on the right. Abbreviations as in Fig. 1a. Note that African-ancestry samples show more variant calls, as expected. Note also that there is some residual variability in variant counts owing to differences in data from each sequencing centre, but that this is mainly limited to small tandem duplications (see **a**), primarily at STRs. **c**, SV length distribution by variant class. **d**, Distribution of the number of singleton SVs detected in samples from different ancestry groups. Only groups with ≥1,000 samples in the B38 callset are shown, and each group was subsampled down to 1,000 individuals before recalculation of the allele frequency. **e**, Histogram showing the resolution of SV breakpoint calls, as defined by the length of the 95% confidence interval of the breakpoint-containing region defined by LUMPY, after cross-sample merging and refinement using svtools. Data are from $n = 360{,}614$ breakpoints, 2 per variant. **f**, Distribution of the number of SVs detected per sample in WGS data from each sequencing centre (*x* axis) for African and non-African (non-AFR) samples, showing all variants (left), and those larger (middle) and smaller (right) than 1 kb in size. Per-centre counts are as follows: centre A, 1,527 AFR, 2,080 non-AFR; centre B, 408 AFR, 2,745 non-AFR; centre C, 2,953 AFR, 2226 non-AFR; centre D, 150 AFR, 2,534 non-AFR. **g**, Plots of Mendelian error (ME) rate (*y* axis) by MSQ for each variant class. Dot size is determined by point density (right) and the threshold used to determine high and low confidence SVs are shown by the vertical lines. All box plots in indicate the median (centre line) and the first and third quartiles (box limits); whiskers extend 1.5 × IQR.

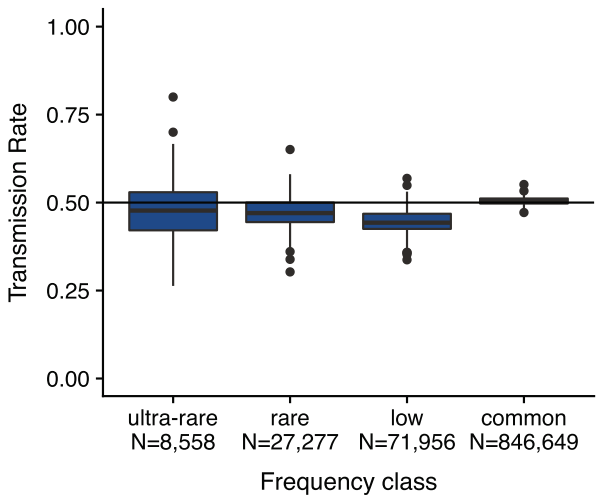


Extended Data Fig. 4 | PCA for the B37 callset. PCA was performed using a linkage disequilibrium-pruned subset of high-confidence DEL and MEI variants, with MAF > 1%. Self-reported ancestry is shown using the colour scheme on the right, with abbreviations as in Fig. 1a.



Extended Data Fig. 5 | Validation of SV calls by PacBio long reads in nine control samples. $n = 9{,}905$ variants. **a**, Validation rates in variant carriers (*y* axis) versus validation rates in non-carriers (that is, false validations; *x* axis) for each method of determining variant overlap, for a range of supporting-read-count thresholds. Ultra-rare variants ($n = 133$) are shown separately on the right. For each variant overlap method, each data point represents a distinct read-count threshold (≥1, 2, 3, 5, 10, 15 or 20 PacBio reads) that was used to determine validation of SV calls by long-read alignments. Two methods were used for determining overlap between SV coordinates and long-read alignments while accounting for positional uncertainty: (1) BEDTools pairtopair, requiring overlap between the pair of breakpoint intervals predicted by short-read SV mapping and the pair of breakpoint intervals predicted by long-read alignment, allowing 100 bp or 200 bp of 'slop'; and (2) BEDTools intersect, requiring 90% or 95% reciprocal overlap between the coordinates spanned by the SV predicted by short-read SV mapping and the SV predicted by long-read alignment. Here, we plot the first criteria by themselves, and in pairwise combination with the latter (see key on the right of the figure). Note that Supplementary Table 3 is based on the '100 bp slop or 90% reciprocal overlap' method, requiring at least two PacBio reads. **b**, Validation rates by frequency class for variant carriers and non-carriers with increasing PacBio supporting-read thresholds, shown using the same overlap method as in Supplementary Table 3. Variant counts per frequency class are as follows: ultra-rare, $n = 133$; rare, $n = 734$; low frequency, $n = 1{,}361$; common, $n = 7{,}677$.





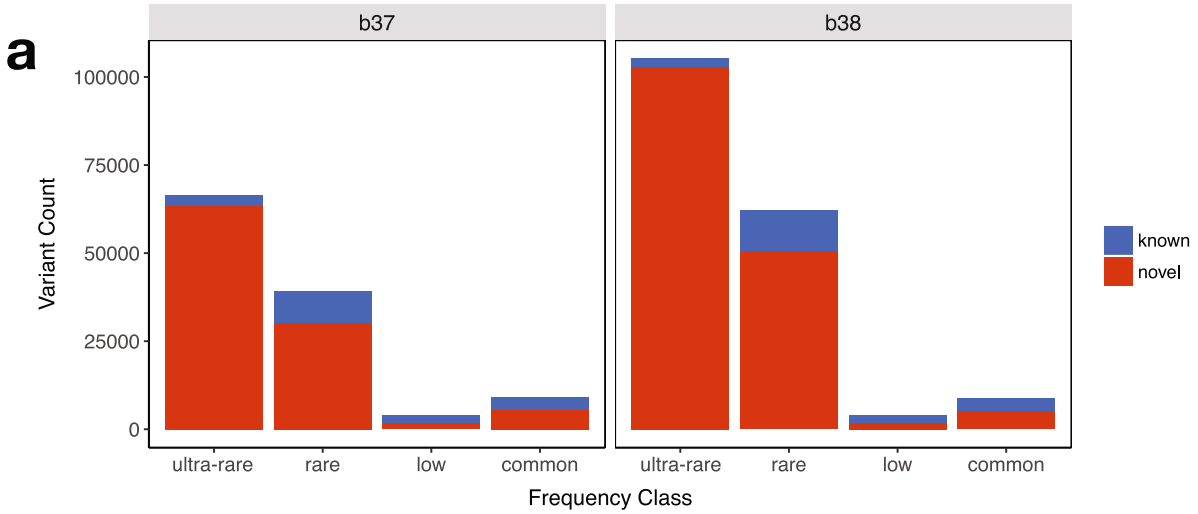
c

Frequency bin	total count	Mendelian errors	Mendelian error rate
common	1,446,555	45,553	0.031
low	31,612	2,927	0.093
rare	11,105	480	0.043
ultra-rare	3,441	82	0.024

d

SV type	Mendelian errors	Total count	Mendelian error rate
DEL	55	2,291	0.024
DUP	21	495	0.042
INV	0	7	0.000
MEI	4	4	1.000
BND	2	644	0.003
All	82	3,441	0.024

Extended Data Fig. 6 | Mendelian inheritance analysis in a set of three-generation CEPH pedigrees comprising 409 parent–offspring trios. **a**, Example structure of a single CEPH pedigree indicating nomenclature of the parental (P_0), first (F_1) and second (F_2) generations. **b**, Transmission rate of SVs from different allele-frequency classes including SVs that are unique to a single family (ultra-rare), rare (<1%), low frequency ('low'; 1–5%) and common (>5%). **c**, Table showing the number and rate of Mendelian errors by allele-frequency class. **d**, Table showing the number and rate of Mendelian errors for SVs that are unique to a single family, for each SV type.

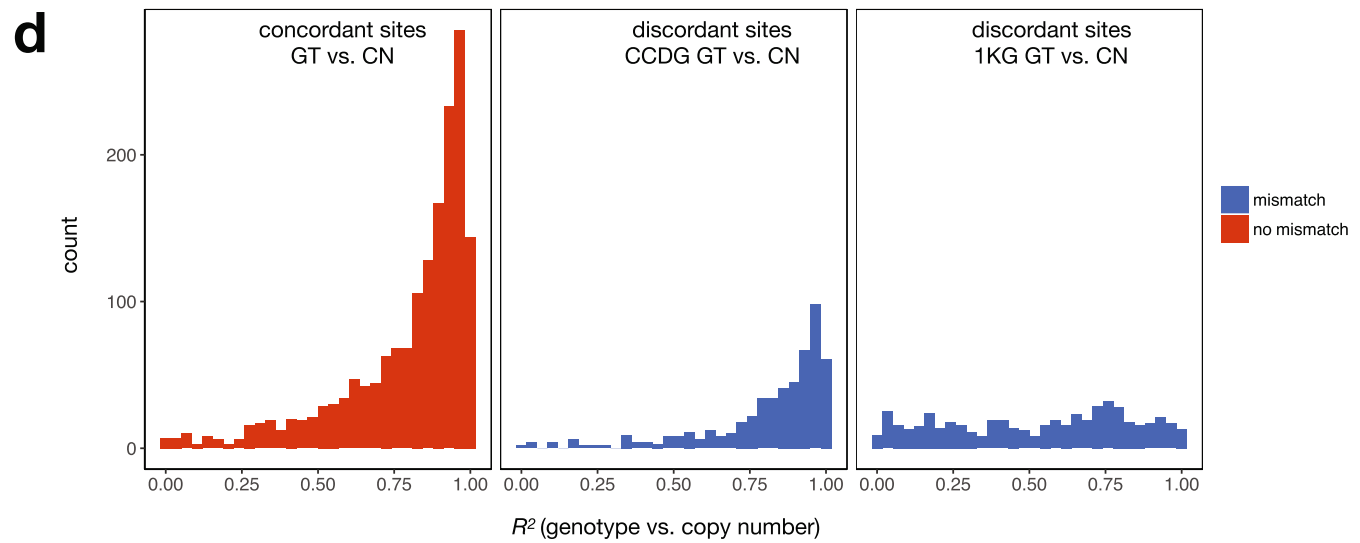


b

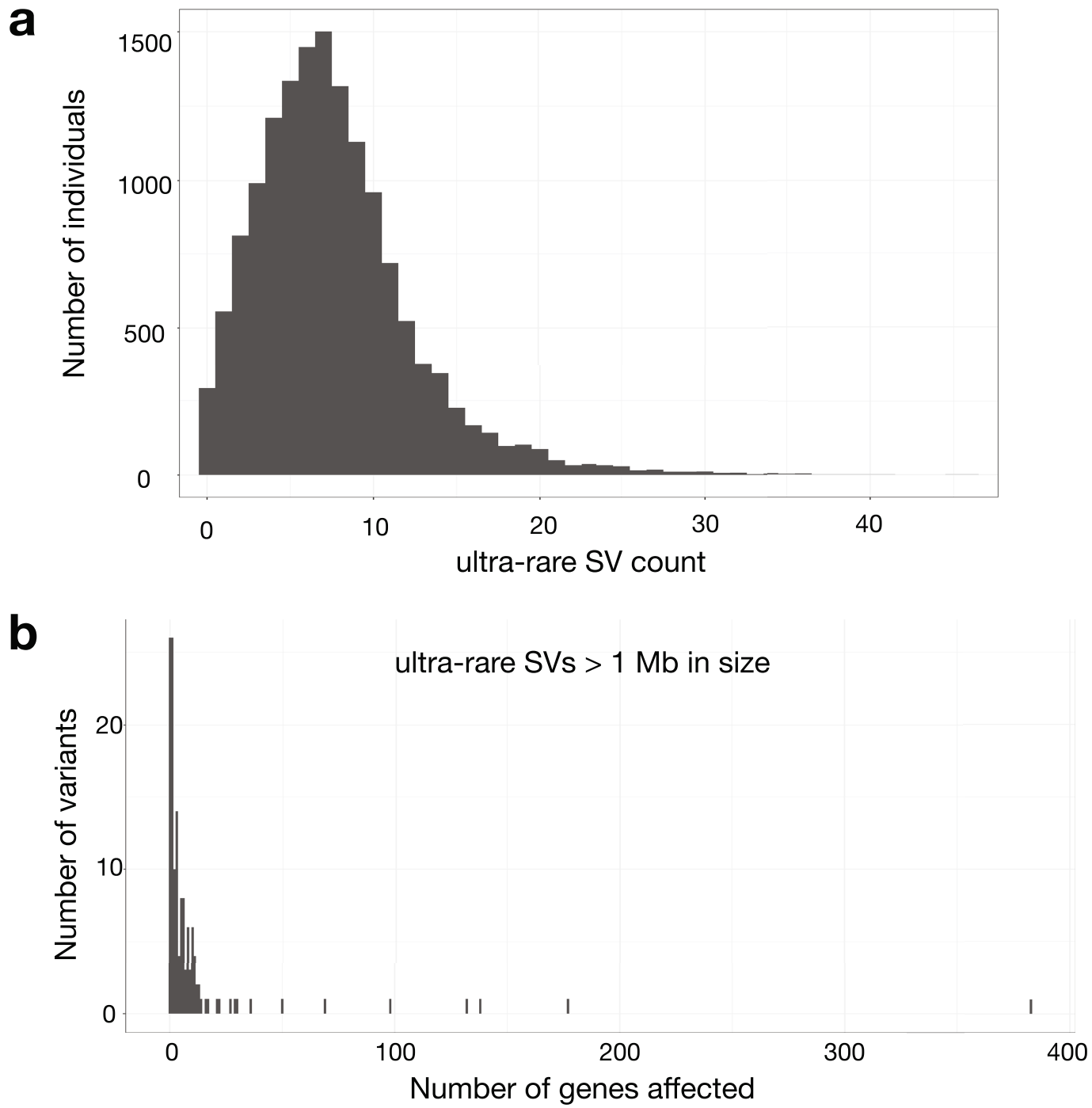
		1000 Genomes GT		
		0/0	0/1	1/1
B38 GT	0/0	6283	42	1
	0/1	138	3813	29
	1/1	15	992	1888

c

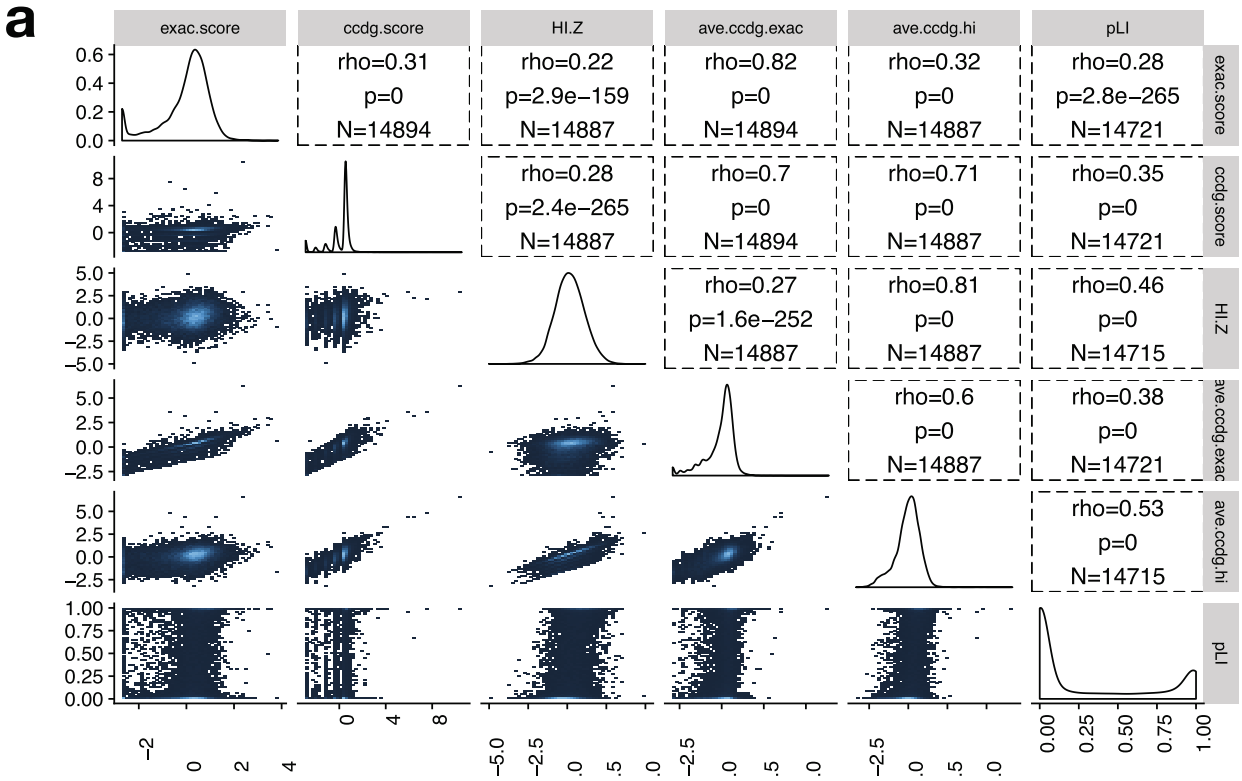
svtype	total	kappa	frac concordant
DEL	7150	0.964	0.981
BND	1495	0.879	0.929
MEI	4556	0.685	0.787
all	13201	0.852	0.908

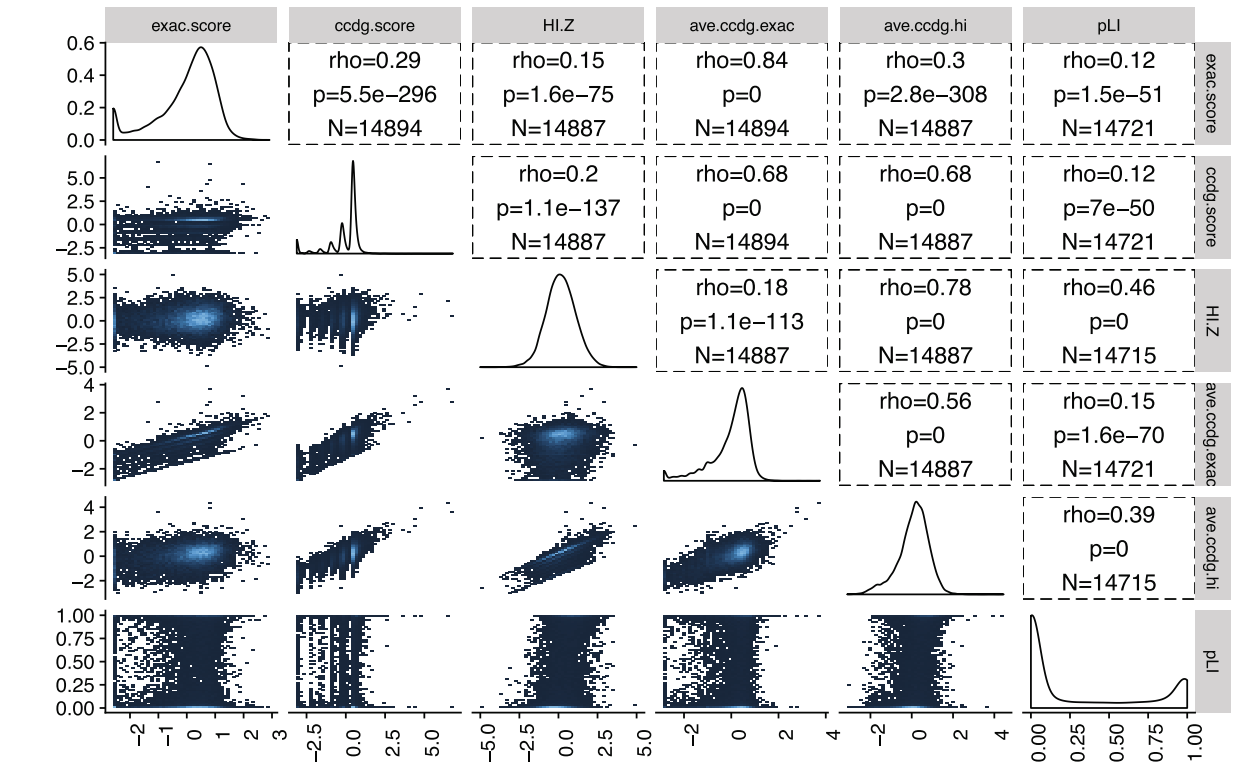


Extended Data Fig. 7 | Comparison of SV calls and genotypes to the 1KG phase 3 callset. a, Number of known and novel SVs in the B37 (left) and B38 (right) callsets, shown by frequency class. **b**, Table showing the genotypes (GT) reported in our B38 callset[5] (rows) versus the 1KG callset (columns) at SVs identified by both studies among the five samples included in both callsets. **c**, Table showing genotype concordance by SV type including the fraction of concordant calls and Cohen's *κ* coefficient. **d**, Distribution of correlation (R^2) between genotype information determined by breakpoint-spanning reads and estimates of copy number (CN) determined by read-depth analysis for the SVs shown in **b**, **c** when genotype information between the B38 and the 1KG callset is concordant (left) or discordant (middle, right). At sites with discordant genotypes, correlation with copy-number information is typically higher for genotypes from the B38 callset (middle) than the 1KG callset (right).



Extended Data Fig. 8 | Ultra-rare SVs in the B38 callset. $n = 14{,}623$. **a**, Histogram showing the number of ultra-rare SVs per individual (ultra-rare is defined as singleton variants private to a single individual or nuclear family). **b**, Histogram showing the number of genes affected by ultra-rare SVs larger than 1 Mb in size.





Extended Data Fig. 9 | Correlations between dosage sensitivity scores for CNV in the combined callset. $n = 17{,}795$. **a**, Results for deletion variants. The ExAC score is the published ExAC DEL intolerance score[45]; the CCDG score is similarly calculated from our data, using CCDG deletions; pLI is the published loss-of-function intolerance score from ExAC[27]; 'HI.Z' is the negative of the inverse-normal transformed haploinsufficiency score from DECIPHER[46]; 'ave. ccdg.exac' is the arithmetic mean of the CCDG and ExAC DEL intolerance scores; and 'ave.ccdg.hi' is the arithmetic mean of the CCDG and HI-Z scores. The correlations shown are Spearman rank correlations (rho); *P* values are calculated by two-sided Spearman rank correlation test; and *N* represents the number of genes included in the test. **b**, Results for duplication variants, using the same naming conventions as in **a**.

Extended Data Table 1 | Ancestry, ethnicity and continental origin of the samples analysed in this study

a

Ancestry	Build 37	Build 38	Combined
AFR	3683	5501	6170
AMR	698	4165	4186
EAS	65	929	972
FE	2898	1207	2884
NFE	682	9588	10254
Not Specified	105	428	436
Other	123	751	777
PI	110	87	110
SAS	62	519	558

b

Ethnicity	Build 37	Build 38	Combined
Hispanic	586	2829	2829
Non-Hispanic	3758	8022	10559
Not Specified	4082	12324	12959

c

Continent	Build 37	Build 38	Combined
African	66	24	66
Asian	32	1272	1272
Caribbean	279	1815	1815
East Asian	43	0	43
European	2985	1219	2971
North American	4641	18563	19800
Oceanic	41	18	41
Central Asian/Siberian	26	0	26
South American	274	264	274
South Asian	39	0	39

For each table, the number of samples in the B37 and B38 callsets are shown separately, and the non-redundant combined set is shown on the right. Abbreviations as in Fig. 1a.

nature research

Corresponding author(s): Hall

Last updated by author(s): Apr 4, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	no software was used
Data analysis	All sequence data were aligned and processed as described in the methods section. For the 'b37' callset, data were processed using the speedseq pipline. For the 'b38' callset data were processed according to the functional equivalence standard. We used LUMPY (v0.2.13) for per-sample SV calling followed by cohort-level merging, re-genotyping, etc, using the svtools (v0.3.2) workflow as detailed in the Methods section to produce a joint, cohort-level vcf. Dataset qc was performed using bcftools (v1.3.1) and vawk (https://github.com/cc2qe/vawk). The SNV/indel callset was produced using GATK HaplotypeCaller (v3.5-0-g36282e4) as detailed in the methods and annotated using vep and LOFTEE (v0.2.2-beta). Validation of SV by PacBio long reads was performed using custom code in (https://github.com/abelhj/long-read-validation/tree/master). All further analysese were performed using bedtools (v2.23.0) and R (v3.3.3).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The sequencing data can be accessed through dbGaP (https://www.ncbi.nlm.nih.gov/gap) under accession numbers provided in Supplemental Table 7. PacBio long read data used for SV validation can be accessed through SRA, under accession numbers provided in Supplemental Table 2. The set of high-confidence HGSVC long-read derived SV calls, validated by our independent PacBio data and used as a truth set can be found in Supplementary_File4.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was determined based on the number of distinct individuals in the callsets.
Data exclusions	As detailed in the Methods sections, samples with per-sample variant counts of any type exceeding the median+6*MAD were excluded (per our standard qc practice). A set of 64 samples were excluded because they appeared to be duplicates (or monozygotic twins) of other samples in the callset. (One per duplicate pair was excluded at random.) Additional samples were excluded because we could not obtain consent for aggregate sharing. (See methods for details.)
Replication	This was an observational study, there was no attempt at replication.
Randomization	This was an observational study, there was no randomization.
Blinding	This was an observational study, there was no blinding.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging