nature genetics

LETTERS

https://doi.org/10.1038/s41588-021-00865-4

Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits

Doruk Beyter[1], Helga Ingimundardottir[1], Asmundur Oddsson[1], Hannes P. Eggertsson[1,2], Eythor Bjornsson[1,3,4], Hakon Jonsson[1], Bjarni A. Atlason[1], Snaedis Kristmundsdottir[1,5], Svenja Mehringer[6], Marteinn T. Hardarson[1], Sigurjon A. Gudjonsson[1], Droplaug N. Magnusdottir[1], Aslaug Jonasdottir[1], Adalbjorg Jonasdottir[1], Ragnar P. Kristjansson[1], Sverrir T. Sverrisson[1], Guillaume Holley[1], Gunnar Palsson[1], Olafur A. Stefansson[1], Gudmundur Eyjolfsson[7], Isleifur Olafsson[8], Olof Sigurdardottir[9], Bjarni Torfason[10,11], Gisli Masson[1], Agnar Helgason[1,12], Unnur Thorsteinsdottir[1,3], Hilma Holm[1], Daniel F. Gudbjartsson[1,2], Patrick Sulem[1], Olafur T. Magnusson[1], Bjarni V. Halldorsson[1,5] ✉ and Kari Stefansson[1,3] ✉

Long-read sequencing (LRS) promises to improve the characterization of structural variants (SVs). We generated LRS data from 3,622 Icelanders and identified a median of 22,636 SVs per individual (a median of 13,353 insertions and 9,474 deletions). We discovered a set of 133,886 reliably genotyped SV alleles and imputed them into 166,281 individuals to explore their effects on diseases and other traits. We discovered an association of a rare deletion in *PCSK9* with lower low-density lipoprotein (LDL) cholesterol levels, compared to the population average. We also discovered an association of a multiallelic SV in *ACAN* with height; we found 11 alleles that differed in the number of a 57-bp-motif repeat and observed a linear relationship between the number of repeats carried and height. These results show that SVs can be accurately characterized at the population scale using LRS data in a genome-wide non-targeted approach and demonstrate how SVs impact phenotypes.

Human sequence diversity is partially due to SVs[1]: genomic rearrangements affecting at least 50 bp of sequence in the form of insertions, deletions, inversions or translocations. The number of SVs carried by each individual is less than the number of SNPs and short (<50 bp) insertions and deletions (indels), but their greater size makes them more likely to have a functional role[2], as evident by their disproportionately large impact on diseases and other traits[2,3].

Extensive characterization of three parent–offspring trios sequenced using several technologies[4] and an annotated set based on one sample (HG002)[5] indicate that humans carry 23,000–31,000 SVs per individual. Most studies using whole-genome sequence data are based on short-read sequencing (SRS), for which reads are typically 100–200 bp in length, allowing SNPs and small indels to be reliably identified[6,7]. However, short reads make the discovery, genotyping and characterization of SVs difficult[8], and the number of SVs found per individual has been limited to 2,000–11,000 in large-scale studies using SRS[3,9–11]. LRS, with read lengths of several kilobases, allows SVs to be detected with greater accuracy. The typical process for identifying SVs involves mapping and comparing sequence reads to a reference genome. Due to their greater length, LRS reads can be mapped more accurately than SRS reads[8]. LRS reads are also more likely to cover entire SVs, enabling better determination of their breakpoints and length. However, LRS reads have a relatively high sequencing error rate (often more than 10%) that varies depending on sample quality, sequencing technology and protocol[8]. High error rates can result in artifacts[8,12], as well as failure in SV identification. Artifacts can be especially challenging in large-scale studies, in which accumulating false positives (FPs) may dominate results and hinder downstream analysis, such as genome-wide associations. Although there are studies on detecting and characterizing SVs in human genomes using long reads[8,13–16] on select small datasets, analysis at scale has not been reported.

We present a study applying LRS at a population scale, focused on identifying a set of reliable SVs consistently called across individuals that can be used for downstream analysis in the context of diseases and other traits. We sequenced 3,622 Icelanders by using Oxford Nanopore Technologies (ONT), including 441 parent–offspring trios, recruited for various studies at deCODE genetics[17]. DNA was isolated from whole blood ($n = 3{,}524$) and heart tissue ($n = 102$) and sequenced with ONT PromethION instruments (Methods). SRS and DNA-chip data were also available for all of these individuals[18].

[1]deCODE genetics/Amgen, Inc., Reykjavik, Iceland. [2]School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland. [3]Faculty of Medicine, University of Iceland, Reykjavik, Iceland. [4]Department of Internal Medicine, Landspitali—the National University Hospital of Iceland, Reykjavik, Iceland. [5]School of Technology, Reykjavik University, Reykjavik, Iceland. [6]Algorithmic Bioinformatics (ABI), Freie Universitat Berlin, Berlin, Germany. [7]The Laboratory in Mjodd, Reykjavik, Iceland. [8]Department of Clinical Biochemistry, Landspitali—the National University Hospital of Iceland, Reykjavik, Iceland. [9]Department of Clinical Biochemistry, Akureyri Hospital, Akureyri, Iceland. [10]Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland. [11]Department of Cardiothoracic Surgery, Landspitali—the National University Hospital, Reykjavik, Iceland. [12]Department of Anthropology, University of Iceland, Reykjavik, Iceland. ✉e-mail: bjarnih@decode.is; kstefans@decode.is

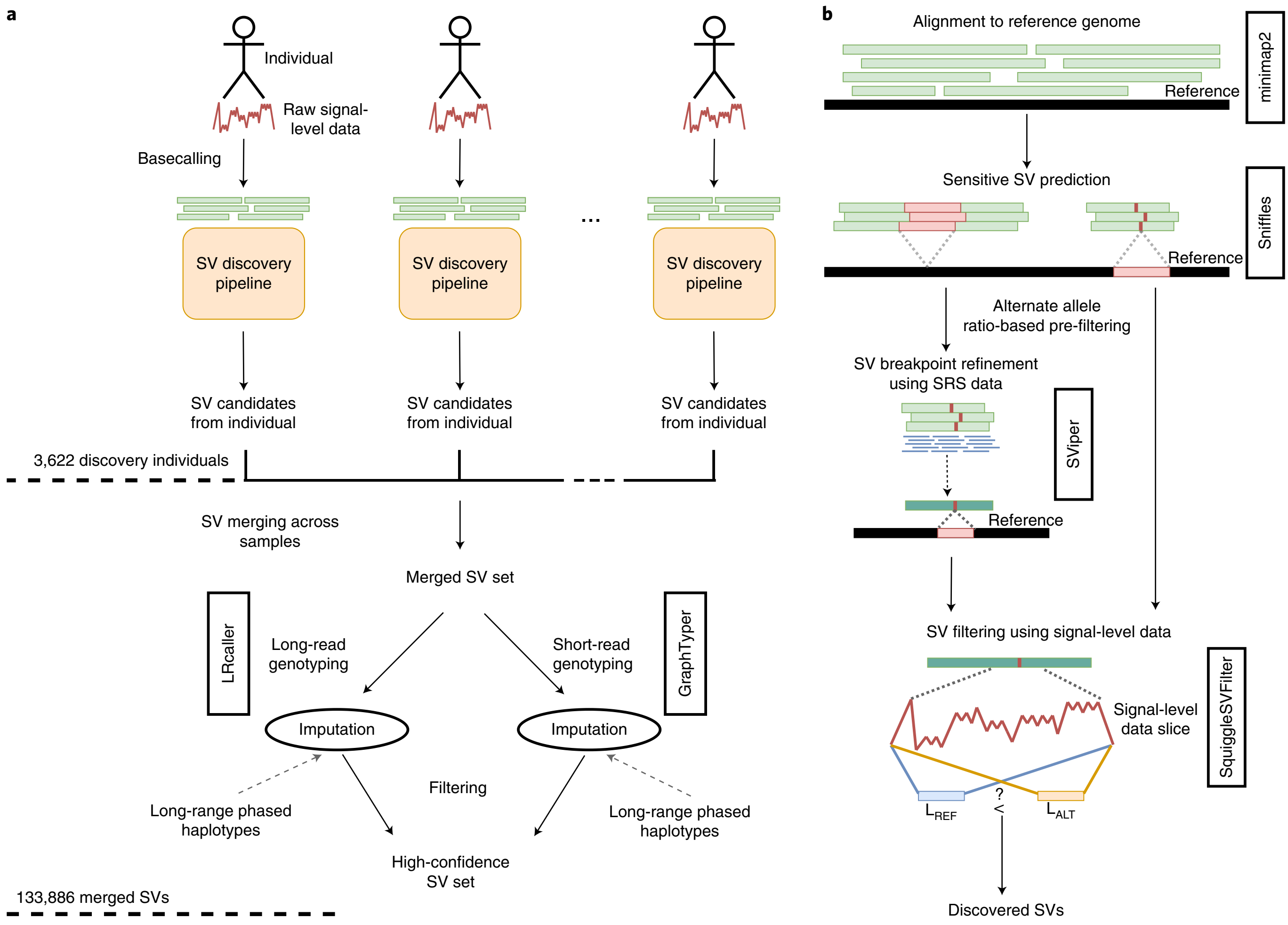


Fig. 1 | SV analysis workflow. a, Each individual is basecalled and their SVs are discovered independently. SV sets are merged across all individuals, and the merged SV set is used to genotype individuals using both SRS and LRS data separately. Finally, genotyped variants are imputed into long-range phased haplotypes, and variants that pass SV filtering are accepted as high-confidence variants. **b**, For the SV discovery pipeline, reads are mapped to the human reference genome (GRCh38) using minimap2, followed by sensitive SV predictions using Sniffles. SV predictions are then pre-filtered based on their alternate allele ratio, and SV breakpoints are refined using SRS data, if possible, with SViper. Finally, candidate SVs are compared against the raw-signal-level data using SquiggleSVFilter for further verification. ALT, alternate; REF, reference.

We introduced a number of tools and approaches to facilitate SV analysis using long reads characterized by a high error rate at scale, including SV filters and heuristics for merging SVs. Finally, to illustrate the power of population-based LRS, we developed a tool to perform joint genotyping on LRS data.

We basecalled raw sequence data from 4,757 flow cells, in which half of all sequenced basepairs (N50) belonged to reads longer than 19,940 bp (Supplementary Data 1 and Extended Data Fig. 1a). We mapped[19] all reads to the human reference genome GRCh38 (ref. [20]) and observed a median LRS aligned coverage of 17.2× (range, 10.0–94.3×; ‘Sequencing statistics’ in the Supplementary Information and Extended Data Fig. 1b) per individual. A median of 87.6% of basepairs aligned to the reference (Extended Data Fig. 1c), and the median sequencing error rate was 11.6% (3.3% for insertions, 4.5% for deletions and 3.8% for mismatches, Extended Data Fig. 1d).

We generated a high-confidence SV set in four stages: (1) discovery, (2) merging across individuals, (3) genotyping and (4) imputation (Fig. 1a). We began (Fig. 1b) by discovering SVs with high sensitivity[8] and refined them at predicted breakpoints using SRS data, when possible[21] (Methods). Their presence was confirmed using the raw-signal-level data[22] (Methods and Extended Data Fig. 2) to alleviate potential basecalling and alignment errors. We did not attempt to discover translocations and inversions. The SVs discovered across individuals were then merged and genotyped using two independent datasets: 3,622 and 10,000 Icelanders with LRS and SRS data, respectively[7,23] (‘Individual selection for short-read genotyping’ in the Supplementary Information). Finally, we imputed the genotyped variants into the long-range phased haplotypes of a total of 166,281 SNP-chip-typed Icelanders[18,24,25] and defined a set of high-confidence SVs, based on imputation accuracy and other filters (Methods).

We identified 133,886 high-confidence SV alleles (75,050 insertions, 55,649 deletions and 3,187 unresolved insertions or deletions, Supplementary Data 2, https://github.com/DecodeGenetics/LRS_SV_sets and Fig. 2a), avoiding double counting of alleles of similar length at similar positions (Methods). We observed more insertions than deletions[4,16,26,27]. This contrasts with results based on SRS[3,9], in which deletions are typically more frequent and easier to identify. We were able to impute 120,108 SV alleles (67,673 insertions, 49,845 deletions and 2,590 unresolved insertions or deletions) into long-range phased haplotypes from 166,281 chip-typed Icelanders. We identified a median of 22,636 SVs per individual (a median of 13,353 insertions and 9,474 deletions, ‘Calculating SV counts per individual’ in the Supplementary Information and Fig. 2a), of

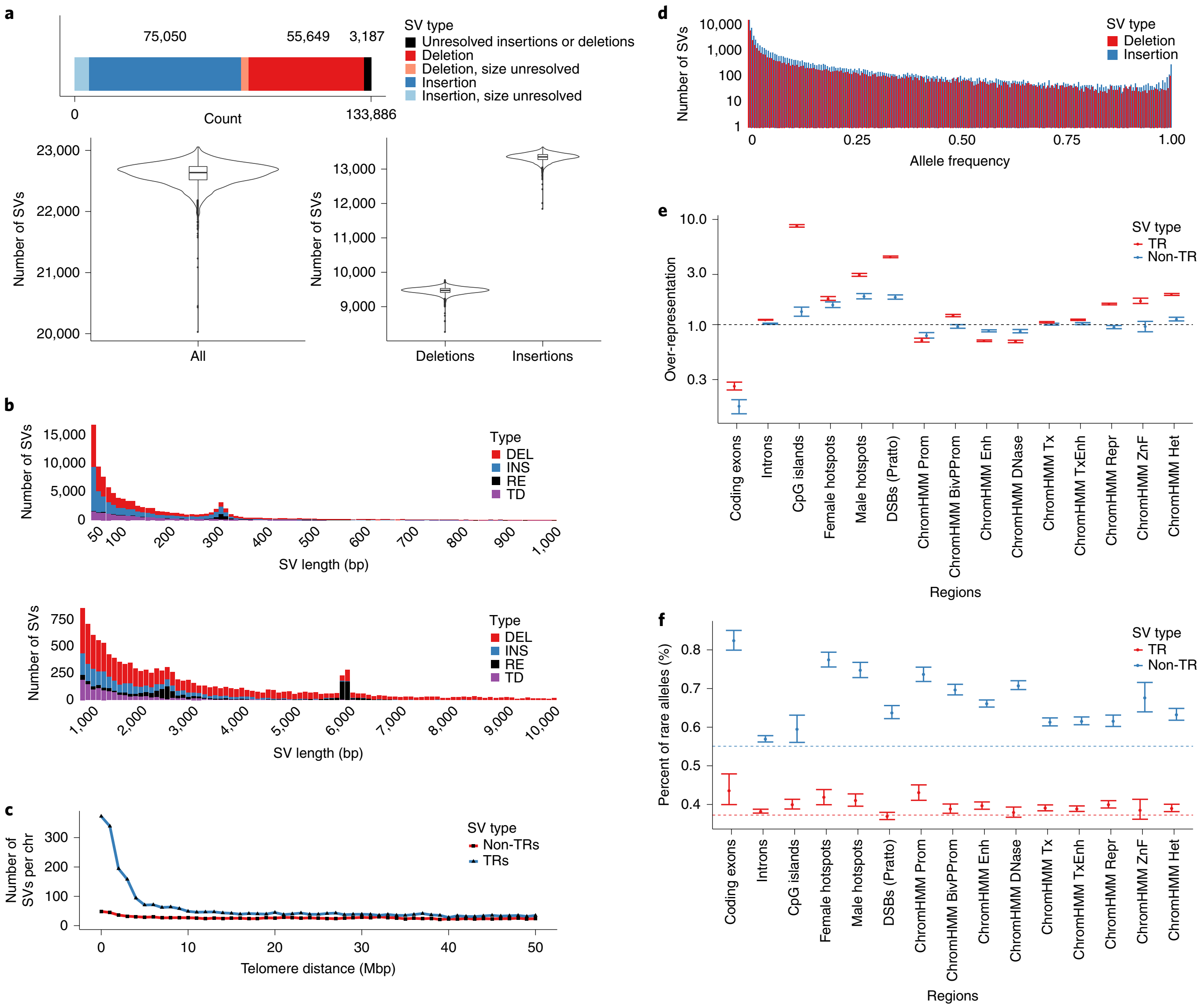


Fig. 2 | Merged SV set characteristics. a, Number of merged SVs and distribution of phased SVs per individual, total and stratified by SV type ($n=3,204$; 412 individuals were not genotyped using SRS data, and six individuals in which <90% of SV alleles could be phased were omitted. In total, 6,455 insertions and 3,574 deletions were of unresolved size. Box limits indicate upper and lower quartiles, center lines indicate medians, and whiskers indicate ±1.5× the interquartile range). **b**, Stacked SV length distributions in ranges (50 bp, 1 kb) and (1 kb, 10 kb), stratified by SV type. Insertions are subset into REs and TDs. Remaining insertions are shown as INS. The peaks observed around 300 bp, 2.5 kb and 6 kb correspond to SINE, SVA and LINE elements, respectively, highlighted by the increase in REs ($n=120,670$; 13,216 SVs without a specified size were omitted). DEL, deletion. **c**, Number of TR and non-TR SVs per chromosome (chr) as a function of telomere distance (binned at 1 Mb). **d**, Allele frequency distribution of SVs binned at 0.05%. **e**, Over-representation of SVs across genomic regions. DSB, double-strand break; enh, enhancer; Prom, promoter regions (PromU/D1/D2); BivPProm, bivalent and poised promoters; Enh, enhancer states (EnhA1/2/AF/W1/W2/Ac); DNase, deoxyribonuclease (DNase)-only states; Tx, transcribed regions (Tx5′/Tx/Tx3′/TxWk); TxEnh, enhancers within transcribed regions (TxEnh5′, TxEnh3′, TxEnhW, and TxReg); Repr, polycomb-group-repressive complex 2; ZnF, enriched over zinc-finger genes and repeats; Het, heterochromatin. **f**, Percent of rare SVs across genomic regions. In **e**,**f**, centers are means, and error bars indicate 95% confidence intervals from 1,000-fold bootstrapping ('Confidence intervals and *P*-values' in the Supplementary Information). $n=120,670$; 13,216 SVs without a specified size were omitted; SV numbers in each category are provided in Supplementary Table 1.

which a median of 20,891 SVs were imputed, spanning a cumulative median length of 10.02 Mb per haploid genome. We estimated the false negative (FN) and FP rates of our high-confidence SV set by comparing it to public SV datasets. Comparison to an LRS SV dataset from Audano et al.[16] ($n=15$) and an SRS SV dataset, gnomAD-SV[11] ($n=14,891$), using SVs within genome in a bottle tier 1 regions of HG002 (ref. [5]), suggested FN rates of 2.6% and 3.4%, respectively, for our dataset (Methods). We estimated an FP rate of 8.2% for our dataset by considering our common SVs absent from the dataset in Audano et al.[16] and their observed versus expected allele counts in HG002. We also estimated FP rates of 6.3–7.6% for the gnomAD-SV dataset, which is comparable to the rate we estimated for our call set (Methods). These estimates may be upwardly biased due to population-specific drift. In an attempt to validate 70 of the SVs using PCR, an SV was confirmed for 60 of the successful 63 assays (seven assays failed), suggesting an FP rate of 4.8% ('Polymerase Chain Reaction (PCR) verification of SVs' in the Supplementary Information and Supplementary Data 3).

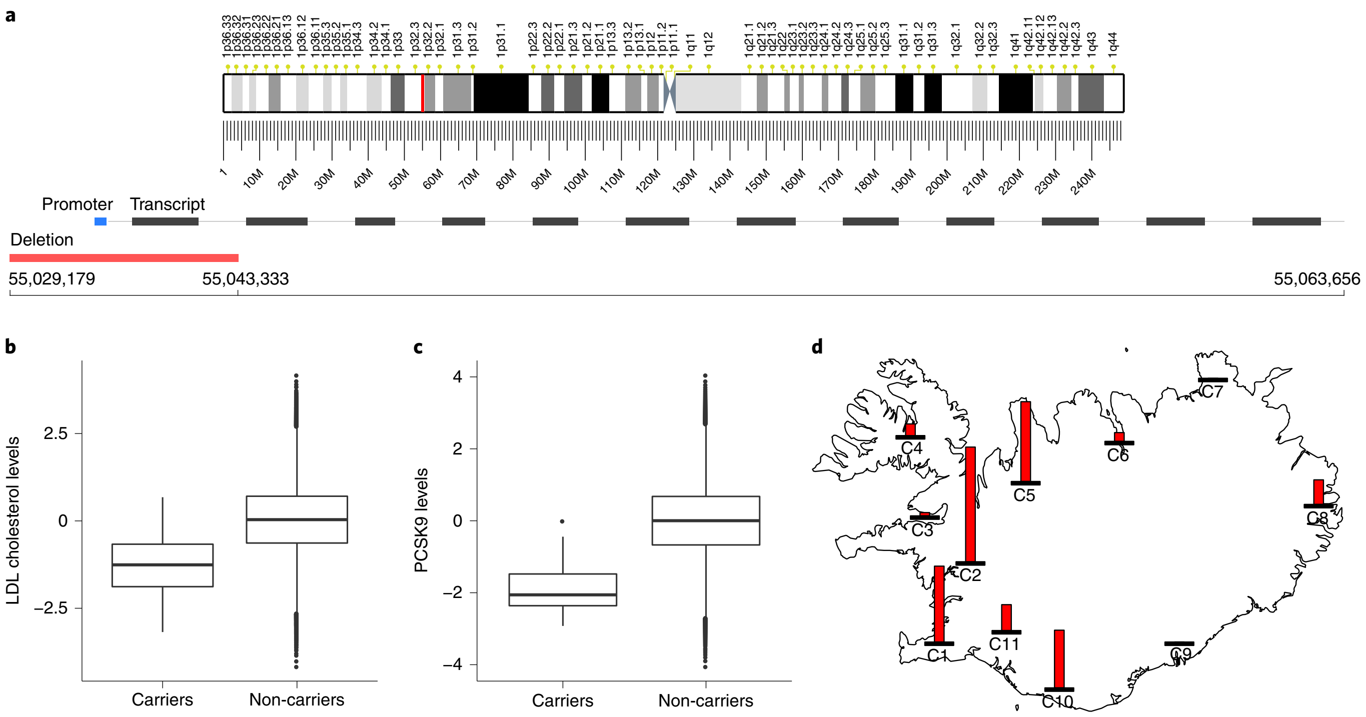


Fig. 3 | Large deletion in *PCSK9* associated with lower LDL cholesterol levels. a, Chromosome 1 ideogram with cytobands, highlighting the SV site in red. A 14,154-bp (allele frequency, 0.037%) deletion removes the promoter and the first coding exon of *PCSK9* at 55,029,179–55,043,333 bp (GRCh38). **b**, LDL cholesterol levels in carriers and non-carriers ($n = 72$ carriers, $n = 96{,}840$ non-carriers, effect = −1.31 s.d., $P = 7.0 \times 10^{-20}$, two-sided linear regression). **c**, PCSK9 protein levels in carriers and non-carriers using SOMAscan ($n = 20$ carriers, $n = 38{,}385$ non-carriers, effect = −1.99 s.d., $P = 3.1 \times 10^{-13}$, two-sided linear regression). **d**, Geographical distribution of the *PCSK9* deletion in 166,281 chip-typed Icelanders. Each bar shows the allele frequency of the variant relative to the geographical region with the highest frequency. Icelandic counties are grouped into 11 regions, C1–C11, as shown in the panel ('Calculating the geographic distribution of the carriers of the PCSK9 deletion' in the Supplementary Information and Supplementary Table 2). In **b**,**c**, box limits indicate upper and lower quartiles, center lines indicate medians, and whiskers indicate ±1.5× the interquartile range.

To measure the relative merits of LRS versus SRS in SV discovery, we assessed whether SV alleles discovered using LRS were also found by gnomAD-SV[11], which only uses SRS data. Comparing the gnomAD-SV dataset to the SV calls made by Audano et al.[16] with allele frequency greater than 50% suggested a 41.3% FN rate for the gnomAD-SV dataset (Methods). Similarly, among our set of 46,352 imputed SV alleles with frequency greater than 10%, 19,430 (41.9%) were not found in the gnomAD-SV dataset. Repeating this analysis for subsets of SV alleles within or outside of tandem-repeat (TR) regions, we observed FN rates of 47.4% and 27.4%, respectively, for the gnomAD-SV dataset. LRS data also improved the genotyping of SVs in our data. Of 120,108 imputed SV alleles, 76,857 (64.0%) and 3,917 (3.2%) SVs were imputed only from LRS and SRS genotyping, respectively. Furthermore, 74.2% and 38.6% of SV alleles within and outside of TR regions, respectively, could not be imputed using genotype calls from SRS data. These results show that SV discovery and genotyping at the population scale from LRS data are more accurate and reliable than those from SRS[11] data. The difference is particularly pronounced for SVs in TR regions, which have mutation rates one to four orders of magnitude higher than those for other genomic loci[28,29].

The number of variants in our SV set rapidly decreased with length (Fig. 2b), consistent with previous reports[11,14,16]. To better characterize insertions, we classified them into three groups: tandem duplications (TD), retrotransposable elements (REs) and other insertions (INS), corresponding to 30%, 7% and 63% of insertions, respectively. We observed three noticeable peaks at sizes around 300 bp, 2.5 kb and 6 kb, due to REs, corresponding to short interspersed nuclear elements (SINEs), SINE/VNTR/Alu (SVA) and long interspersed nuclear elements (LINEs) (Fig. 2b), as expected. We found more SVs, particularly TR SVs, near telomeres[16] (Fig. 2c), a reflection of the sequence content of telomeres and the high mutation rate of TRs[16,30]. The number of alleles detected decreased with increasing allele frequency, with 40.1% of them at a frequency less than 1%, which is rare (Fig. 2d and Extended Data Fig. 3). The small number of variants that stand out for being fixed or near fixed in frequency are most likely examples in which the reference sequence carries a derived allele rather than the ancestral state.

In general, frequency reflects the age of variants, such that younger variants are rarer than those that are older. As a result, differences in the relative allele frequencies of SVs by genomic region can provide information about the strength of negative selection that has acted against them. We observed both an under-representation of SVs and an elevated fraction of rare SVs in coding exons and non-coding regulatory regions, such as enhancers and promoters, compared to the genomic average ($P < 0.002$, in coding exons, enhancers and promoters, for both TR and non-TR SVs, bootstrap test, Fig. 2e,f). We also found that SVs in TRs tended to be observed at higher frequencies than those outside of TRs (particularly when compared to those in coding exons), suggesting a higher tolerance for SVs in TRs. In regulatory elements, although SVs in TRs were more under-represented than SVs outside TRs, they similarly had a higher fraction of common alleles than SVs outside TRs. In accordance with the notion that recombination plays a role in SV formation[16], we found that SVs were enriched in double-strand break regions[31] ($P < 0.002$, bootstrap test) and recombination hotspots[32] ($P < 0.002$, bootstrap test), particularly in TR alleles within male hotspots ($P < 0.002$, bootstrap test) (Fig. 2e). Interestingly, we also observed an elevated rate of rare alleles in recombination hotspots (Fig. 2f).

Variants inside coding exons that are not multiples of three in length generally result in translational frameshift and non-functional proteins. Among the 549 variants contained within a single coding exon, we observed a deficit in variant lengths that were not multiples of three: 187 (34.1%) compared to the two-thirds (362) expected ($P=4.9\times10^{-55}$, two-sided binomial test, Extended Data Fig. 4), in line with results using indels[25]. These results are consistent with the hypothesis that SVs that result in translational frameshifts are selected against due to their phenotypic impact.

We used two strategies to determine the impact of SVs on phenotypes. First, we inferred association of SVs based on linkage disequilibrium (LD) with variants previously reported to be associated with phenotypes. Of the 116,479 unique variants reported in the genome-wide association study (GWAS) catalog, 11,194 were in strong LD ($R^2\geq0.8$) with 5,238 SV alleles in our dataset, suggesting possible functional explanations of these associations (Supplementary Data 4). A subset of 34 high-impact and 54 moderate-impact SVs, overlapping exons or splice regions, were in strong LD with 198 GWAS catalog variants and were therefore plausible causal variants for the reported associations. Among these are examples in which the presence of an SV was previously established using alternate methods, including a deletion in *LCE3B*[33] associated with psoriasis and a deletion in *CTRB2* associated with diabetes[34,35] and age-related macular degeneration[36]. Another example is a rare 2,460-bp deletion that removes two exons of *COL4A3* and is associated with hematuria[37]. We also found loci where the occurrence of an SV at a GWAS locus has not been reported, including a deletion that overlaps the first exon of *SLC25A24* and is in strong LD with a SNP associated with white blood cell count[38] and a 120-bp inframe deletion in *KAT2B* that removes 40 amino acids from the translated protein and is in strong LD with a variant associated with systolic blood pressure[35,39].

Second, we performed direct tests of association with phenotypes of a cohort of Icelanders (Methods). We found an association with a rare 14,154-bp deletion overlapping the first exon of *PCSK9* (Fig. 3a) and LDL cholesterol levels (adjusted effect = −1.31 s.d. and $P=7.0\times10^{-20}$, Fig. 3b). LDL cholesterol levels were 0.93 mmol l^{-1} lower in carriers ($n=75$) than in non-carriers ($n=98{,}081$). We observed 13, 56 and 119 heterozygous carriers of the deletion in our LRS, SRS and imputation datasets, respectively, corresponding to an allele frequency of 0.037%. No homozygous carrier was identified. *PCSK9* encodes the enzyme proprotein convertase subtilisin–kexin type 9 (PCSK9), a key regulator of LDL cholesterol metabolism[40] and a target of cholesterol-lowering drugs[41]. Loss-of-function variants in *PCSK9* are known to result in lower levels of LDL cholesterol and reduced cardiovascular risk[42–44], consistent with the association observed here. We next tested this deletion for association with 4,792 plasma proteins measured in 38,405 Icelanders by using SOMAscan[45]. PCSK9 levels in carriers ($n=20$) were, on average, 1.99 s.d. below the population mean ($P=3.1\times10^{-13}$, Fig. 3c). The geographical distribution of the carriers in Iceland suggests that the variant is more prevalent in western Iceland than in other parts of the country (Fig. 3d). Two carriers were found in the gnomAD-SV dataset (one of 9,534 Africans and one of 7,624 Europeans), and one carrier with low LDL levels was found in a study with Dutch participants[46], showing that the deletion is not specific to the Icelandic population.

The SVs discussed above, although detected using LRS data, also could have been detected using SRS data, as they occurred in genomic locations where short reads can be reliably mapped. Below, we present three examples of common multiallelic SVs within TRs that are not found in the gnomAD-SV dataset and are difficult to detect from SRS. These include the repeat region of an exon in *ACAN*, a proline-rich repeat region in *NACA* and the zinc-finger (ZnF) domain of *PRDM9*.

ACAN contains an exonic variable number TR (VNTR), with a 57-bp motif (19 amino acids in aggrecan, the translated protein) within one of its chondroitin sulfate (CS) attachment sites[47], in which we identified 11 SV alleles: the reference allele along with deletions of one, three, four, five, six, eight and 14 motif(s) and insertions of one, two, three and four motif(s). (Fig. 4a,b). We found the five-motif (285 bp) deletion to be highly correlated (LD $R^2=0.96$) with a synonymous SNP (rs16942341[T]; allele frequency, 3%) reported to associate with decreased height (effect = −0.13 s.d., $P=4\times10^{-27}$) in a large GWAS analysis of 183,000 individuals of European decent[48]. Both the reported synonymous variant and the SV were strongly associated with height in our data (effect = −0.13 s.d., $P=1.02\times10^{-10}$; and effect = −0.12 s.d., $P=1.35\times10^{-9}$, respectively). We observed a stronger association and a linear relationship (effect = 0.016 s.d. per motif inserted, $P=6.2\times10^{-18}$, Methods) between the number of motifs carried and height (Fig. 4c and Table 1), suggesting that this SV plays a causal role in the association with height. The variance in height explained by the number of motifs carried ($R^2=9.9\times10^{-4}$) was also higher than the variation explained by the SNP correlated with the five-motif deletion ($R^2=5.7\times10^{-4}$). The number of TRs results in a change in the number of CS attachment sites and thereby the number of attached CS chains on the aggrecan molecule[49,50]. Aggrecan is the most abundant proteoglycan in cartilage[51], and the negatively charged CS chains were shown to move water into cartilage[52]; thus the varying number of CS chains can affect the function of this protein. Following the submission of this manuscript, this association was also observed in the UK Biobank[53].

NACA contains an exonic VNTR, with a repeat length of 69 bp (23 amino acids in NAC-α, the translated protein), repeated 18 times in GRCh38, in which we identified deletions of one and two motifs (69 and 138 bp, respectively, Fig. 4d). The intergenic SNPs rs2860482[A] and rs7978685[T] were reported to be associated with atrial fibrillation[54], and both SNPs were in strong LD ($R^2=0.85$) with the reference allele of the VNTR. This exon of *NACA* is transcribed in skNAC, a muscle-specific, alternately spliced transcript, the importance of which was demonstrated in the developing heart in animal models[55]. This suggests that the multiallelic SV in *NACA* is the likely explanation of the observed associations with atrial fibrillation[55,56].

Within the ZnF domain of the recombination hotspot positioning gene *PRDM9*, we identified deletions of one, four and five ZnF(s) and insertions of one, two and four ZnF(s), resulting in the removal

Fig. 4 | Multiallelic SVs in repeat regions within exons of *ACAN*, *NACA* and *PRDM9*, difficult for SV detection using SRS. a,**d**,**e**, Ideograms of respective chromosomes with cytobands, highlighting the SV site in red. *SRS mean depth of coverage tracks across respective genes were accessed via the gnomAD browser. Motif lengths and TR begin and end sites are not to scale. Inserted or deleted motifs are shown with arbitrary begin sites within the TR region. **a**, SV alleles with inserted or deleted 57-bp motifs within an exon of *ACAN*. The SRS-coverage track shows the absence of reads with a reliable mapping across the TR region. **b**, Allele frequencies of the SV alleles, including the reference allele. **c**, Effects of SV alleles and the reference allele on height, showing a linear relationship (effect = 0.016 s.d. per motif inserted, $P=6.2\times10^{-18}$, two-sided linear regression). Error bars indicate 95% confidence intervals, and centers show the average change in height, expressed in s.d. units. SV alleles with frequency less than 0.01% were omitted in **a**,**b** due to their large confidence interval for effect values. Carrier numbers for each allele are given in Supplementary Table 1. **d**, SV alleles with deleted 69-bp motifs within an exon of *NACA*. **e**, SV alleles with inserted or deleted 84-bp ZnF motifs within the last exon of *PRDM9*. SRS-coverage tracks in **a**,**d**,**e** indicate reduced numbers of reliably mapped reads across the TR regions, compared to those of remaining regions of the gene.

or addition of a number of ZnF motifs from or to the encoded protein, methyltransferase PRDM9 (Fig. 4e). The ZnF domain of PRDM9 is the DNA-binding domain, and the SV alleles thus introduce alterations in the DNA-binding motif of PRDM9 and consequently change the locations of meiotic recombination[57,58]. All the different ZnF-motif lengths showed a strong association with

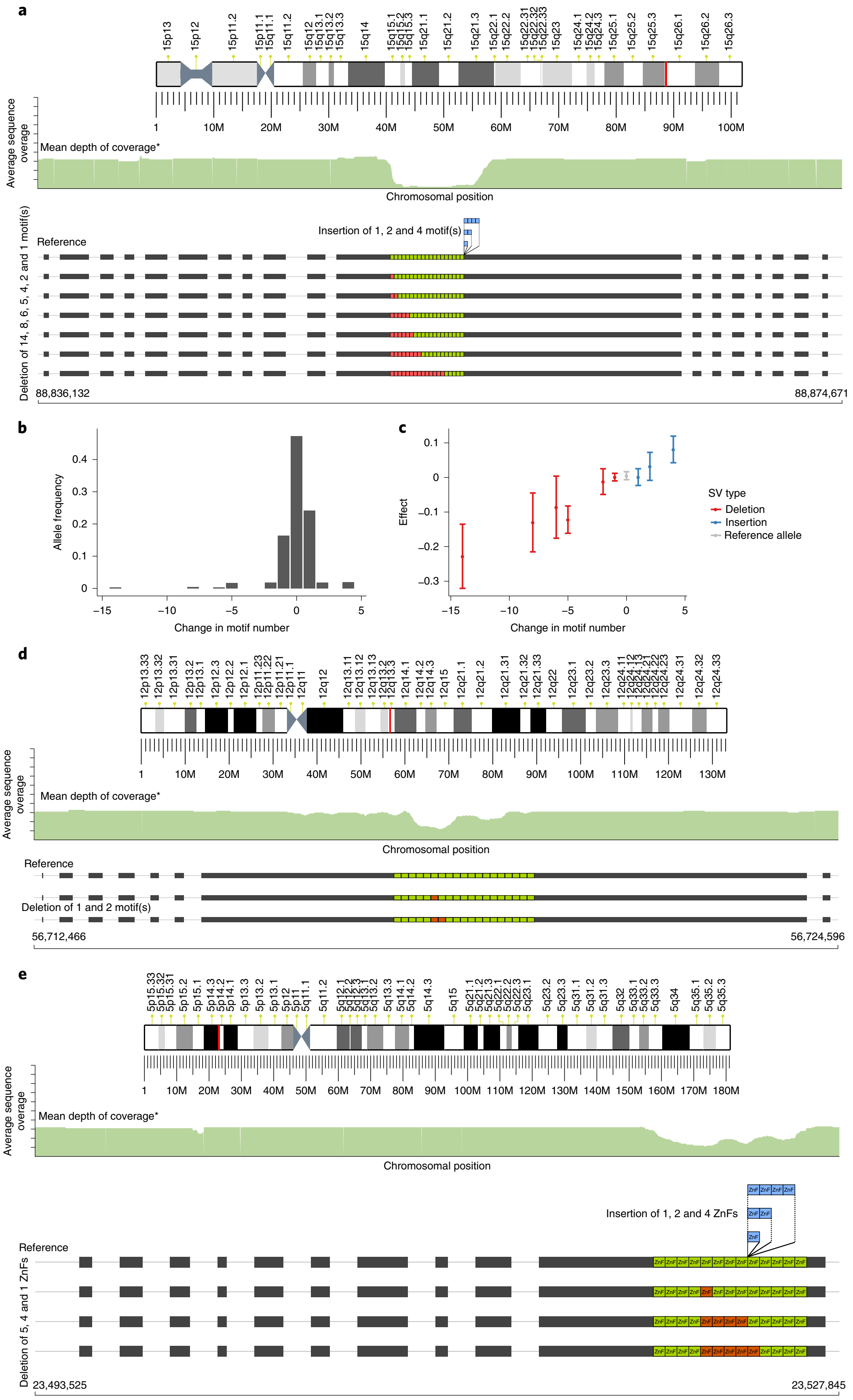


Table 1 | *ACAN* SV allele frequencies, lengths, *P* values (two-sided linear regression) and effects on height and SV types

Frequency	Length (bp)	*P* value	Effect (s.d.)	Event
3.0×10^{-3}	−798	1.5×10^{-6}	−0.228	DEL
3.9×10^{-3}	−456	2.8×10^{-3}	−0.130	DEL
3.2×10^{-3}	−342	6.0×10^{-2}	−0.086	DEL
1.7×10^{-2}	−285	1.4×10^{-9}	−0.122	DEL
9.8×10^{-4}	−154	0.49	−0.058	DEL
1.9×10^{-2}	−114	0.53	−0.012	DEL
0.17	−57	0.86	0.001	DEL
0.47	0	0.40	0.005	Reference allele
0.24	56	0.94	0.001	INS
1.8×10^{-2}	117	0.12	0.032	INS
7.4×10^{-5}	171	9.0×10^{-2}	0.504	INS
1.9×10^{-2}	228	3.6×10^{-5}	0.081	INS

Table 2 | *PRDM9* SV allele frequencies, lengths, *P* values (two-sided linear regression) and effects on recombination hotspots and SV types

Frequency	Length (bp)	*P* value	Effect (s.d.)	Event
1.3×10^{-2}	−420	2.9×10^{-12}	−0.204	DEL
5.6×10^{-3}	−336	1.6×10^{-5}	−0.201	DEL
2.3×10^{-3}	−84	9.8×10^{-6}	−0.318	DEL
9.1×10^{-5}	331	1.8×10^{-7}	−1.679	INS
5.4×10^{-3}	168	4.9×10^{-305}	−1.700	INS
2.6×10^{-2}	83	$7.9\times10^{-1,717}$	−1.732	INS

the location of crossovers as measured by the fraction of crossovers that occur in recombination hotspots (Table 2). These results are consistent with previous results, which were indirectly ascertained via SNPs tagging multiple motif counts[32], while the current results allow us to directly ascertain the effects of each motif count individually.

We cataloged genes with rare homozygous loss-of-function SVs, as predicted by the Ensembl Variant Effect Predictor[59], and found 181 genes with a rare loss-of-function SV, for which at least one homozygous carrier was observed (151 of these were not found in the gnomAD-SV set, Supplementary Data 5). A number of these were reported to cause diseases under recessive inheritance. One example is a 57-kb deletion, overlapping the genes *CTNS* and *SHPK*, originally associated with cystinosis[60], a lysosomal storage disease characterized by the abnormal accumulation of the amino acid cystine, in which homozygous carriers of the deletion generally develop cystinosis. We identified a single homozygous carrier of this deletion in our imputation set, who was not in our genotyping set and was diagnosed with cystinosis. We also observed a 31.7-kb deletion (exons 11–17; allele frequency, 0.45%) in *GALC*, which encodes galactosylceramidase. This is the most common mutation causing Krabbe disease in Europeans[61–63]. We identified a single homozygote for this deletion, a boy with a diagnosis of Krabbe disease.

In this study, we demonstrate the application of LRS at the population scale and describe how it can be used to accurately identify SVs and to assess their impact on human disease and other traits. We identified over 22,636 SVs per individual, three to five times more than those found in SRS data[10,11]. We show that LRS is more sensitive than SRS in detecting SVs across the genome. This advantage is most pronounced in repeat regions, such as TRs. We report 5,238 SVs in strong LD with variants in the GWAS catalog that are associated with a disease or other traits. This constitutes an increase of over twofold from the number of SVs using SRS data alone[11]. These results show that LRS can further our understanding of disease mechanisms and the effects of sequence variation on human traits.

LRS technology and accompanying data analysis methods are still being developed and can be improved upon. Although we detected a large number of SVs per individual, we did not attempt to discover all forms of SVs. We detected fewer long insertions than expected, possibly due to sequencing bias or limitations of the LRS analysis algorithms. We also expected an under-representation of very rare SV alleles, as it is more difficult to phase and impute alleles with few carriers accurately. Although we highlighted SVs that overlap coding exons due to their established functional impact, other SVs may still affect the individual, for example, those altering regulatory regions or changing RNA secondary structure. A better understanding of the biochemical causes and consequences of SVs will be essential to understand human evolution and disease. These will in turn also lead to better analysis methods and increase our ability to identify SVs and assess their impact.

SVs have frequently been found using targeted approaches, often relying on discovered SNPs or indels in a disease-association context. We demonstrate that our method can identify SVs in a genome-wide, non-targeted fashion. We show that SVs affecting protein function are disproportionately rare. As a result, large-scale SV studies will be essential to characterize their role in the genetics of disease. This study, based on LRS data from 3,622 Icelanders, lays down an important foundation for further large-scale SV studies, allowing investigation of their full frequency spectrum, including those in genomic regions thus far inaccessible to SRS technologies.

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/s41588-021-00865-4.

Received: 18 November 2019; Accepted: 5 April 2021;
Published online: 10 May 2021

References

1. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
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. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
4. Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
5. Zook, J. M. et al. A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38**, 1347–1355 (2020).
6. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
7. Eggertsson, H. P. et al. Graphtyper enables population-scale genotyping using pangenome graphs. *Nat. Genet.* **49**, 1654–1660 (2017).
8. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
9. Kloosterman, W. P. et al. Characteristics of de novo structural changes in the human genome. *Genome Res.* **25**, 792–801 (2015).
10. Abel, H. J. et al. Mapping and characterization of structural variation in 17,795 deeply sequenced human genomes. *Nature* **583**, 83–89 (2020).
11. Collins, R. L. et al. A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).

12. Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
13. Stancu, M. C. et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat. Commun.* **8**, 1326 (2017).
14. De Coster, W. et al. Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome Res.* **29**, 1178–1187 (2019).
15. Gilpatrick, T. et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat. Biotechnol.* **38**, 433–438 (2020).
16. Audano, P. A. et al. Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663–675 (2019).
17. Gudbjartsson, D. F. et al. Sequence variants from whole genome sequencing a large group of Icelanders. *Sci. Data* **2**, 150011 (2015).
18. Jónsson, H. et al. Whole genome characterization of sequence diversity of 15,220 Icelanders. *Sci. Data* **4**, 170115 (2017).
19. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
20. Schneider, V. A. et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
21. Mehringer, S. et al. SViper: a tool for SV polishing. *Prep.* (2019).
22. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12**, 733–735 (2015).
23. Eggertsson, H. et al. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* **10**, 5402 (2019).
24. Kong, A. et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40**, 1068–1075 (2008).
25. Gudbjartsson, D. F. et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
26. Pendleton, M. et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).
27. Seo, J. S. et al. De novo assembly and phasing of a Korean human genome. *Nature* **538**, 243–247 (2016).
28. Sulovari, A. et al. Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proc. Natl Acad. Sci. USA* **116**, 23243–23253 (2019).
29. Duitama, J. et al. Large-scale analysis of tandem repeat variability in the human genome. *Nucleic Acids Res.* **42**, 5728–5741 (2014).
30. Sun, J. X. et al. A direct characterization of human mutation based on microsatellites. *Nat. Genet.* **44**, 1161–1165 (2012).
31. Pratto, F. et al. Recombination initiation maps of individual human genomes. *Science* **346**, 1256442 (2014).
32. Halldorsson, B. V. et al. Human genetics: characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**, eaau1043 (2019).
33. De Cid, R. et al. Deletion of the late cornified envelope *LCE3B* and *LCE3C* genes as a susceptibility factor for psoriasis. *Nat. Genet.* **41**, 211–215 (2009).
34. Onengut-Gumuscu, S. et al. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47**, 381–386 (2015).
35. Kichaev, G. et al. Leveraging polygenic functional enrichment to improve GWAS power. *Am. J. Hum. Genet.* **104**, 65–75 (2019).
36. Fritsche, L. G. et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat. Genet.* **48**, 134–143 (2016).
37. Benonisdottir, S. et al. Sequence variants associating with urinary biomarkers. *Hum. Mol. Genet.* **28**, 1199–1211 (2018).
38. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429 (2016).
39. Evangelou, E. et al. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nat. Genet.* **50**, 1412–1425 (2018).
40. Horton, J. D., Cohen, J. C. & Hobbs, H. H. PCSK9: a convertase that coordinates LDL catabolism. *J. Lipid Res.* **50**, S172–S177 (2009).
41. Raal, F. et al. Low-density lipoprotein cholesterol-lowering effects of AMG 145, a monoclonal antibody to proprotein convertase subtilisin/kexin type 9 serine protease in patients with heterozygous familial hypercholesterolemia: the Reduction of LDL-C with PCSK9 Inhibition in Heterozygous Familial Hypercholesterolemia Disorder (RUTHERFORD) randomized trial. *Circulation* **126**, 2408–2417 (2012).
42. Cohen, J. C., Boerwinkle, E., Mosley, T. H.Jr & Hobbs, H. H. Sequence variations in *PCSK9*, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* **354**, 1264–1272 (2006).
43. Willer, C. J. et al. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
44. Kent, S. T. et al. *PCSK9* loss-of-function variants, low-density lipoprotein cholesterol, and risk of coronary heart disease and stroke: data from 9 studies of Blacks and whites. *Circ. Cardiovasc. Genet.* **10**, e001632 (2017).
45. Saevarsdottir, S. et al. *FLT3* stop mutation increases FLT3 ligand level and risk of autoimmune thyroid disease. *Nature* **584**, 619–623 (2020).
46. Balder, J. W. et al. Genetics, lifestyle, and low-density lipoprotein cholesterol in young and apparently healthy women. *Circulation* **137**, 820–831 (2018).
47. Doege, K. J., Sasaki, M., Kimura, T. & Yamada, Y. Complete coding sequence and deduced primary structure of the human cartilage large aggregating proteoglycan, aggrecan. Human-specific repeats, and additional alternatively spliced forms. *J. Biol. Chem.* **266**, 894–902 (1991).
48. Allen, H. L. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
49. Doege, K. J., Coulter, S. N., Meek, L. M., Maslen, K. & Wood, J. G. A human-specific polymorphism in the coding region of the aggrecan gene: variable number of tandem repeats produce a range of core protein sizes in the general population. *J. Biol. Chem.* **272**, 13974–13979 (1997).
50. Roughley, P. J., Alini, M. & Antoniou, J. The role of proteoglycans in aging, degeneration and repair of the intervertebral disc. *Biochem. Soc. Trans.* **30**, 869–874 (2002).
51. Schwartz, N. B. & Domowicz, M. Chondrodysplasias. In *Reference Module in Biomedical Sciences* https://doi.org/10.1016/b978-0-12-801238-3.03764-8 (Elsevier, 2014).
52. Kiani, C. et al. Structure and function of aggrecan. *Cell Res.* **12**, 19–32 (2002).
53. Mukamel, R. E. et al. Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. Preprint at *bioRxiv* https://doi.org/10.1101/2021.01.19.427332 (2021).
54. Nielsen, J. B. et al. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat. Genet.* **50**, 1234–1239 (2018).
55. Park, C. Y. et al. SkNAC, a Smyd1-interacting transcription factor, is involved in cardiac development and skeletal muscle growth and regeneration. *Proc. Natl Acad. Sci. USA* **107**, 20750–20755 (2010).
56. Roselli, C. et al. Multi-ethnic genome-wide association study for atrial fibrillation. *Nat. Genet.* **50**, 1225–1233 (2018).
57. Kong, A. et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099–1103 (2010).
58. Hinch, A. G. et al. The landscape of recombination in African Americans. *Nature* **476**, 170–175 (2011).
59. McLaren, W. et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).
60. Touchman, J. W. et al. The genomic region encompassing the nephropathic cystinosis gene (*CTNS*): complete sequencing of a 200-kb segment and discovery of a novel gene within the common cystinosis-causing deletion. *Genome Res.* **10**, 165–173 (2000).
61. Rafi, M. A., Luzi, P., Chen, Y. Q. & Wenger, D. A. A large deletion together with a point mutation in the *GALC* gene is a common mutant allele in patients with infantile Krabbe disease. *Hum. Mol. Genet.* **4**, 1285–1289 (1995).
62. Luzi, P., Rafi, M. A. & Wenger, D. A. Characterization of the large deletion in the *GALC* gene found in patients with Krabbe disease. *Hum. Mol. Genet.* **4**, 2335–2338 (1995).
63. Tappino, B. et al. Identification and characterization of 15 novel *GALC* gene mutations causing Krabbe disease. *Hum. Mutat.* **31**, E1894–E1915 (2010).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Participants. A set of 3,622 individuals (1,656 males and 1,966 females) was selected for ONT sequencing, including 441 parent–offspring trios. Individuals were selected from a large set of Icelandic samples collected as part of disease-association efforts at deCODE genetics. The earliest year of birth (YOB) among both males and females was 1900. The latest YOB among males was 2003 and 2014 among females. The median YOB was 1955 for males and 1957 for females. The samples constitute a database of DNA-sequence variation in the Icelandic population combined with extensive phenotypic data, including information on blood levels of lipids for up to 113,355 genotyped individuals (approval no. VSN-15-023). The study population was described in detail previously[17,18,64,65]. All participants were Icelanders who donated biological samples for genotyping and provided informed consent as part of various genetic programs at deCODE genetics. The study was approved by the National Bioethics Committee of Iceland (approval nos. VSN-15-023 and VSN-05-097, with amendments). A subset of individuals ($n=102$) provided heart samples, with approval no. VSN-05-097.

DNA source. Most of the DNA samples sequenced in this study were isolated from whole blood ($n=3{,}524$). DNA from whole blood was extracted using the Chemagic method (PerkinElmer), an automated procedure that involves the use of M-PVA magnetic beads (https://chemagen.com). DNA samples were also isolated from heart tissue ($n=102$); four individuals provided both heart and whole-blood samples. Samples were received and subsequently stored in liquid nitrogen. Samples were cut to a smaller size on dry ice if needed. Lysis buffer and a sterile 5-mm steel bead were added to each sample before homogenization on a TissueLyser LT (Qiagen). DNA was extracted from the homogenized lysates using the MasterPure DNA Purification kit (Epicentre) following the manufacturer's protocol but with overnight proteinase K digestion. Isolated DNA samples were quantified using a Trinean DropSense, and integrity was assessed using the Fragment Analyzer capillary system from AATI.

Sample preparation. Sequencing libraries were generated using the SQK-LSK109 ligation kit from ONT. Sample input varied from 1 to 5 µg DNA, depending on the exact version of the preparation kit and the flow cell type used for PromethION sequencing. In total, 1,322 of 4,757 flow cells underwent partial DNA shearing using the Covaris g-TUBE to a mean fragment size of 10−15 kb. The remainder of the samples were not sheared (Supplementary Fig. 1, from March 2019 and onward). Library preparation started with DNA repair and A tailing using the NEBNext FFPE repair mix (M6630) and the NEBNext End repair/dA-tailing module (E7546), followed by clean up with AMPure XP beads. Adaptor ligation was performed using NEB T4 ligase (NEBNext Quick Ligation Module, E6056) and the ONT/LSK109 adaptor mix (AMX) and ligation buffer, respectively. Samples were again purified using AMPure XP beads, using the Long Fragment Buffer for the wash steps. Final sample elutions from the beads were performed using 15 µl elution buffer. Samples were quantified using a Qubit fluorimeter and diluted appropriately for loading onto the flow cells.

Sequencing. Samples were loaded onto PromethION R9.4.1 flow cells following ONT standard operating procedures. Sequencing was performed on PromethION devices. Data acquisition varied from 48 to 60 h per flow cell.

Basecalling. Squiggle data from PromethION sequencers were basecalled using Guppy (3,622 individuals, 4,757 flow cells). We ran Guppy Sequencing Pipeline Software for GPU machines, version 3.2.2 (643 flow cells) and version 3.3.0 (4,114 flow cells), using the 'flipflop' model (configuration file template_r9.4.1._450bps_large_flipflop.json from Guppy version 2.3.1) for PromethION flow cells with firmware 2.0.4 and the 'hac' model (using the corresponding configuration file template_r9.4.1._450bps_hac_prom.json) for firmwares 2.0.10, 2.0.12 and 2.0.14 (Supplementary Fig. 1).

Our oldest flow cells were originally basecalled using Albacore (a now deprecated basecaller from ONT). Mid-year 2019, we upgraded to Guppy version 3.2.2. This was reported in our preliminary study based on 1,817 individuals[66]. In autumn 2019, we upgraded Guppy to version 3.3.0 due to new PromethION firmware and re-basecalled all flow cells previously basecalled with Albacore to reduce error rates[67].

All 3,622 individuals basecalled with Guppy had a minimum reference-genome-aligned sequencing coverage of at least 10× at the time of analysis for SV discovery.

Read mapping. Basecalled reads were mapped to the human reference genome GRCh38 (ref. [20]) with minimap2 (ref. [19]) (version 2.14-r883), using the recommended option for ONT sequence-to-reference mapping (-x map-ont). In addition, we used the parameters '--MD -Y'. The aligned reads were sorted using SAMtools sort[68] and stored in a BAM file.

Generating SV candidates from an individual. As shown in Fig. 1b, after basecalling the raw reads, we aligned them to the GRCh38 reference genome using minimap2 and performed a sensitive SV prediction using Sniffles. We filtered the SV candidates using their alternate allele ratios and refined their breakpoints using SViper[21]. Next, we used SquiggleSVFilter on both the breakpoint refined SVs and filtered Sniffles SV calls (breakpoint unrefined) to verify their presence using the raw-signal-level data. Supplementary Fig. 2 shows the number of discovered SVs per individual, sorted by the aligned coverage of the individual.

SV prediction and breakpoint refinement. A set of preliminary variant predictions was obtained using Sniffles[8] (version 1.0.10) for each genome, in a highly sensitive fashion (using -s 3 and --ignore_sd) to minimize FNs due to the existence of low-coverage regions. Up to 30 supporting reads were reported per variant. Other optional parameters were left as default. Indels with different start and end chromosomes and of size larger than 1 Mb were discarded.

Next, deletions and insertions with alternate allele ratios below 0.2 and 0.05, respectively, were discarded as a pre-filter from raw Sniffles calls. A higher value was used for deletions, as the basecaller is deletion biased. We calculated the alternate allele ratio as the number of reads supporting the variant divided by the coverage at the variant site. SVs were then breakpoint refined with SViper[21] (https://github.com/DecodeGenetics/SViper/tree/cornercases) using SRS data when possible ('Breakpoint and variant refinement with SViper' in the Supplementary Information).

SV filtering using squiggles (SquiggleSVFilter). We developed SquiggleSVFilter (https://github.com/DecodeGenetics/nanopolish/tree/squigglesv) to filter false SV predictions using the signal-level raw ONT sequencing data, that is, the squiggle. SquiggleSVFilter employs the squiggle-versus-sequence log-likelihood-score function provided by Nanopolish[22] and compares the log-likelihood scores of the predicted alternate allele versus the reference allele on the squiggle around both of the SV breakpoints. The likelihood score is essentially the probability of the signal-level raw data given a candidate sequence[22]. Nanopolish uses the events, which are the step-wise changes in the measured electrical currents, as the signal data in its log-likelihood-score function. Accessing an event interval over a predicted SV breakpoint requires a mapping of the read sequence indices to reference-genome coordinates (that is, a reference-aligned BAM file) and to event indices, called an 'event table'. To achieve this, we generated basecalls and event tables for reads that support a predicted SV, using a modified version of Scrappie (https://github.com/DecodeGenetics/scrappie/tree/v1.3.0.events) and mapped these reads to the reference genome using minimap2 with parameters as described in Read mapping. Using the event table, we found the 'event slices' corresponding to the read regions spanning the SV breakpoints ('Accessing event-slices of interest in SquiggleSVFilter' in the Supplementary Information). Finally, we calculated the raw signal-versus-sequence log-likelihood scores using the reference and alternate allele sequences for both event slices and used their difference to support or reject a candidate variant. We supported a variant if at least three reads obtained a log-likelihood score difference of at least 1.92 for either of the event slices. A sample execution is provided in https://github.com/DecodeGenetics/SquiggleSV_samplerun.

SV merging. Most SVs are carried by multiple individuals and thus will be rediscovered, potentially with slightly different representations across carriers, varying in length and location. To eliminate such redundancies, we applied the following SV-merging approach, in which we represent SVs as vertices in a graph and find cliques representing merged SVs.

1. Preprocessing: we first identified TR SVs and then preclustered them into disjoint SV groups to reduce the input sizes for the following step ('SV merging preprocessing' in the Supplementary Information).
2. Finding SV cliques: we found cliques within each detached SV group generated in (1), independently for TR SVs and non-TR SVs as defined in (1). We used an undirected graph $G(V,E)$, where each vertex $v \in V$ represents an SV, and each edge $e \in E$ between vertex v_i and vertex v_j is drawn if distance $d(v_i, v_j)$ is at most D and assigned $d(v_i, v_j)$ as the edge length, where distance is

$$d\left(v_i, v_j\right) = 1 - \begin{cases} \frac{\min\left(v_i^e, v_j^e\right) - \max\left(v_i^b, v_j^b\right)}{\max\left(l\left(v_i\right), l\left(v_j\right)\right)} & \text{if } v_i \text{ and } v_j \text{ overlap}, \\ 0 & \text{otherwise}, \end{cases}$$

where v_i and v_j represent two SVs, $l(v_i)$ is the length of v_i, and v_i^e and v_i^b represent the ending and beginning sites of v_i.

We then formulated the SV merging as a corrupted cliques problem, where, given a graph G, the aim is to transform G into a clique graph with the smallest number of edge additions and removals, such that a clique represents a single merged SV. To solve this, we employed the Cluster Affinity Search Technique algorithm[69] (https://github.com/DecodeGenetics/sv-merger). We computed the average distance of vertex v_i to the cluster C as a weighted average, such that

$$m\left(v_i, C\right) = \frac{\sum_{v_j \in C} d\left(v_i, v_j\right) \times w_{I(v_j)}^{C}}{\sum_{v_j \in C} w_{I(v_j)}^{C}}$$

where $I(v_j)$ is the individual v_j is discovered in, and $w_{I(v_j)}^{C}$ is the weight of v_j for cluster C, set as the inverse of the number of SVs from individual $I(v_j)$ found in the cluster *C*. The weighted average limits bias in the clique formation toward individuals with multiple SV calls at similar positions (detected with different approaches, for example, breakpoint refined and unrefined SV calls). Likewise, the degree of a vertex v_i was also calculated as the number of individuals $I(v_j)$ that it has an edge with. We used $D = 0.5$ for non-TR SVs and $D = 0.15$ for TR SVs. We decided to be more conservative in the merging of TR SVs than that for non-TR SVs to be able to distinguish alleles with different numbers of repeats within the TR regions and because we artificially increased their overlap by changing their begin sites ('SV merging pre-processing' in the Supplementary Information).

3. Finding SV clique representatives: we represented a clique using an SV with the most common (begin site, length) attribute set among all the clique SVs. If there was no such single most common attribute set, we sorted the SVs with the most common attribute sets using the frequency of their begin and end site and length among all the clique SVs, separately, in the given order. In this step, we used the original begin and end sites for TR SVs. For all TR SV clique representatives within the same TR region, we assigned their position as the most common original begin or end site among all SVs of all cliques within the TR region. All further ties were first broken by prioritizing begin sites over end sites and then using the alternate allele ratios. If ties could not be broken, the median begin site was used.

SV clique representatives were finally presented as merged SVs. We note that variants discovered with length between 30 bp and 50 bp were not filtered out during the SV-merging step to prevent clique formation with incomplete data.

SRS genotyping with GraphTyper. We provided the merged SV set to GraphTyper[7,23] version 2.6, which generates an augmented graph genome using SV predictions, together with previously discovered SNPs and indels[7], for population-scale genotyping. The variants were genotyped on the set of 10,000 individuals with SRS data, using three genotyping models for deletions and insertions. For insertions, we use the models first breakpoint (B1), second breakpoint (B2) and their aggregate (AG). For deletions, we use the models breakpoint (B), coverage (C) and their aggregate (AG). As GraphTyper does not support multiallelic SV genotyping, we added multiallelic SVs as separate biallelic variants.

LRS genotyping (LRcaller). LRcaller is a proof-of-concept genotyping algorithm that genotypes SVs directly from ONT sequencing reads. We introduced LRcaller version 0.1 in our previous study[66], in which we genotyped 1,817 individuals with LRS data. In this study, we introduce LRcaller version 0.2 (https://github.com/DecodeGenetics/LRcaller), which now allows for the genotyping of multiallelic variants, better treatment of TRs and additional genotyping models compared to version 0.1.

Each breakpoint was genotyped independently, resulting in two sets of genotypings for the canonical deletion and insertion variants identified in this study, corresponding to the left and right breakpoints (Extended Data Fig. 2). Note that the algorithm processes each variant independently; that is, each variant is genotyped without considering other variants in the region, which may lead to suboptimal behavior when there are multiple neighboring variants.

To capture multiple types of information that could represent an SV, we used five genotyping models: direct alignment (AD), variant alignment (VA), reference-aware variant alignment (VAr), presence (PR) and joint (J). We used the reads overlapping a breakpoint and two sets of evidence for genotyping: (AD) from an alignment of a subread to the reference and alternate alleles and (VA, VAr, PR) from the alignment present in the BAM file as aligned by minimap2. The joint model (J) uses both sets of evidence. See 'LRS genotyping (LRcaller) models' in the Supplementary Information for further explanation of the models.

Genotyping with LRcaller. Variants were genotyped independently for the left and right breakpoint using the five different models presented above for each variant, producing a total of ten genotypes per individual–marker pair.

Phasing and imputation of structural variants. For each marker, we produced a total of 13 different genotypes, three from GraphTyper (Illumina) and ten from LRcaller (ONT). We phased and imputed all genotyped variants into the haplotypes of 166,281 Icelanders, using a previously described methodology[18,24,25]. We considered variants with an imputation information score greater than 0.9 and a leave-one-out R^2 greater than 0.5 as 'imputed'. After imputing our SVs, we also acquired the allele frequencies of the variant alleles and the reference allele spanning a multiallelic SV. We used the reference alleles to deduce haplotypes carrying a variant allele of unresolved size or type. Carriers of a variant allele of unresolved type were determined as those not containing the reference allele of a secondary-form ('Creating a Variant Call Format (VCF) file for the merged SV set' in Supplementary Information) multiallelic SV comprising both insertion and deletion variant alleles. We refer to variant alleles with an unresolved size or type as 'non-reference alleles'.

SV filtering. After genotyping and imputing our merged SV set using both LRS and SRS sequencing, we filtered the data using imputation accuracy and other filters in five stages as described in 'SV filtering stages' in the Supplementary Information. We report the set of variant alleles from our final filtering step as 'high-confidence' variants.

Comparison of the merged SV set to other SV datasets. To calculate FP and FN rates of our high-confidence SV alleles, we developed a statistic, variants inconsistent with HG002 (VIH), and compared our dataset to the SV sets provided by Audano et al.[16], gnomAD-SV[11] and Zook et al.[5] on HG002 (https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_SVs_Integration_v0.6/HG002_SVs_Tier1_v0.6.vcf.gz).

We expect most high-frequency variants to be present in populations of similar heritage, such as European populations. Variants found in population A but not found in population B can be due to (1) fixation of the variant in population B, (2) different representation of the variants between the two datasets, (3) FP calls in dataset A and (4) FN calls in dataset B.

As variant representation differs between variant callers, we used a relaxed method for determining whether a variant discovered in dataset A was also present in dataset B, counting the variant as present if there was a variant in dataset B occurring within 500 bp of the start position of the variant discovered in dataset A (given GOR files for datasets A and B, the command run was gorpipe 'A.gor | join --snpseg --f 500 B.gor').

Variants inconsistent with HG002 statistic. As HG002 has been extensively characterized within tier 1 regions, we can use it to estimate FP rates within these regions. We developed the statistic VIH. VIH takes as input two studies, A and B, along with their variants and the variant frequencies from study A. VIH assumes that (1) the same classes of variants have been characterized within HG002 as in study A and (2) no drift in variant frequency has occurred between study A and HG002. We compute

$$\mathrm{VIH}\,(A,B) = \frac{|A-B|}{|A|} \times \left(1 - \frac{c}{E}\right)$$

where $|.|$ represents cardinality, $|A-B|$ the number of variants in study A missed by study B. c is the number of variants in $|A-B|$ found in HG002, and E is the expected number of variants in $|A-B|$ found in HG002, given the variant frequencies in study A. $\frac{c}{E}$ is therefore an estimated true positive rate of the variants in $|A-B|$, and $\left(1-\frac{c}{E}\right)$ is an estimate of the FP rate. We use this statistic as a surrogate for the FP rate for study A, although it may in part be explained by population drift, differences in variant classification between studies and FNs in the HG002 truth set.

Association testing. We tested our SVs for association with LDL levels based on the linear mixed model implemented in BOLT-LMM[70]. We used BOLT-LMM to calculate leave-one-chromosome-out residuals, which we then tested for association using simple linear regression. A generalized form of linear regression was used to test for association of phenotypes with SVs. We assume that the phenotypes follow a normal distribution with a mean that depends linearly on the expected allele at the variant and a variance–covariance matrix proportional to the kinship matrix[71]. We used LD score regression to account for distribution inflation in the dataset due to cryptic relatedness and population stratification[72]. The inflation factors were computed from a set of SNP and indel sequence variants. Using a set of about 1.1 million SNP and indel sequence variants, we regressed the χ^2 statistics from a genome-wide association scan against LD score and used the intercept as a correction factor. Effect sizes based on leave-one-chromosome-out residuals were shrunk, and we rescaled them based on the shrinkage of the 1.1 million variants used in LD score regression.

Comparison to the GWAS catalog. We downloaded version 1.0 of the GWAS catalog with all associations (https://www.ebi.ac.uk/gwas/docs/file-downloads) on 23 July 2020 (gwas_catalog_v1.0-associations_e100_r2020-07-14.tsv). SNPs and indels in the GWAS catalog were matched with in-house SNPs using exact-coordinate matching, and two markers were assumed to be the same if they had the exact same coordinate in GRCh38.

An in-house tool was used to compute correlations between SNPs and indels imputed into 166,281 Icelanders and SVs imputed into the same set. Correlations were limited to windows of 500 kb, such that a correlation between a SNP or indel and an SV was observed if and only if they were within 500 kb of each other. We provide our results in Supplementary Data 4.

Association testing for the number of motifs in *ACAN*. Using SV alleles that we identified in a VNTR in *ACAN* (Table 1), we calculated expected motif change per haplotype in an individual as

$$c\,(h) = \sum_{v \in V} n\,(v) \times P(v,h)$$

where $v \in V$ represents the SV alleles, and $n(v)$ represents the number of motif change in SV allele v, which we calculated by rounding the allele lengths divided by the motif length to an integer (negative for deletions). $P(v,h)$ is haplotype carrier probability for haplotype h carrying the SV allele v, calculated during imputation

(Phasing and imputation of structural variants). We performed a unit-based normalization of the $c(h)$ values to use as haplotype carrier probabilities in association testing.

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

Data availability

Access to these data is controlled; the sequence data cannot be made publicly available because Icelandic law and the regulations of the Icelandic Data Protection Authority prohibit the release of individual-level and personally identifying data. Data access can be granted only at the facilities of deCODE genetics in Iceland, subject to Icelandic law regarding data usage. Anyone wishing to gain access to the data should contact K.S. (kstefans@decode.is). Icelandic law allows for unimpeded sharing of summary-level data. Data access consists of Supplementary Data 1–5 as described below, alongside the VCF and index files for the high-confidence SV alleles at https://github.com/DecodeGenetics/LRS_SV_sets.

Code availability

Codes are available as follows: SViper, modified, used in this study (https://github.com/DecodeGenetics/SViper/tree/cornercases); SViper, original repository (https://github.com/smehringer/SViper); Scrappie, modified, used in this study (https://github.com/DecodeGenetics/scrappie/tree/v1.3.0.events); Scrappie, original repository (https://github.com/nanoporetech/scrappie); SquiggleSVFilter (https://github.com/DecodeGenetics/nanopolish/tree/squigglesv); sample execution of SquiggleSVFilter with input and expected output data (https://github.com/DecodeGenetics/SquiggleSV_samplerun); sv-merger, to form SV cliques using the Cluster Affinity Search Technique algorithm (https://github.com/DecodeGenetics/sv-merger); LRcaller (https://github.com/DecodeGenetics/LRcaller).

References

64. Nioi, P. et al. Variant *ASGR1* associated with a reduced risk of coronary artery disease. *N. Engl. J. Med.* **374**, 2131–2141 (2016).
65. Helgadottir, A. et al. Variants with large effects on blood lipids and the role of cholesterol and triglycerides in coronary disease. *Nat. Genet.* **48**, 634–639 (2016).
66. Beyter, D., Ingimundardottir, H., Eggertsson, H. P. & Bjornsson, E. Long read sequencing of 1,817 Icelanders provides insight into the role of structural variants in human disease. Preprint at *bioRxiv* https://doi.org/10.1101/848366 (2019).
67. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 129 (2019).
68. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
69. Ben-Dor, A., Shamir, R. & Yakhini, Z. Clustering gene expression patterns. *J. Comput. Biol.* **6**, 281–297 (1999).
70. Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
71. Benonisdottir, S. et al. Epigenetic and genetic components of height regulation. *Nat. Commun.* **7**, 13490 (2016).
72. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

Acknowledgements

We thank J. Simpson, our colleagues from deCODE genetics/Amgen Inc. and employees of ONT for their help and advice. We also thank all research participants who provided a biological sample to deCODE genetics.

Author contributions

D.B. implemented software, with additional software implemented by H.I., H.P.E., S.K., S.M., G.H. and B.V.H. D.B. and B.V.H. wrote the paper with input from H.I., A.O., H.P.E., E.B., H.J., B.A.A., S.K., M.T.H., S.A.G., R.P.K., G.H., G.P., O.A.S., A.H., U.T., H.H., D.F.G., P.S., O.T.M. and K.S. H.I. implemented the analysis pipelines, with input from D.B., S.K., S.A.G., S.T.S., G.M. and B.V.H. D.N.M. and O.T.M. performed ONT sequencing. Aslaug Jonasdottir and Adalbjorg Jonasdottir performed PCR validation experiments. G.E., I.O. and O.S. acquired LDL measurements. H.H. and B.T. acquired heart tissues. B.V.H. and K.S. conceived and supervised the study. All authors approved the final version of the manuscript.

Competing interests

D.B., H.I., A.O., H.P.E., E.B., H.J., B.A.A., S.K., M.T.H., S.A.G., D.N.M., Aslaug Jonasdottir, Adalbjorg Jonasdottir, R.P.K., S.T.S., G.H., G.P., O.A.S., G.M., A.H., U.T., H.H., D.F.G., P.S., O.T.M., B.V.H. and K.S. are employees of deCODE genetics/Amgen. The remaining authors declare no competing interests.

Additional information

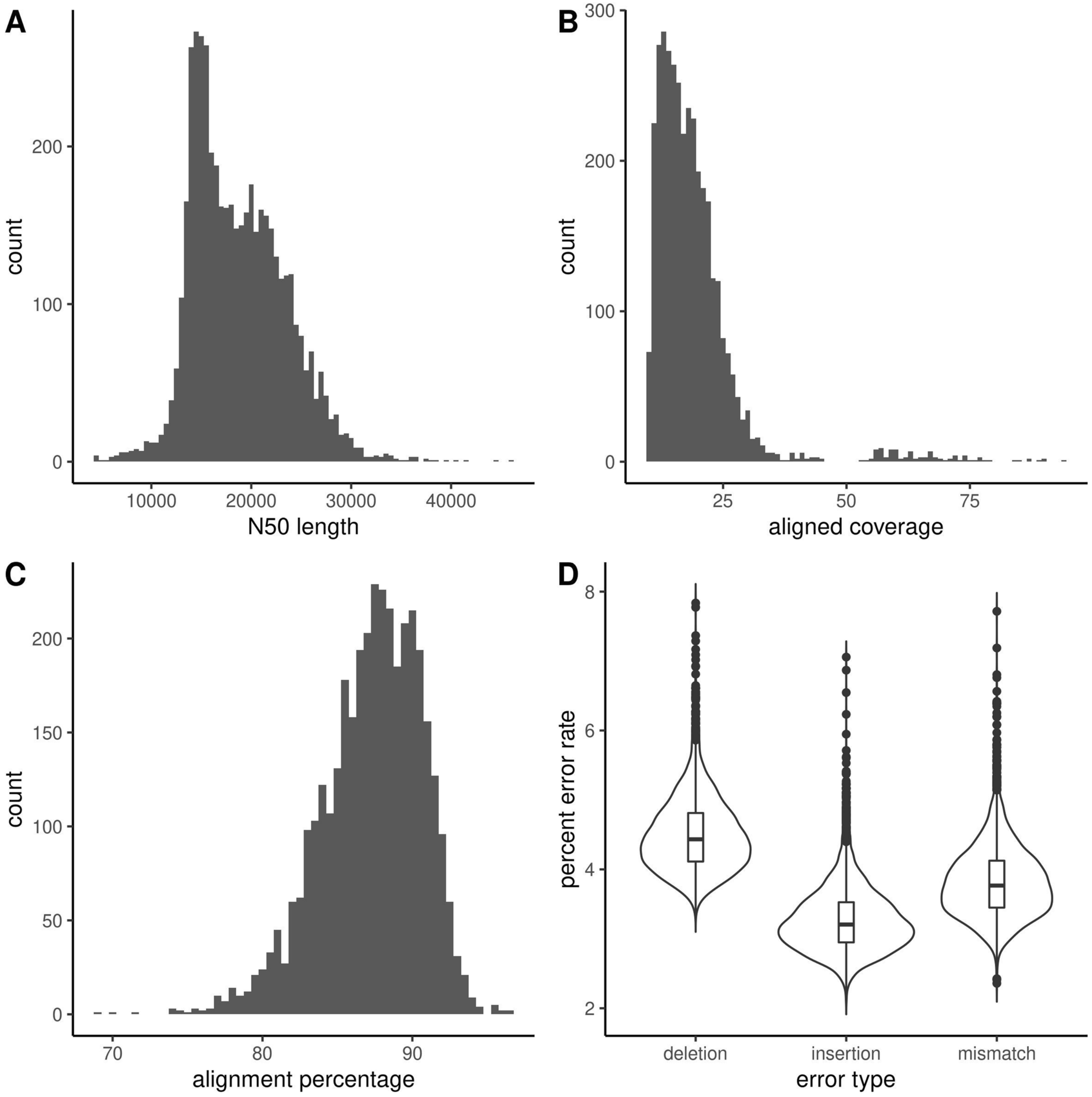
Extended data is available for this paper at https://doi.org/10.1038/s41588-021-00865-4.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00865-4.

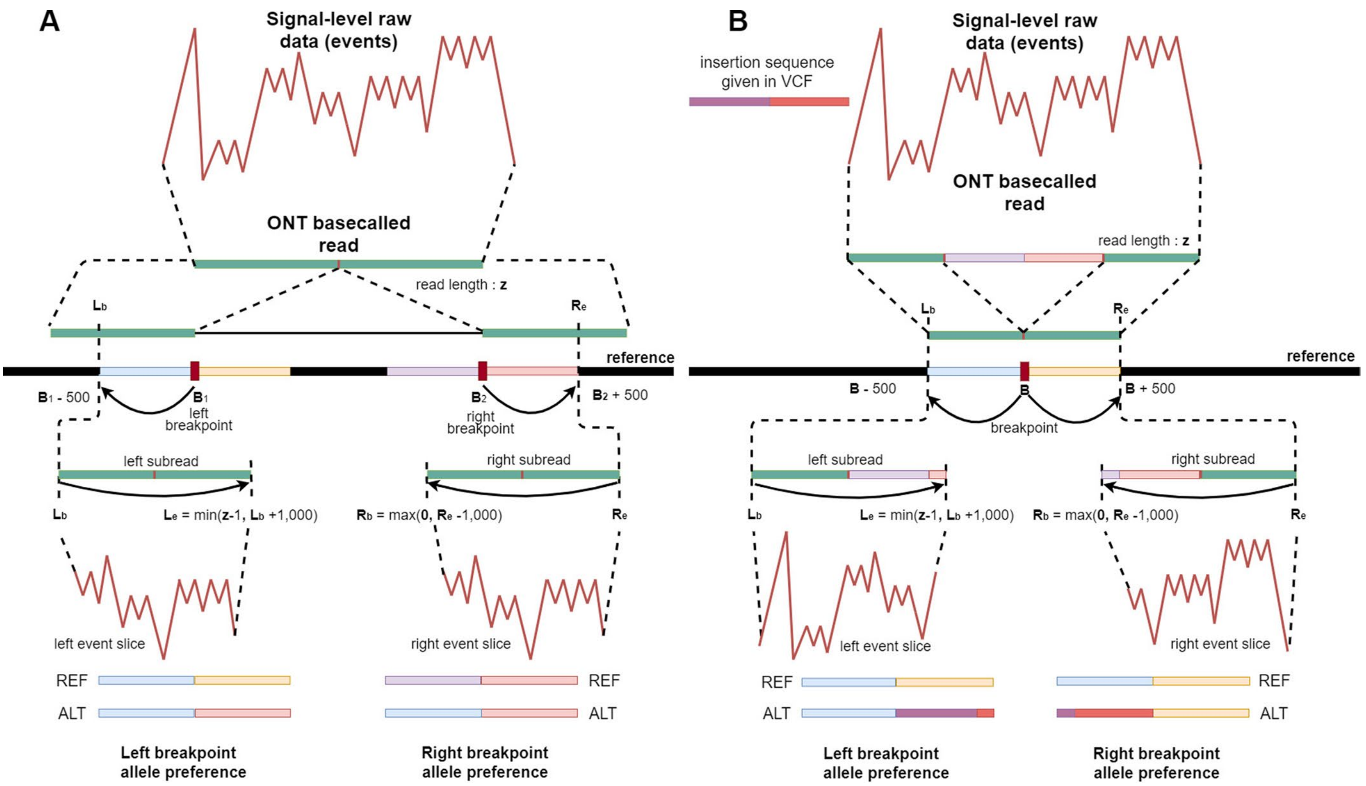
Correspondence and requests for materials should be addressed to B.V.H. or K.S.

Peer review information *Nature Genetics* thanks Mark Chaisson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

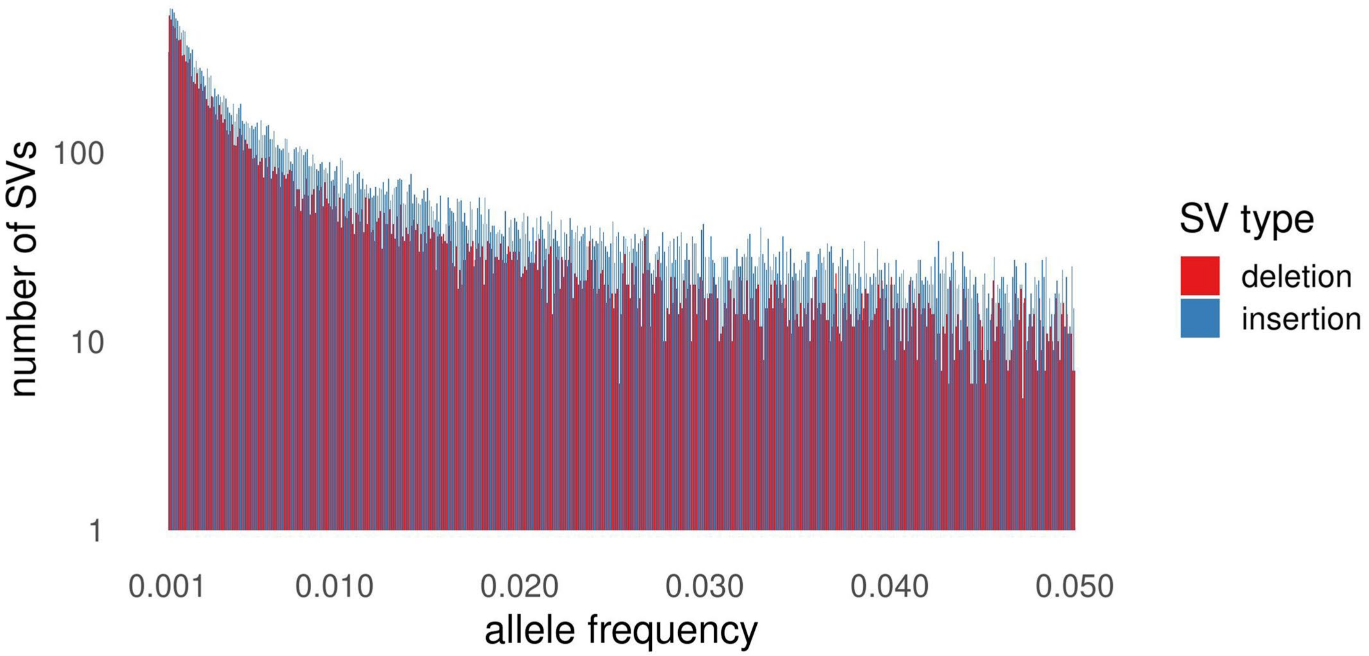
Reprints and permissions information is available at www.nature.com/reprints.



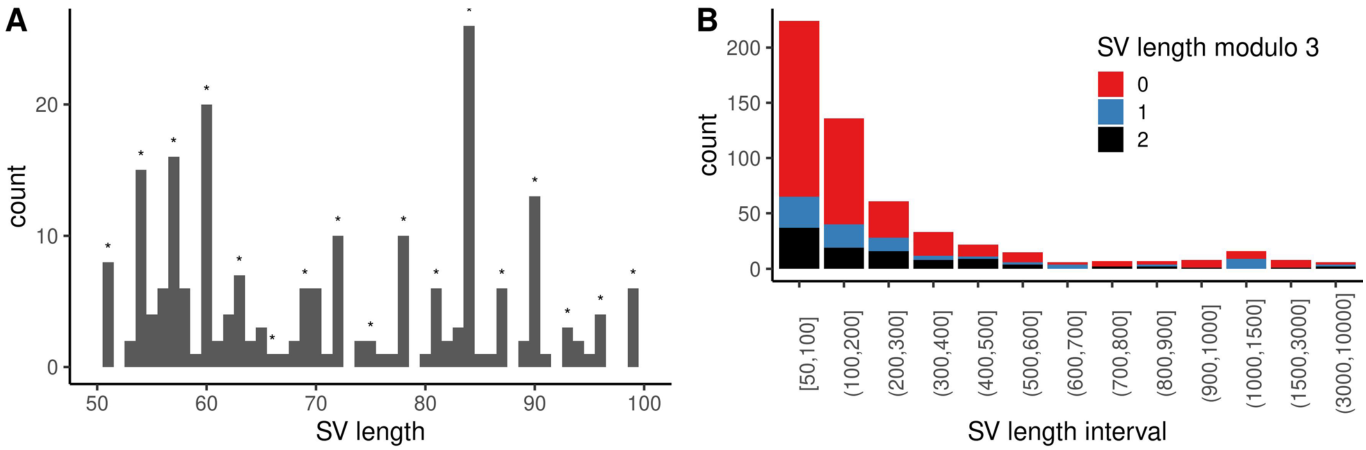
Extended Data Fig. 1 | Oxford Nanopore Technologies (ONT) long-read sequencing statistics. a, N50 length per flowcell (N = 4,757 flowcells) prior to GRCh38 alignment. **b,c,d**, Aligned coverage, alignment percentage, and error rates stratified by type, per individual (N = 3,622 individuals). Statistics are computed over sequenced reads longer than 3000 bp. In panel **d**, box limits indicate upper and lower quartiles, centre line indicates median, and whiskers indicate ±1.5 times the interquartile range.



Extended Data Fig. 2 | SquiggleSVFilter overview. Given a candidate structural variant (SV), and an SV supporting read, SquiggleSVFilter first identifies the subread of the ONT basecalled read overlapping the SV, using the reference alignment BAM file. Next it finds the squiggle slice of the identified subsequence using the event table. For both the left and right flanks around the variant, it determines the reference and alternative sequences given the candidate variant, and computes their raw data-vs-sequence log likelihood scores with the squiggle slice. A sufficiently high log likelihood score difference for the alternate allele marks the read as an SV supporting read.



Extended Data Fig. 3 | Allele frequency distribution of SVs at low frequency. SVs are binned at 0.01% for alleles with 0.1% to 5% frequency.



Extended Data Fig. 4 | Length and modulo distributions of structural variants (SVs) that are contained within exons. a, Length distribution of SVs with lengths between 50 and 100. Stars denote lengths divisible by 3. (N = 224 markers). **b**, Modulo distribution of SV lengths across length intervals. (N = 549).

nature research

Corresponding author(s): Bjarni V. Halldorsson

Last updated by author(s): Mar 9, 2021

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	Description of data collection and software used is available in sections Participants, DNA source, Sample preparation, Sequencing, Basecalling and Read mapping under Methods, for long read sequencing (LRS) data. See reference 18, "Sequence variants from whole genome sequencing a large group of Icelanders" for short read sequencing (SRS) data. Samples were loaded onto PromethION R9.4.1 flowcells and basecalled using Guppy Sequencing Pipeline Software for GPU machines v3.2.2 (643 flowcells) and v3.3.0 (4,114 flowcells) using the flipflop model (configuration file template_r9.4.1._450bps_large_flipflop.jsn from guppy version 2.3.1) for PromethION flowcells with firmware 2.0.4 and the hac model (using the corresponding configuration file template_r9.4.1._450bps_hac_prom.jsn) for firmwares 2.0.10, 2.0.12 and 2.0.14. Basecalled reads were mapped using minimap2 v2.14-r883 and sorted using samtools v1.9.
Data analysis	Description of data analysis code is included in Methods, and github repositories provided in section Code availability. SVs are detected on mapped reads using Sniffles v1.0.10, and breakpoint refined using SViper (https://github.com/DecodeGenetics/SViper/tree/cornercases). Next, breakpoint refined and unrefined SVs are filtered using SquiggleSVFilter (https://github.com/DecodeGenetics/nanopolish/tree/squigglesv). SVs are merged (see Methods and Supplementary Information, https://github.com/DecodeGenetics/sv-merger) and genotyped using LRcaller v0.2 (https://github.com/DecodeGenetics/LRcaller) and Graphtyper v2.6.

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 SVs discovered in this study and correlations with GWAS catalog markers are included as supplementary data. Icelandic law allows for unimpeded sharing of summary-level data. However, the law does not allow the sharing of individual-level data on genotypes and phenotypes outside of Iceland.

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	The sample sizes used in the study were not predetermined. The number of LRS individuals is sufficiently large to perform accurate imputation using LRS genotypes, demonstrated by the imputed 120,108 SV alleles out of 133,886 high-confidence SV alleles.
Data exclusions	No data were excluded from the analyses.
Replication	We provide parent support rates on discovered structural variants in 441 children (Fig. S4) to estimate our false positive rates. Also, we include the code on tools we developed in Github in the section Code availability. PCR and agarose gel electrophoresis were carried out using standard protocols on DNA from two individuals of each expected genotype: non-carriers, heterozygotes and homozygotes (when available).
Randomization	Not relevant. We did not allocate samples/participants into experimental groups.
Blinding	Not relevant as there were no group allocation of samples.

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	Among 3,622 long-read sequenced individuals 1,656 are male, and 1,966 are female. Earliest year of birth (YOB) among both males and females is 1900. Latest YOB among is 2003 among males, and 2014 among females. Median YOB is 1955 for males, and 1957 for females.
Recruitment	All participants were Icelanders who donated biological samples to explore the interplay between the genetic variation and phenotypic diversity, and were provided informed consents as part of various genetic programs at deCODE genetics, Reykjavik, Iceland. The sample set may be enriched for families/close-relatives, and cases compared to normals, however is sufficiently large to represent the Icelandic population.

Ethics oversight	All participants were Icelanders who donated biological samples for genotyping and provided informed consents as part of various genetic programs at deCODE genetics, Reykjavik, Iceland. The study was approved by The National Bioethics Committee of Iceland (approval no. VSN-15-023 and VSN 05-097, with amendments). A subset of individuals (N = 102) provided heart samples, with approval no. VSN-05-097.

Note that full information on the approval of the study protocol must also be provided in the manuscript.