Long read sequencing of 1,817 Icelanders provides insight into the role of structural variants in human disease

Authors: Doruk Beyter[1], Helga Ingimundardottir[1], Hannes P. Eggertsson[1,2], Eythor Bjornsson[1,3,4], Snaedis Kristmundsdottir[1,5], Svenja Mehringer[6], Hakon Jonsson[1], Marteinn T. Hardarson[1], Droplaug N. Magnusdottir[1], Ragnar P. Kristjansson[1], Sigurjon A. Gudjonsson[1], Sverrir T. Sverrisson[1], Guillaume Holley[1], Gudmundur Eyjolfsson[7], Isleifur Olafsson[8], Olof Sigurdardottir[9], Gisli Masson[1], Unnur Thorsteinsdottir[1,3], Daniel F. Gudbjartsson[1,2], Patrick Sulem[1], Olafur T. Magnusson[1], Bjarni V. Halldorsson[1,5]*, Kari Stefansson[1,3]*

Affiliations: 1 deCODE genetics/Amgen, Inc., Sturlugata 8, 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 Science and Engineering, 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

*Correspondance to: Bjarni V. Halldorsson, deCODE genetics/Amgen Inc., Sturlugata 8, 101 Reykjavik, Iceland. bjarnih@decode.is, Phone: 354-5701808, fax 354-5701901

Kari Stefansson, deCODE genetics/Amgen Inc., Sturlugata 8, 101 Reykjavík, Iceland. kstefans@decode.is, Phone: 354-5701900, fax 354-5701901.

ABSTRACT: Long-read sequencing (LRS) promises to improve characterization of structural variants (SVs), a major source of genetic diversity. We generated LRS data on 1,817 Icelanders using Oxford Nanopore Technologies, and identified a median of 23,111 autosomal structural variants per individual (a median of 11,506 insertions and 11,576 deletions), spanning cumulatively a median of 9.9 Mb. We found that rare SVs are larger in size than common ones and are more likely to impact protein function. We discovered an association with a rare deletion of the first exon of *PCSK9*. Carriers of this deletion have 0.93 mmol/L (1.36 sd) lower LDL cholesterol levels than the population average (p-value = $2.4 \cdot 10^{-22}$). We show that SVs can be accurately characterized at population scale using long read sequence data in a genomewide non-targeted fashion and how these variants impact disease.

Human sequence diversity is partially shaped by structural variants[1] (SVs); genomic rearrangements affecting at least 50 bp of sequence in forms of insertions, deletions, inversions, or translocations. The number of SVs carried by each individual is less than the number of single nucleotide polymorphisms (SNPs) and short (< 50 bp) insertions and deletions (indels), but their greater size gives them a higher probability to have a functional role[2], as they potentially affect cumulatively a similar (or greater) number of base-pairs[3].

Large-scale genetic studies mostly rely on whole genome short read sequencing (SRS, commonly 100-200 bp), where SNPs and indels can be fairly reliably identified[4,5]. The size limit of the short reads, however, makes the discovery, genotyping and characterization of SVs difficult[6]. The number of SVs found per individual (2-8k) in large scale studies using SRS[3,7–9] is much smaller than the 23-31k SVs observed in recent efforts, including fifteen individuals sequenced with long reads at high coverage (70x)[10] and three trios sequenced with several technologies[11].

Long-read sequencing (LRS), with read lengths of several kilobases (kb), aims to detect SVs with greater accuracy: SVs are commonly found by mapping and comparing sequence reads to a reference genome, and LRS reads can be mapped more accurately than SRS reads[6]. Long-reads are also more likely to cover entire SVs, enabling better breakpoint resolution and length determination. Long-reads, however, currently have high sequencing error rate, frequently over 10%, depending on samples, sequencing technology, and protocol[6]. In particular, insertion and deletion error rate imbalances in ONT sequencing can result in artifacts[6,12], as well as failure in SV identification. The introduction of artifacts can be especially challenging in large scale studies, as accumulating false-positives (FP) may result in FP-dominated results and hinder downstream analysis, such as genome-wide associations. Although previous work on detecting and characterizing SVs in human genomes using long-reads[6,10,13–15] is available on select small datasets, analysis at large scale has not been reported.

We present the first application of Oxford Nanopore Technologies (ONT) sequencing at a population scale. We sequenced 1,817 Icelanders, including 369 trios, recruited as part of various studies at deCODE genetics[16]. DNA was isolated from whole blood (n = 1,698) and heart tissue (n = 119) and sequenced with ONT GridION and PromethION sequencing machines (Methods). SRS and DNA chip data were also available for all but 24 of these individuals[17]. We introduce a number of tools and approaches to facilitate long-read SV analysis at scale. We developed SV filters to analyze sequences with high error rates, and a heuristic for merging SVs detected using ONT long-reads across a large dataset. We also developed an LRS genotyper for joint genotyping.

We observed a median LRS aligned coverage of 14.7x (range: 0.05-55.4x, Methods, Fig. 1A). The raw sequence data were basecalled and mapped[18] to the human reference genome. A median of 88% of base-pairs aligned to the reference and the median sequencing error rate was estimated to be 15.2% per individual (Fig. 1A). Half of all sequenced base-pairs (N50)

belonged to reads longer than 14,763 bps (Fig. 1A). Sequencing errors were biased towards deletions; an average error of 3.8% for insertions, 6.7% for deletions, and 4.8% for substitutions (mismatches) was observed per individual (Fig. 1B). As bias may lead us to miss SVs[14], we re-basecalled our data using an insertion biased configuration (Fig. S1), where we observed the converse error rate behavior (Fig. 1C).

We generated a high-confidence SV set in four stages: (i) discovery, (ii) merging across individuals, (iii) genotyping, and (iv) imputation (Fig. 2A). We discovered SVs using an initial set of 822 LRS individuals available at the time, independently for both the deletion and insertion biased basecalls. We began (Fig. 2B) by finding a set of highly sensitive SV predictions[6] and refined these variants at predicted breakpoint sites using SRS data, if possible[19] (Methods). We then verified their presence using the raw signal-level data[20] (Methods, Fig. S6), to alleviate potential basecalling and alignment errors. We merged SVs discovered across individuals and genotyped them twice, independently using 1,817 LRS and 5,000 SRS Icelanders[5,21] (Methods). We finally imputed the genotyped variants into a total of 166,261 Icelanders[17,22,23] and constructed a set of high-confidence SVs, based on imputation accuracy.

We identified 68,050 SVs after merging SVs across individuals, avoiding double counting variants occurring at a similar location and of similar size. We saw a more balanced number of insertions (32,354) and deletions (35,696), compared to the greater deficit of insertions discovered using SRS[3,7], where identification of deletions is generally easier. Almost all high frequency variants in existing SV call sets[3,10] were also found in our dataset (Fig. 3A). We categorized 47,936 SVs (22,970 insertions and 24,966 deletions) as high-confidence, in addition to 523 fixed (100% frequency) variants, by imputing the genotyped variants into long-range phased haplotypes from 166,261 chip typed Icelanders. Imputation was not attempted for the 2,998 variants on the X chromosome. We identified a median of 23,111 autosomal SVs

(a median of 11,506 insertions and 11,576 deletions, Methods, Fig. 3E-F) per individual, of which a median of 15,605 (a median of 7,286 insertions and 8,304 deletions) are high-confidence or fixed, spanning a median of 9.9 Mb.

The number of variants in our high-confidence SV set rapidly decreases with length, with two noticeable peaks at sizes around 300bp and 6kb, corresponding to SINE and LINE elements (Fig. 3B-C, Fig. S2), consistent with previous reports[10,14]. We found more SVs, particularly tandem repeats (TRs), in telomeres[10] (Fig. 3D, Fig. S3), a reflection of the sequence content of telomeres and the high mutation rate of TRs[10,24]. The number of variants detected decreases with increased allele frequency, with the exception of variants that are fixed or close to fixed (Fig. 3H).

The LRS data improved the genotyping and discovery of SVs. Of 47,936 high-confidence SVs 17,590 and 3,996 SVs were imputed only from LRS and SRS genotyping, respectively. The number of variants discovered in LRS per individual increases with sequence coverage, but appears to saturate at around 30x (Fig. S4). Variants missed in a single individual may be recovered by genotyping SVs discovered in other individuals[21,25], particularly if there are many carriers. Rare variants are more likely to be missed entirely.

As harmful variants are subjected to negative selection, we expect that rare SVs (allele frequency < 1%) are more likely to have a functional impact. Indeed, while rare variants account for 30.2% of high-confidence SVs, they represent more than half (55.6%) of the SVs overlapping a coding exon (as annotated by RefSeq[26]), (p-value: $2.2 \cdot 10^{-33}$, Fisher's exact test). Consequently 2.0% and 0.7% of rare and common (allele frequency ≥ 1%) variants, respectively, overlap a coding exon. Excluding TRs, the high confidence rare variants are disproportionately large; 15.6% of common variants were longer than 1 kb, but 26.2% of rare variants (p-value: $1.3 \cdot 10^{-114}$, Fisher's exact test).

Variants inside coding exons that are multiples of three in length generally result in addition or removal of amino-acids from translated proteins, while those that are not result in translational frameshift and generally non-functional proteins. In keeping with results[23] for indels, among 212 variants contained in a single coding exon, we observed a deficit in variant lengths that are not multiples of three, 74 (35%) compared to two thirds (141) expected (p-value: $3.8 \cdot 10^{-21}$, binomial test). These results show that SVs that result in translational frameshift are selected against due to their phenotypic impact.

We asked how the SVs identified impact phenotype and disease. To answer this, we correlated the high-confidence SVs with variants reported to associate with phenotypes in the GWAS catalog[27]. Of the 90,440 unique markers reported in the GWAS catalog 6,324 are correlated ($r^2 \geq 0.8$) with 3,725 SVs in our dataset, suggesting possible functional explanations for these associations (Supplementary Table 1). A subset of 30 SVs impacting coding exons that are correlated with 82 GWAS catalog markers are plausible causal variants for the associations. These variants include loci where the presence of an SV has been previously established using alternate methods, including a deletion in *LCE3B*[28] associating with psoriasis and a deletion in *CTRB2* associating with diabetes[29,30] and age related macular degeneration[31]. We also find loci where the co-occurrence of an SV and a GWAS locus has not been reported; including a 3,930bp deletion that overlaps the first exon of *SLC25A24* and correlates with a SNP associated with white blood cell count[32] and a 120 bp inframe deletion in *KAT2B*, that removes 40 amino acids from the translated protein and correlates with a variant associated with systolic blood pressure[30,33].

Targeted approaches have identified several associations between SVs and phenotypes. We recapitulate a number of these in a genomewide non-targeted fashion. The first example is a rare 2,476 bp deletion which deletes two exons of *COLA3* and we have shown to associate with hematuria[34]. A second example is an 84bp insertion in the *PRDM9* gene, which results in an

addition of a zink-finger motif to the encoded protein, PRDM9. The insertion results in a change in the DNA binding motif of PRDM9 and consequently changes the locations of meiotic recombinations[35,36]. Our third example is a 57 kb deletion, overlapping the genes *CTNS* and *SHPK*, originally associated with cystinosis[37] where homozygous carriers of the deletion generally develop cystinosis, a lysosomal storage disease characterized by the abnormal accumulation of the amino acid cystine. We identified a single homozygous carrier of this deletion in our imputation set, not included in our genotyping set, who had been diagnosed with cystinosis.

We associated the high-confidence SVs with phenotypes available at deCODE genetics (Methods). We found an association with a rare 14,154 bp deletion overlapping the first exon of *PCSK9* (Fig. 4) and LDL cholesterol levels (adjusted effect = −1.36 SD and $p = 2.4 \cdot 10^{-22}$). LDL cholesterol levels were 0.93 mmol/L lower in carriers (n = 75) than in non-carriers (n = 98,081). We observed a single carrier of the deletion in our LRS dataset, 40 in our SRS dataset and 123 heterozygous carriers in our imputation data, corresponding to an allele frequency of 0.041%. No homozygous carrier was identified. *PCSK9* encodes the enzyme proprotein convertase sub tilisin/kexin type 9, a key regulator of LDL cholesterol metabolism[38] and a target of cholesterol lowering drugs[39]. Loss-of-function variants in *PCSK9* are known to result in lower levels of LDL cholesterol and reduced cardiovascular risk[40–42], consistent with the association observed here.

In this study, we demonstrate the first application of ONT sequencing at population scale, and describe how it can be used to identify SVs to assess their impact on disease and other phenotypes. LRS technology and its data analysis methods are still maturing and can be improved upon. Nevertheless, we were able to use LRS data to identify 23,111 SVs per individual, most of the SVs estimated to be present and three times more than reported to be found with SRS data.

SVs alter on average more basepairs than SNPs and indels and the relative impact of these variations is an important avenue of research. Although we have highlighted SVs that overlap coding exons due their established functional impact, other SVs may still affect the individual, e.g. by removing regulatory regions or changing RNA secondary structure. A better understanding of the biochemical mechanisms that lead to and are affected by SVs will be essential to understand human evolution and disease. These will in turn also lead to better 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 show that our method can identify SVs in a genomewide non-targeted fashion. We show that SVs impacting protein function are disproportionately rare; as a result, we believe that large scale SV studies will be essential to understand their role in the genetics of disease. Having generated and analyzed LRS data for 1,817 Icelanders, we believe that this work sets a foundation for further large-scale studies of SVs, allowing investigation of their full frequency spectrum.

Methods

Participants

A set of 1,817 individuals was selected for ONT sequencing, including 369 trios (an offspring and both parents). The individuals were selected from a large set of Icelandic samples collected as part of disease association efforts at deCODE genetics. 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. The study population has been described in detail previously[16,17,44,45]. 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 (approvals no. VSN-15-023 and VSN 05-097-v6, with amendments).

DNA source

Most of the DNA samples sequenced in this study were isolated from whole blood (n = 1698). DNA from whole blood was extracted using the Chemagic method (Perkin Elmer), an automated procedure which involves the use of M-PVA magnetic beads (URLs). The remaining DNA samples (n = 119) were isolated from heart tissue. Samples were received and subsequently stored in liquid nitrogen. Samples were cut to smaller size on dry-ice if needed. Lysis buffer and a sterile 5mm steel bead were added to each sample prior to homogenisation on a TissueLyser LT (Qiagen). DNA was extracted from the homogenised lysates using the MasterPure DNA Purification kit (epicentre) following the manufacturers protocol, but with overnight Proteinase K digestion. Isolated DNA samples were quantified using a Trinean DropSense^{TM} and integrity asessed 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 − 5µg of DNA, depending on the exact version of the prep kit and flowcell type used for sequencing (GridION/PromethION). Approximately two thirds of the sample cohort underwent partial DNA shearing using the Covaris g-TUBE™ to a mean fragment size of 10 − 15 kbp. The remainder of the samples were not sheared (cf. Fig. S5 from March 2019 and onward). Library preparation started with DNA repair/A-tailing using the NEBNext FFPE repair mix (#M6630) and the NEBNext End repair/dA-tailing module (E7546), followed by AMPure XP bead clean-up. Adapter ligation was performed using the NEB T4 ligase (NEBNext quick Ligation Module, #E6056) and the ONT/LSK109 adapter mix (AMX) and ligation buffer, respectively. Samples were again purified using AMPure XP beads, using the Long Fragment Buffer (LFB) for the wash steps. Final sample elutions from the beads were done using 15µL of elution buffer (EB). Samples were quantified using a Qubit fluorimeter and diluted appropriately for loading onto the flowcells.

Sequencing

Samples were loaded onto either GridION R9.4 or PromethION R9.4.1 flowcells following ONT standard operating procedures. Sequencing was performed on either GridION X or PromethION devices, respectively. Data acquisition varied from 48 − 60 hours.

Basecalling

The squiggle data from the sequencers was basecalled using Albacore (995 individuals, 1, 290 flowcells) and Guppy (905 individuals, 1, 069 flowcells). We used Albacore Sequencing Pipeline Software, version 2.3.1, with model configuration file template `template_r9.4.1_450bps_5mer_raw_prom.jsn` for PromethION flowcells and configuration file `template_r9.4_450bps_5mer_raw.jsn` for GridION flowcells. We ran Guppy Sequencing Pipeline Software for GPU machines, version 2.1.3, for PromethION flowcells using the corresponding configuration file

`template_r9.4.1._450bps_5mer_raw_prom.jsn`.

An initial set of 822 individuals basecalled with Albacore that had reached a reference genome aligned sequencing coverage of 8x at the time of analysis (end of January 2019) were used in SV discovery.

Using an initial set of ten R9.4.1 flowcells we observed that the Albacore model configuration is deletion biased (Fig. 1D). We observed a mean deletion error 7.7%, CI = [5.7, 11.4], whereas the mean insertion error was 3.9%, CI = [2.3, 7.1]. The deletion error rate was thus 4.19%, CI = [1.3, 8.5] higher than insertion error. In order to limit the effects of potential false negative SVs due to deletion bias, we altered the stay penalty of the basecaller from a deletion biased to an insertion biased model. We observed that changing the stay penalty changed the relative error rate of deletion and insertions but not the total error rate. We used `--stay_penalty = 1.25` and `--stay_penalty = 1.00` for Albacore and Guppy, respectively, for our deletion biased (db) model and `--stay_penalty = 0.05` for our insertion biased (ib) model.

Read mapping

The basecalled reads were mapped to human reference genome GRCh38[43] with minimap2[18] (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[46] and stored in a bam file.

Sequencing statistics

To estimate sequencing error we used the following terminology:

M_a = Number of correct bases in matched basepairs

M_i = Number of mismatching (incorrect)bases in matched basepairs

I = Number of inserted basepairs

D = Number of deleted basepairs

C = Number of soft clipped basepairs present in read

Then we have the quantities

$M = M_a + M_i$	Number of matched basepairs
$S = M + I + C$	Sequence length
$L_{seq} = M + I$	Aligned read length w.r.t. sequence
$L_{ref} = M + D$	Aligned read length w.r.t. refererence
$L_{max} = max(L_{seq}, L_{ref})$	

Note, all secondary alignments are ignored. Moreover, when there were supplementary alignments available for a read, we took the first alignment present and sum the quantities M_i, I, D, L_{seq}, L_{ref} and L_{max} as long as the subsequent alignments do not overlap the last accepted alignment for the combined read.

These quantities were calculated for all reads for a basecalled flowcell, except for reads that are less than 3000 bp. Shorter reads often have an ambiguous mapping to the reference genome and are therefore of less use for SV calling. We did not omit reads labelled as "FAIL", omitting those reads from analysis would result in lower error rates, higher mapping rates but lower sequencing coverage.

Let r be a read and R be a set of reads. We report the error rate, E, as the total sum of erroneous basepairs, normalized by alignment length,

$$E_I = \frac{\sum_{r \in R} I(r)}{\sum_{r \in R} L_{max}(r)} \cdot 100\% \qquad \text{insertion error (1)}$$

$$E_D = \frac{\sum_{r \in R} D(r)}{\sum_{r \in R} L_{max}(r)} \cdot 100\% \qquad \text{deletion error (2)}$$

$$E_{M_i} = \frac{\sum_{r \in R} M_i(r)}{\sum_{r \in R} L_{max}(r)} \cdot 100\% \qquad \text{mismatch error (3)}$$

$$E_{total} = E_I + E_D + E_{M_i} \quad \text{total error} \quad (4)$$

and the alignment accuracy, A, as alignment length normalized by its read length,

$$A = \frac{\sum_{r \in R} L_{seq}(r)}{\sum_{r \in R} S(r)} \cdot 100\% \quad (5)$$

Mean error rates and alignment accuracy for both basecallers, using either deletion or insertion bias, is given in Table S2. On average 1.67% lower overall error rate was observed for flowcells basecalled with Guppy than Albacore using our deletion biased model.

In addition we report aligned coverage as:

$$X_i = \frac{\sum_{R_j \in \{R_i\}} \sum_{r \in R_j} L_{seq}(r)}{3 \cdot 10^9} \cdot 100\% \quad (6)$$

For all reads in a set of flowcells R_i belonging to the same individual i.

Preliminary structural variant predicting with Sniffles

A set of preliminary variant predictions was obtained using Sniffles[6] for each genome, in a highly sensitive fashion (using `-s 3`, and `--ignore_sd`) to minimize false negatives due to the existence of low coverage regions. Up to 30 supporting reads were reported per variant. Other optional parameters were left as default. Insertions and deletions with different start/end chromosomes and larger than 1 Mb are discarded.

Next, deletions and insertions with alternate allele ratio below 0.2 and 0.05, respectively, are discarded if detected using a deletion biased basecaller. Similarly, deletions and insertions with an alternate allele ratio below 0.05 and 0.2, respectively, are discarded if detected using an insertion biased basecaller, as a pre-filter from raw Sniffles calls.

Breakpoint and variant refinement with SViper

The variant predictions were either breakpoint refined or not with SViper[19] (URLs) using SRS data. All optional parameters were left as default. SViper first identifies the ONT long-reads supporting the candidate variant and forms consensus sequences flanking the candidate breakpoints. It then selects SRS reads near the predicted SV breakpoint(s), if available, and

aligns to the ONT consensus sequence for polishing. The polished consensus is re-aligned to the human reference to provide the refined breakpoints.

SV filtering using signal-level raw Nanopore data (SquiggleSVFilter)

We developed SquiggleSVFilter to filter false SV predictions using the signal-level raw ONT sequencing data, i.e. the squiggle. SquiggleSVFilter employs the squiggle-vs-sequence log likelihood score function provided by Nanopolish[20], and compares the log likelihood scores of the predicted alternate allele vs. 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[20]. 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 (i.e. a reference alignment 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 (URLs), and mapped these reads to the reference genome, using minimap2 with parameters as described in "Read mapping". Any read with a basecall score less than 0.1 or a length less than 1000 bp is discarded.

For any SV supporting read, reported by Sniffles, we start by calculating where the predicted SV breakpoints correspond in the read using the reference alignment bam file (Fig. S6). Next, we determine the read regions spanning the SV breakpoints, which we refer to as "subreads". An alignment may not contain an anchor on the reference on both sides of a breakpoint, and may instead be soft-clipped on one of the sides. We therefore determine a left and right subread by approaching from the left and right flanks of the variant. Figure S6A and S6B, depicts a sample deletion and insertion, respectively. We find a left subread using an alignment anchoring the reference from the left side of the breakpoint B (left breakpoint B1 for

a deletion). Using a flank size of 500 bp in the reference, we compute the read index L_b mapping to reference coordinate B − 500 (B1 − 500 for a deletion), as the begin index of the left subread, and set L_e = min(z − 1, L_b +1000), as the end index, where z is the length of the read sequence. Similarly, we find a right subread using an alignment anchoring the reference from the right side of the breakpoint B (left break-point B2 for a deletion). We compute the read index, R_e, mapping to reference coordinate B + 500 (B2 + 500 for a deletion), as the end index of the right subread, and set R_b = max(0, R_e − 1000), as the begin index.

Using the event table, we find the event indices corresponding to the subreads to access event slices spanning the SV breakpoints. Given the predicted SV breakpoint sites and sequence (for insertions), and the reference alignment bam file, we determine the reference (ref) and alternate (alt) allele sequences spanning the left and right subreads. For a deletion, we set the ref allele sequences as 500 bps of reference sequence flanking B1, and B2 from both sides, for the left and right event slices, respectively. We set the alt allele sequence as 500 bps flanking B1 from the left, followed by 500 bps flanking B2 from the right, for both event slices, as shown in Fig. S6A. Similarly, for an insertion, we set the ref allele sequences as 500 bps of reference sequence flanking B from both sides, for both event slices. We set the alt allele sequence as 500 bps flanking B from the left, appended by min(500, m) bp from the insertion sequence of the predicted SV, where m is the insertion sequence length, followed by max(0, 500 − m) bp flanking B from the right, for the left event slice, and similarly for the right event slice, as shown in Fig. S6B. Finally, we calculate the raw signal-vs-sequence log likelihood scores using the ref and alt allele sequences for both event slices, and use their difference to support or reject the candidate variant. We support a variant if at least 3 reads obtain a log likelihood score difference of at least 50 for either of the event slices.

Figure S7 displays the SV counts per individual and points to the levels of SV filtering in successive stages. The alternate allele ratio based pre-filter is followed by SquiggleSVFilter,

where we observe a lower variance on the number of SVs across individuals.

We estimated false-positive levels for SV discovery per individual by checking whether one of the parents also contain the SVs discovered (see Parent support of proband SVs in trios) in 96 trios. We observed consistent low false-positive levels (Fig. S8), with the exception of a number of individuals where error rates were exceptionally high.

We observed SquiggleSVFilter to be especially effective in individuals with high number of SVs after the alternate allele ratio based pre-filter, a result of high error rates in a fraction of the LRS data. Figure S7A shows this uneven level of filtering achieved by SquiggleSVFilter, across individuals, which is a result of the high mean and variance observed in error rates (Fig. 1B-C).

Constructing a single basecaller SV set

Using a single basecaller, we run SquiggleSVFilter on both the original (Sniffles) variants and their breakpoint refined (SViper) forms. If both forms of the variant are acceptable, we accept the breakpoint refined form if it has a mean log likelihood difference score greater than 2/3 times the mean log likelihood difference score of the original variant. Mean log likelihood difference scores are calculated using only values ≥ 50.

We test both the breakpoint refined and unrefined forms of the variant with SquiggleSVFilter in case the short-read breakpoint refinement caused a distortion of the otherwise correct original form of the variant, resulting in its filtering. We however bias our selection towards the breakpoint refined form in order to be more exact at the breakpoint sites, for genotyping purposes

Constructing an individual's SV set

Per individual, we construct db and ib SV sets separately, and pool all the variants. Next, in any sequence of sorted SVs with a consecutive begin site difference of 250 bp, we pick one representative SV. We give precedence first whether it is a breakpoint refined variant, and

second whether it is a db SV. See Fig. S9 for the effect of using both db and ib basecalling configurations, and their independent contributions.

Parent support of proband SVs in trios

We estimated SV discovery precision upper bounds by computing the rate of parent-supported SVs on a proband, on available trios (Fig. S8), assuming no de novo events. We accept any proband SV discovered at a site close to any SV found in any parent, as parent-supported. An insertion is parent-supported if a parent insertion begin site is within 250 bp to the proband insertion begin site, and similarly, a deletion is parent-supported if a parent deletion begin site is within 250 bp to the proband deletion begin or end site.

Although we ran our complete SV discovery pipeline on the proband, in order to achieve maximal sensitivity on the parent SVs, we ran Sniffles with 2 reads of support (`-s 2`) and ignoring the sd filter (`--ignore_sd`) on parent genomes. Other parameters were left as default. For simplicity, we used a single basecaller for both parents and the proband in calculating the SV parent-support rates.

Merging of the SVs

Most SVs are carried by multiple individuals, and thus will be re-discovered, potentially with slightly different representations across carriers, varying in length and location. In order to eliminate such redundancies, we applied the following SV merging approach, independently for insertions and deletions. We represented SVs as nodes in a graph, and drew an edge between two SVs if they had a minimum mutual overlap of at least 50%, with 1 − (minimum mutual overlap) as edge length. We measured the overlap between two insertions by representing them similar to deletions, where an insertion end coordinate is set as its begin coordinate incremented by its length. 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 (CAST) algorithm[47]. We segregated all discovered SVs into non-overlapping groups, and applied the CAST algorithm on them to find the SV cliques. We picked the most frequent SV, determined by site and length, as clique representative, i.e. an SV centroid. We broke ties by prioritizing the breakpoint refined and deletion biased SVs, in the given order, similar to how we construct an individual's SV set. For insertions, the insertion sequence is arbitrarily picked among the individuals containing the clique representative SV. In a list of successive SV centroids with begin coordinates within 250 bp of each other, we picked the one that represents the SV clique with maximum number of individuals contributing to it, as long as it is ≥ 2. Any SV centroid isolated by at least 250 bp is picked if it represents a clique with at least 2 individuals or a single individual with a minimum alternate allele ratio of 0.3. The alternate allele ratio is measured as the number of SV supporting reads divided by site coverage. The coverage at the site for insertions is determined using the coordinate of the insertion position, and for deletions as the median coverage of equidistant 10 coordinates within the deletion interval.

Construction of the merged SV set VCF file used in SRS and LRS genotyping

Since the SRS genotyping method we used (Graphtyper) is sensitive to unreported indels near SV breakpoints, we provided a VCF file for genotyping purposes where we included all flanking variants, if possible. Any variation from the reference sequence, surrounding the refined SV breakpoints, on the polished ONT consensus re-alignment to the human reference for an SV processed by SViper, is reported in the respective ALT and REF fields on the VCF file by processing the CIGAR string. As a result, the sequences given in the VCF may contain other variations near the discovered SV, such as target site duplications or imperfections in the SV polishing. We note that this approach may change the position and the size of the SVs, thus leading to conditions enforced during merging of the discovered SVs invalid. There was,

however, no duplicate VCF entries given SV coordinates and sequences involved.

Tandem repeat identification

We searched for tandem repeats using Tandem Repeats Finder[48] in both the reference and alternate allele sequence of all variants. Variants were considered to be tandem repeats if a tandem repeat of length at least 50% of the length of the allele was found in either allele sequence.

Comparison of discovered structural variants to available datasets

We used the same comparison approach described in section "Parent support of proband SVs in trios". We show results for various cutoffs as breakpoint distance.

Individual selection for short read genotyping

We selected 5,000 individuals for short read genotyping from our set of Illumina WGS individuals generated previously at deCODE genetics[17]. This set included 1,656 of the 1,817 ONT sequenced individuals, including 736 of the 822 individuals used in the SV discovery set. To increase the probability of finding multiple carriers of rare variants, the set of 5,000 individuals included 1,533 trios (offspring and the two parents).

SRS genotyping with GraphTyper

We provided the merged SV set to GraphTyper[5,21], which generates an augmented graph genome using the SV predictions, together with previously discovered SNPs and indels[5], for population scale genotyping. The variants were genotyped on the set of 5,000 individuals, using three genotyping models. All three models were used in our imputations (see Imputation of SVs).

LRS genotyping (LRcaller)

LRcaller is a proof-of-concept genotyping algorithm that genotypes SVs directly from ONT sequencing reads. Each breakpoint is genotyped independently, resulting in two genotypings

for the canonical deletion and insertion variants identified in this study, corresponding to the left and right breakpoints (Fig. S6). Note that the algorithm processes each variant independently, i.e. each variant is genotyped without considering other variants in the region, which may lead to suboptimal behavior when there are multiple variants present in the same region.

We use 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) from the alignment present in the bam file as aligned by minimap2.

AD genotyping: We start by constructing a sequence for the reference and the alternate allele from a VCF record, in a method analogous to the one described by SquiggleSVFilter. A sequence for the reference allele is constructed as the sequence in reference coordinates (b − 500, b + 500), where b is the position of the breakpoint. A sequence for the alternate allele is constructed analogously, except the sequence in the ALT field of the VCF record is first inserted into the reference and the sequence of the REF field is removed, creating a new reference RA. The alternate sequence constructed is the sequence in coordinates (b − 500, b + 500) in RA.

We next select and crop reads used for genotyping. For left breakpoints we select reads that map to a position in the interval (b − 500, b). We let c_l be the smallest position in the interval (b − 500, b) where the read is aligned, let $i(c_l)$ be the read index aligned to reference position c_l and we crop subsequence $(\max(0, i(c_l) - d_l), \min(i(c_l) - d_l + 1000, \text{readlength}))$ from the read, where $d_l = c_l - b + 500$. For right breakpoints we select reads that map to a position in the interval (b, b + 500). We let c_r be the largest position in the interval (b, b + 500) where the read is aligned and we crop subsequence $(\max(0, i(c_r) + d_r - 1000), \min(i(c_r) + d_r, \text{readlength}))$ from the read, where $d_r = b + 500 - c_r$. Cropped sequences shorter than 500 bp are

ignored.

We then align cropped reads using the function globalAlignment in Seqan[49] with scores 1 for match and -1 for mismatch, insertions and deletions to the reference and the alternate allele, producing alignment scores AS(REF) and AS(ALT), respectively. Alignment scores < 400 are artificially set to 400 and reads where AS(REF) = AS(ALT) are not used. For all other reads, r, a score ad(r) computed as ad(r) = (AS(ALT) − AS(REF))/2. We then cap ad(r) to have absolute value at most 10. ad(r) = min(max(ad(r), − 10), 10).

VA genotyping: We consider the minimap2 alignment to the reference in the interval (b_l −100, b_r + 100), b_l and b_r are the left and right breakpoints, respectively. When considering the left breakpoint we only consider reads overlapping reference coordinate b_l − 100 and are not soft clipped (> 500 bp) at the start of the alignment. Reads that are soft-clipped (> 500 bp) at the end of their alignment within the window (b_l − 100, b_r + 100) are considered to support the alternate variant. For other reads we count, S, the number of inserted or deleted basepairs, in stretches of at least 5bp, in the window (b_l − 100, b_r + 100). Reads where $S > \frac{1}{2} \cdot SV\ length$ are considered to support the deletion and those where $S < \frac{1}{10} \cdot SV\ length$ are considered to support the reference. A read, r, is assigned a score va(r) of −10 or 10 when it supports the alternate or reference, respectively. Reads that support neither the alternate nor the reference are not used in the computation. An analogous procedure is performed for right variants.

Joint model genotyping: The output of the two models are then used in a joint model as: joint(r) = ad(r) + va(r).

Relative log likelihoods: These scores are then used to compute l(r|g), the relative log likelihood of observing a read r given each of the three possible genotypes, g, of an individual; 00, 01, 11, representing homozygous reference, heterozygous and homozygous alternate genotypes, respectively. The scores are then interpreted to represent log2 of l(r|11) − l(r|00),

where a score of x means observing the read is 2x more likely if the individual is homozygous alternate than homozygous reference. We then let $l(g|01) = \max(l(g|00), l(g|11)) - 1$, i.e. observing a read given the individual is heterozygous is half as likely as observing it if the individual is homozygous.

A joint relative likelihood of observing all the reads given the three possible genotypes is then found by multiplying these relative likelihoods, or summing the log relative likelihoods.

Genotyping with LRcaller

The variants were genotyped independently for the left and right breakpoint using the three different models presented above for each variant, producing a total of six genotypes per individual/marker pair.

Imputation of SVs

For each marker a total of 9 different genotypes were produced, 3 from GraphTyper/Illumina and 6 from LRcaller/ONT. All genotyped variants are imputed into the haplotypes of 166,281 Icelanders, using a methodology previously described[17,22,23]. For each variant the genotyping that produced the highest imputation info[23] was selected as the best genotype.

Variants with an imputation info of 0.9 or higher were considered successfully imputed, i.e. high-confidence, and used for genome-wide association analysis.

Calculating autosomal SV counts per individual

We reported the autosomal SV counts per individual by selecting the best imputed genotyping per SV. If an SV had no genotype with an imputation information > 0.9 and at least one genotype with imputed allele frequency of 1.0, the LRS AD (right breakpoint) genotyping model is selected as a default model since it has the highest number of imputed SVs, compared to other models. If the selected genotyping model is from SRS genotyping and with an imputation information ≤ 0.9, we accepted the variant if it had a `PASS` flag, and `PASS_AC` value greater than 0.

Parent-offspring transmission rates of selected SV genotypes, prior to imputation, are given in (Table S1).

Association testing

We tested for association with LDL levels based on the linear mixed model implemented in BOLT-LMM[50]. We used BOLT-LMM to calculate leave-one-chromosome out (LOCO) 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[51]. We used linkage disequilibrium (LD) score regression to account for distribution inflation in the dataset due to cryptic relatedness and population stratification[52]. 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 $\chi 2$ statistics from a genome-wide association scan against LD score and used the intercept as correction factor. Effect sizes based on the LOCO residuals are shrunk and we rescaled them based on the shrinkage of the 1.1 million variants used in the LD score regression.

The SV we report on *PCSK9* (Fig. 4) in association with LDL cholesterol levels was discovered on an earlier SV discovery set available at the time, consisting of 46 LRS individuals. The individual the deletion is discovered in is not among the individuals used in the construction of the merged SV set.

Comparison to GWAS catalog

We downloaded version 1.0.2 of the GWAS catalog with all associations (URLs) on July 12th 2019 (`gwas_catalog_v1.0.2-associations_e96_r2019-07-12.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 inhouse 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 500kb, such that a correlation between a SNP/indel and a SV is observed if and only if they are within 500 kb of each other.

Computation of parent offspring transmissions

All trios included in the study were used to compute parent offspring transmission rates. The genotypes of the child were computed conditioned on the genotypes of the parents. Computations were restricted to SV/trio pairs where a confident call could be made for all members of the trio by requiring the phred scaled genotype likelihood of the second most likely genotype to ≥ 40 for all members of the trio.

URLs

SViper, original repository: `https://github.com/smehringer/SViper`

SViper, modified, used in this study:

`https://github.com/DecodeGenetics/SViper/tree/cornercases`

Scrappie, original repository: `https://github.com/nanoporetech/scrappie`

Scrappie, modified, used in this study:

`https://github.com/DecodeGenetics/scrappie/tree/v1.3.0.events`

SquiggleSVFilter:

`https://github.com/DecodeGenetics/nanopolish/tree/squigglesv`

GWAS catalog: `https://www.ebi.ac.uk/gwas/docs/file-downloads`

LRcaller: `https://github.com/DecodeGenetics/LRcaller`

Chemagen: `https://chemagen.com`

Acknowledgements: We thank Prof. Jared Simpson, our colleagues from deCODE genetics / Amgen Inc., and employees of Oxford Nanopore Technologies for their help and advice. We would also like to thank all research participants who provided a biological sample to deCODE genetics.

Author contributions: DB implemented software with additional software implemented by HI, HPE, SK, SM and BVH. DB and BVH wrote the paper with input from HI, HPE, EB, SK, HJ, MTH, RPK, GH, UT, DFG, PS, OTM, KS. HI implemented the analysis pipelines, with input from DB, SK, SAG, STS, GM and BVH. DNM and OTM performed the ONT sequencing. GIE, IO and OS acquired LDL measurements. BVH and KS conceived and supervised the study. All authors approved the final version of the manuscript.

Competing interests: DB, HI, HPE, EB, SK, SM, HJ, MTH, DNM, SAM, STS, GH, GM, UT, DFG, PS, OTM, BVH and KS are employees of deCODE genetics/Amgen.

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 (2013).

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

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

5. Eggertsson, H. P. *et al.* Graphtyper enables population-scale genotyping using pangenome graphs. *Nat. Genet.* **49**, 1654–1660 (2017).

6. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).

7. Kloosterman, W. P. *et al.* Characteristics of de novo structural changes in the human genome. *Genome Res.* **25**, 792–801 (2015).

8. Abel, H. J. *et al.* Mapping and characterization of structural variation in 17,795 deeply sequenced human genomes. *bioRxiv* 508515 (2018).

9. Collins, R. L. *et al.* An open resource of structural variation for medical and population genetics. *BioRxiv* 578674 (2019).

10. Audano, P. A. *et al.* Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* (2019).

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

12. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).

13. Cretu Stancu, M. *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.* (2019).

15. Gilpatrick, T. *et al.* Targeted Nanopore Sequencing with Cas9 for studies of methylation, structural variants and mutations. *BioRxiv* 604173 (2019).

16. Gudbjartsson, D. F. *et al.* Sequence variants from whole genome sequencing a large group of Icelanders. *Sci. data* **2**, 150011 (2015).

17. Jónsson, H. *et al.* Whole genome characterization of sequence diversity of 15,220

Icelanders. *Sci. data* **4**, 170115 (2017).

18. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

19. Mehringer, S. & others. SViper: a tool for SV polishing. *Prep.* (2019).

20. 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).

21. Eggertsson, H. & others. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* **To Appear**, (2019).

22. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40**, 1068 (2008).

23. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435 (2015).

24. Sun, J. X. *et al.* A direct characterization of human mutation based on microsatellites. *Nat. Genet.* **44**, 1161 (2012).

25. Maretty, L. *et al.* Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature* **548**, 87 (2017).

26. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733--D745 (2015).

27. Buniello, A. *et al.* The {NHGRI-EBI GWAS} Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005--D1012 (2018).

28. 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 (2009).

29. 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 (2015).

30. Kichaev, G. *et al.* Leveraging polygenic functional enrichment to improve GWAS power. *Am. J. Hum. Genet.* **104**, 65–75 (2019).

31. 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 (2016).

32. 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).

33. Evangelou, E. *et al.* Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nat. Genet.* **50**, 1412 (2018).

34. Benonisdottir, S. *et al.* Sequence variants associating with urinary biomarkers. *Hum. Mol. Genet.* **28**, 1199–1211 (2018).

35. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099 (2010).

36. Hinch, A. G. *et al.* The landscape of recombination in African Americans. *Nature* **476**, 170 (2011).

37. 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).

38. Horton, J. D., Cohen, J. C. & Hobbs, H. H. PCSK9: a convertase that coordinates LDL catabolism. *J. Lipid Res.* **50**, S172--S177 (2009).

39. 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 Inhibit. *Circulation* **126**, 2408–2417 (2012).

40. Cohen, J. C., Boerwinkle, E., Mosley Jr, T. H. & Hobbs, H. H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* **354**, 1264–1272 (2006).

41. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274 (2013).

42. 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).

43. 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).

44. Nioi, P. *et al.* Variant ASGR1 associated with a reduced risk of coronary artery disease. *N. Engl. J. Med.* **374**, 2131–2141 (2016).

45. 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 (2016).

46. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

47. Ben-Dor, A., Shamir, R. & Yakhini, Z. Clustering gene expression patterns. *J. Comput. Biol.* **6**, 281–297 (1999).

48. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).

49. Döring, A., Weese, D., Rausch, T. & Reinert, K. SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics* **9**, 11 (2008).

50. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284 (2015).

51. Benonisdottir, S. *et al.* Epigenetic and genetic components of height regulation. *Nat. Commun.* **7**, 13490 (2016).

52. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291 (2015).

Figures

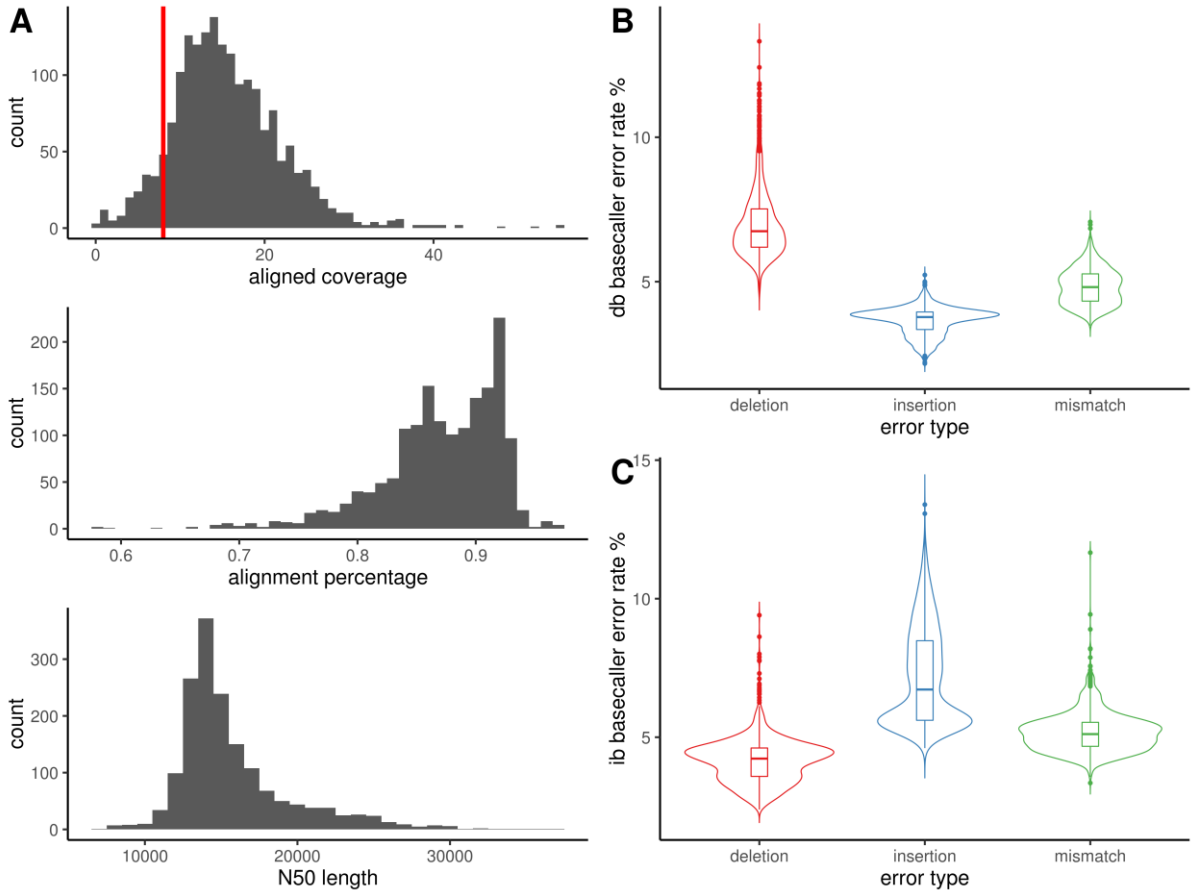

Fig. 1 Oxford Nanopore Technologies (ONT) long-read sequencing statistics. (A) Per individual distributions of aligned coverage (n = 1817, red vertical line denotes 8x coverage), aligned basepairs percentage, and N50 length (n = 1656, data from individuals with aligned coverage < 8x omitted). (B, C) Error rates compared to GRCh38[43] using deletion biased (db), and insertion biased (ib) basecallers, respectively, stratified by error type. Statistics are computed over all sequenced reads longer than 3000bp.

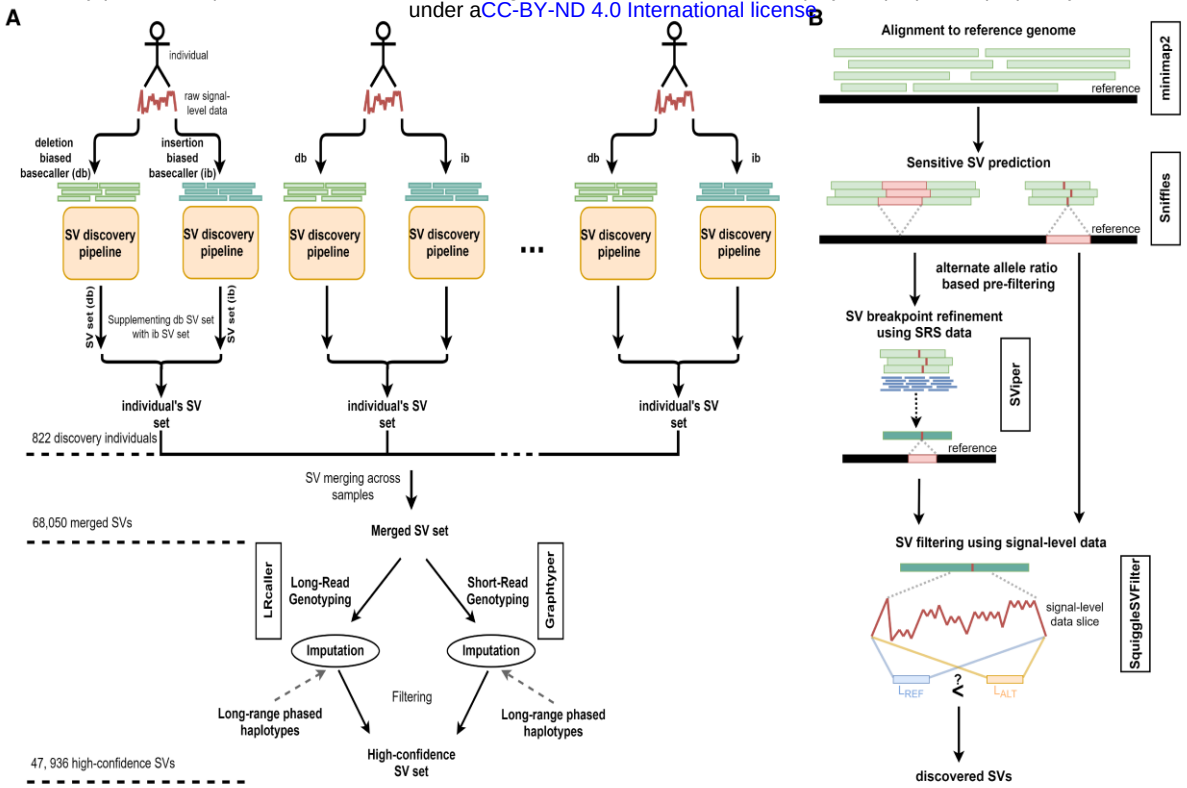

Fig. 2 Structural variant (SV) analysis workflow. (A) Each individual is basecalled using both deletion biased (db) and insertion biased (ib) basecallers, and SVs are identified independently. An individual's SV set is generated by supplementing the db SVs with the ib SV calls. SV sets are merged across all individuals, and the merged SV set is used to genotype individuals on both short read sequencing (SRS) and long read sequencing (LRS) data, separately. Finally, genotyped variants are imputed into long-range phased haplotypes and variants with sufficiently high imputation information (≥ 0.9) are accepted as high-confidence variants. (B) Reads are mapped to human reference genome (GRCh38) using minimap2, followed by the 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.

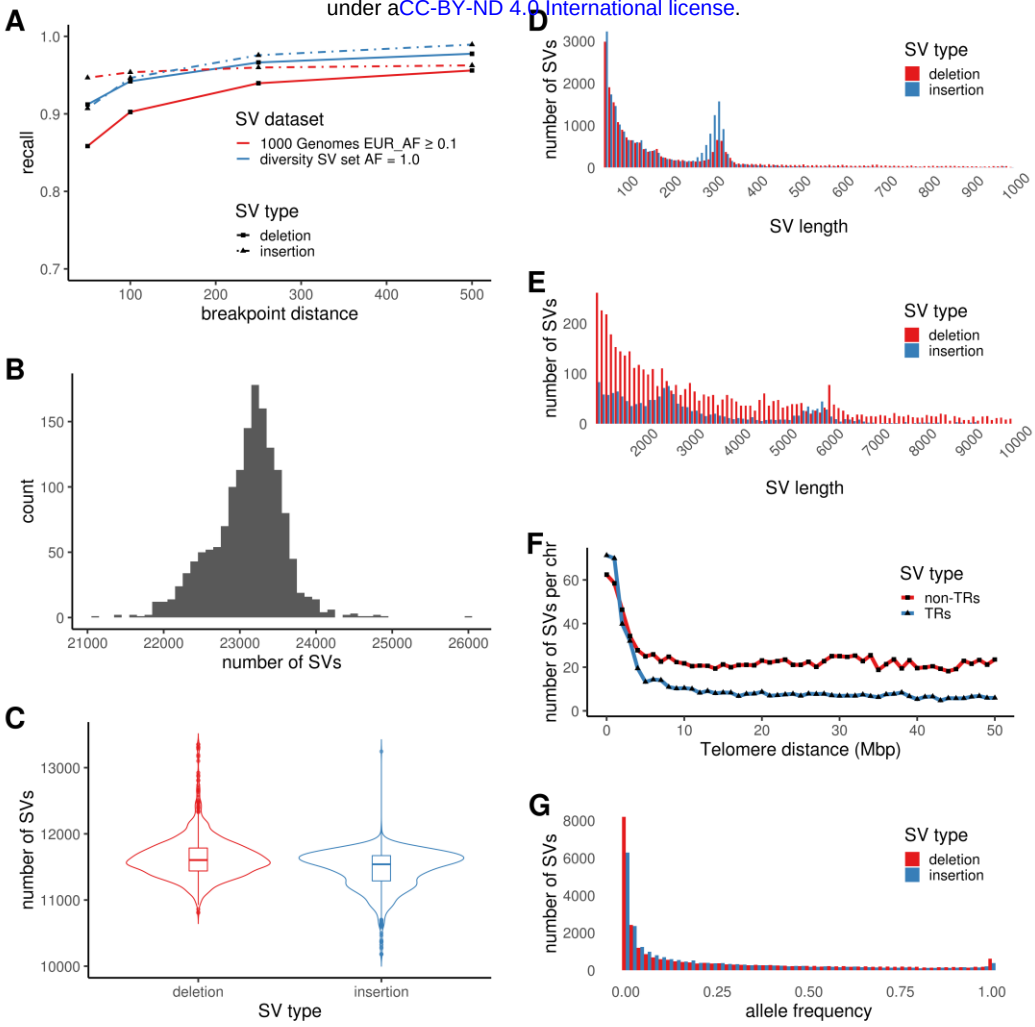

Fig. 3 Merged structural variant (SV) set characteristics. (A) Comparison of the merged SV set to 1000 Genomes SV dataset[3] using variants with European allele frequency (EUR_AF) ≥ 0.1, and the diversity SV set[10] using fixed variants (allele frequency (AF) of 1.0), at different breakpoint distances. (B,C) Distribution of total number of SVs (B) and number of deletions and insertions (C) per individual (n = 1507, individuals with aligned coverage < 8x and those not genotyped using short read sequencing (SRS) data are omitted). (D,E) SV length distributions in ranges [50 bp, 1 kbp], and [1 kbp, 10 kbp], respectively. The peaks observed around 300 bp and 6 kbp correspond to short and long interspersed nuclear elements (SINE and LINE), respectively. (F) Number of tandem repeats (TR) and non-TR SVs located at given telomere distances (binned at 1 Mb). (G) Allele frequency distribution of imputed high-confidence SVs.

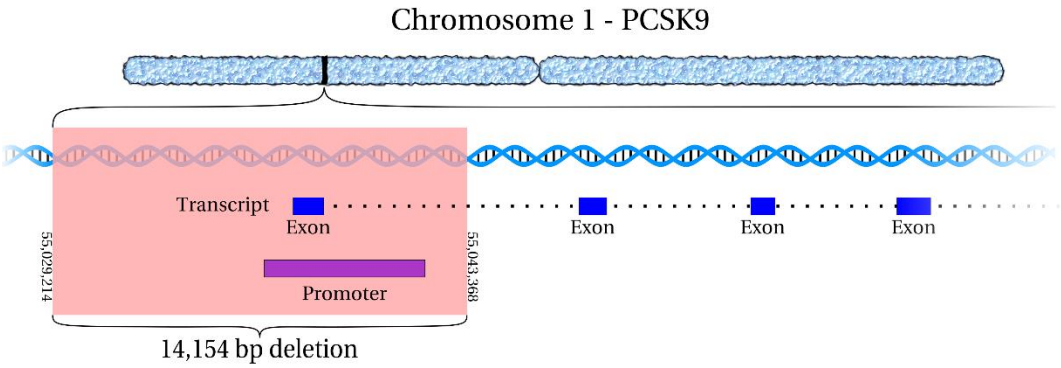

Fig. 4 *PCSK9* deletion. Rare (allele frequency 0.041%) 14,154 bp deletion on chromosome 1, at basepairs 55,029,214 to 55,043,368 (GRCh38) removes the promoter and first coding exon of *PCSK9*.

Supplementary Information

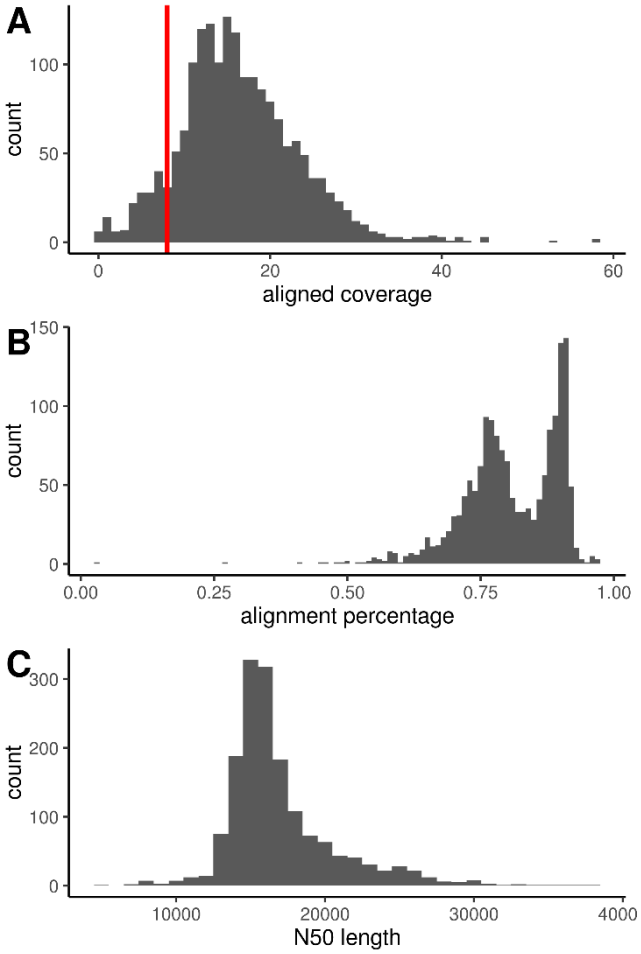

Fig. S1 Insertion biased (ib) basecalling sequencing statistics. Per individual distributions of aligned coverage (A) (n = 1817, red vertical line denotes 8x coverage), aligned basepairs percentage (B), and N50 length (C) (n = 1656, data from individuals with aligned coverage < 8x omitted)

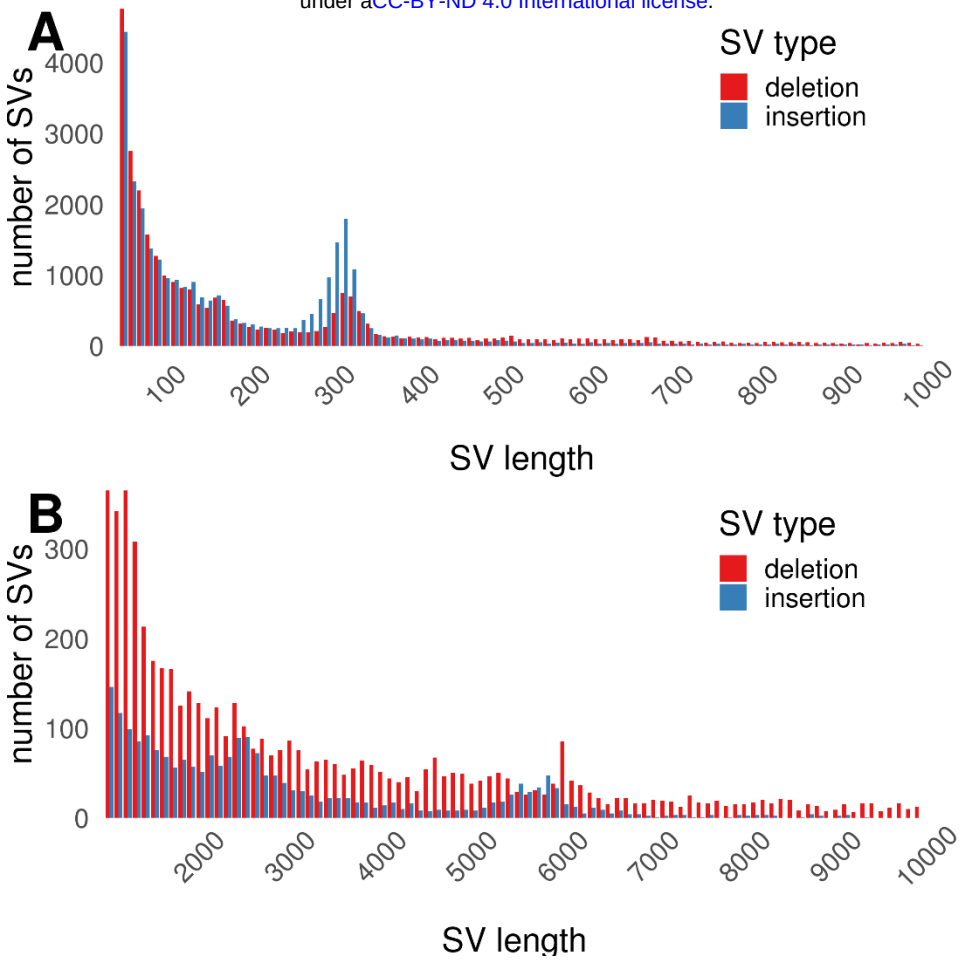

Fig. S2 Structural variant (SV) length distributions in ranges [50 bp, 1 kbp] (A), and [1 kbp, 10 kbp] (B), using all merged SVs.

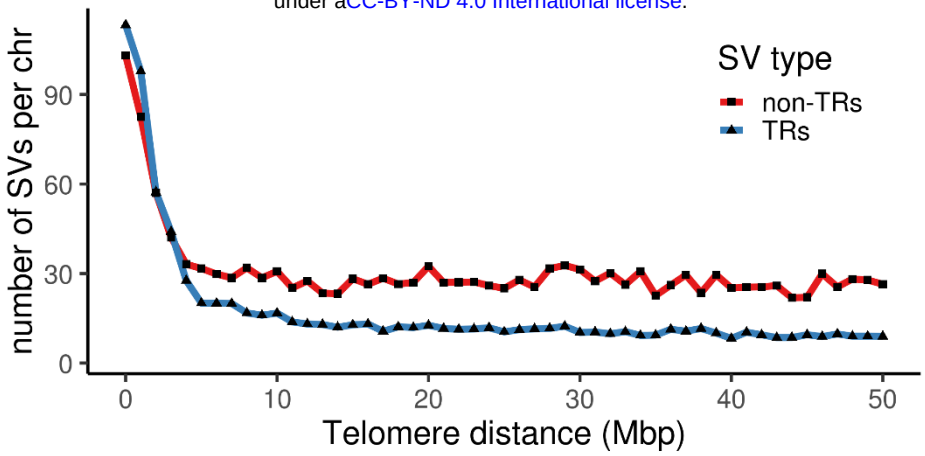

Fig. S3 Number of tandem repeat (TR) and non-TR SVs located at given telomere distances (binned at 1 Mb), using all merged structural variants (SVs).

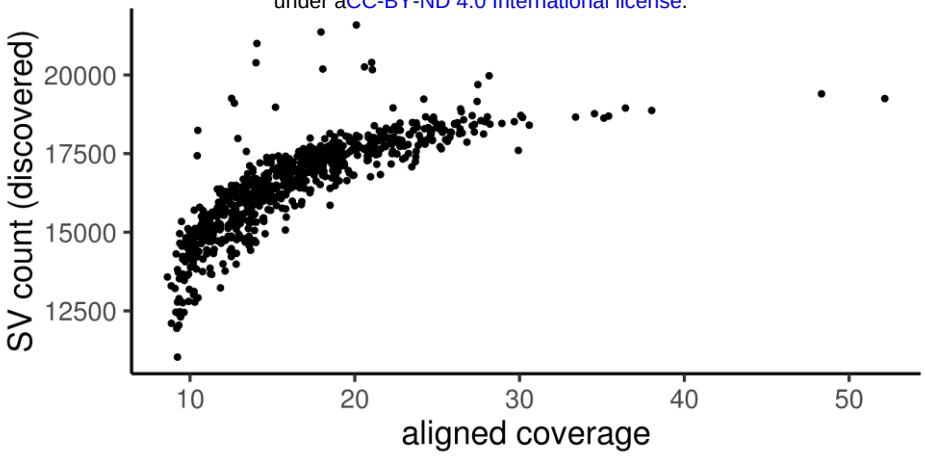

Fig. S4 Aligned coverage vs. number of discovered structural variants (SVs), per individual.

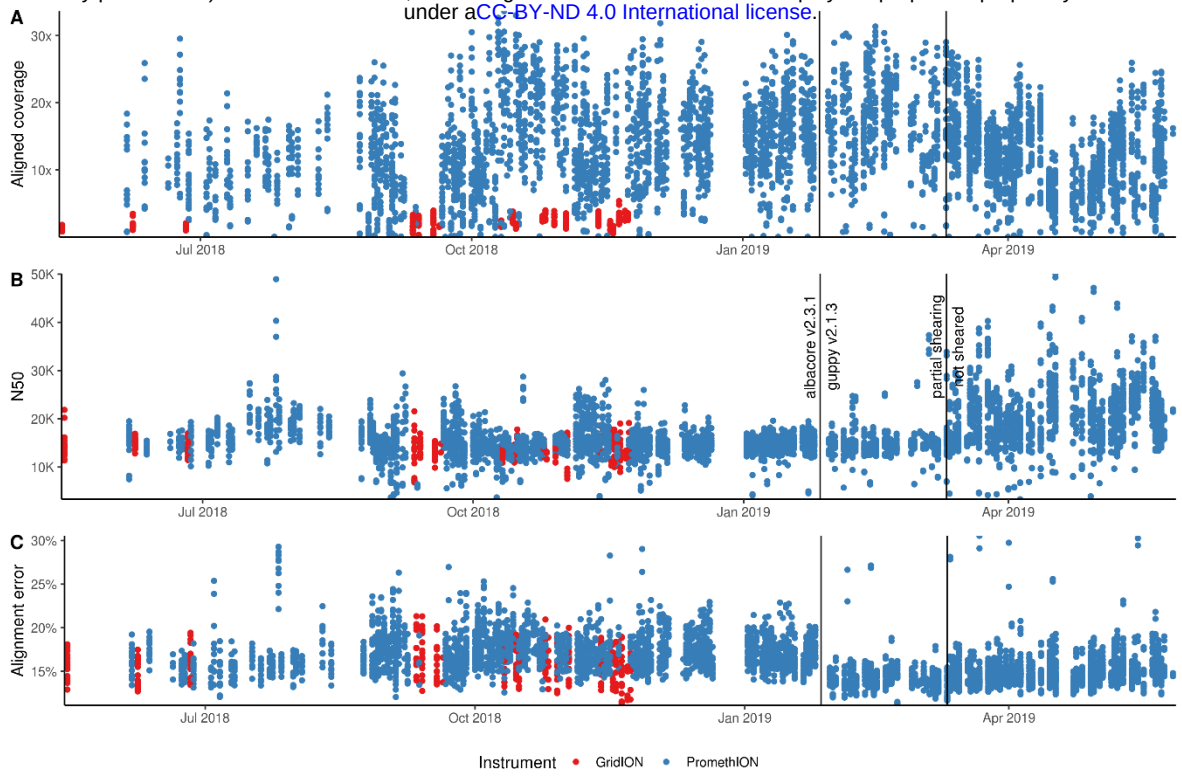

Fig. S5 A grand total of 2,357 flowcells were analysed for 1,817 individuals from May 2018 until June 2019. Thereof 127 were sequenced on GridION (red), and 2,232 on PromethION (blue) instruments. (A) A mean yield for a GridION flowcell is 2x coverage versus 11x for PromethION flowcells. (B, C) In February 2019 we upgraded our basecaller from Albacore to Guppy, resulting in a mean 1.6% decrease in alignment error. Moreover, in March 2019 we stopped partially shearing the DNA, resulting in approximately 2.6x lower coverage per flowcell. However N50 increased by a mean 5,500 basepairs.

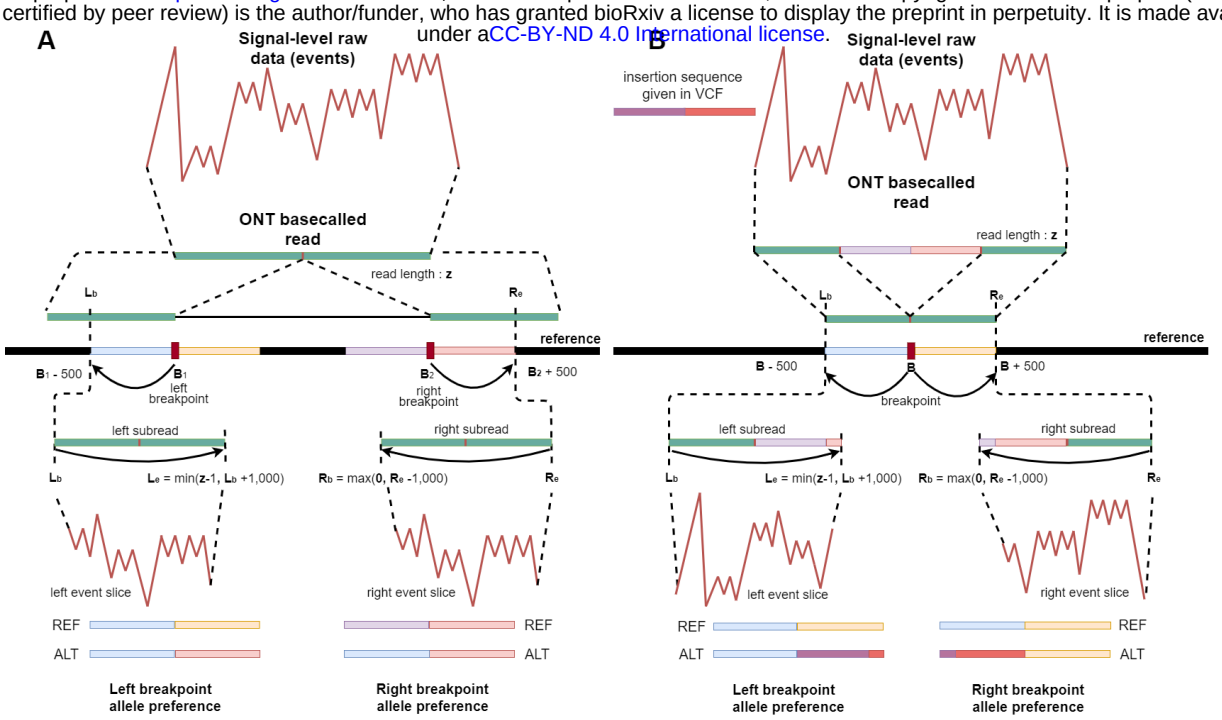

Fig. S6 SquiggleSVFilter overview. Given a candidate structural variant (SV), and an SV supporting read, SquiggleSVFilter first identifies the subread of the ONT basecalled read overlapping the SV, using the reference alignment BAM file. Next it finds the squiggle slice of the identified subsequence 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.

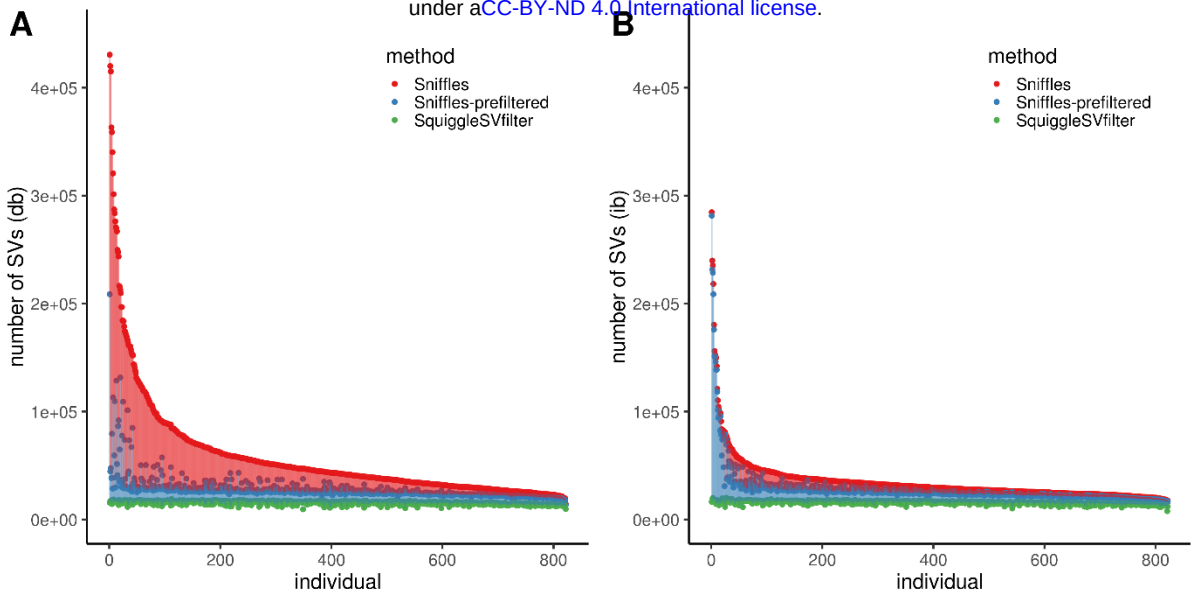

Fig. S7 Structural variant (SV) filtering. Per individual SV calls (A for deletion biased (db) and B for insertion biased (ib) SVs) are shown for Sniffles (red), a local alternative allele ratio based pre-filtering (blue), and SquiggleSVfilter (green), where each subsequent step uses the output of the previous step as input SV set.

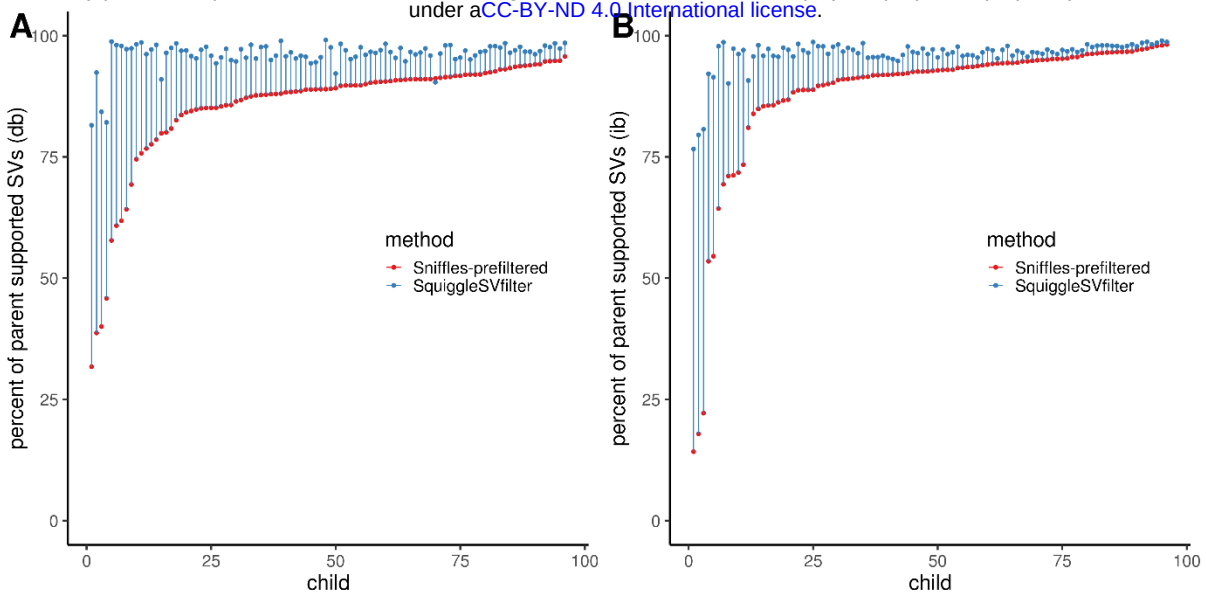

Fig. S8 Parent support rates on discovered structural variants (SVs) in 96 children (panel A for deletion biased (db), and B for insertion biased (ib) SVs). Red denotes the percent of parent supported SVs at 250bps breakpoint distance on the pre-filtered Sniffles variants, whereas blue denotes on SquiggleSVfilter filtered variants.

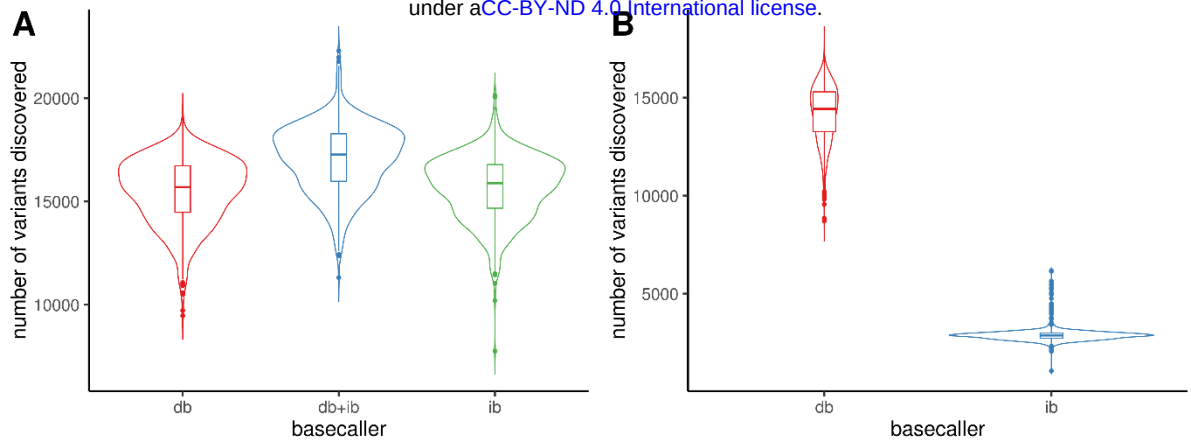

Fig. S9 The effect of supplementing a deletion biased (db) configuration of the basecaller with an insertion biased (ib) configuration. (A) Structural variant (SV) counts using the db and ib basecallers independently, and combined (db+ib), per individual. (B) Contributions from the db and ib SVs to combined used of basecallers (db+ib), per individual.

	0/0	**0/1**	**1/1**
0/0 + 0/0	99.51%	0.49%	0.00%
0/0 + 0/1	45.53%	54.28%	0.18%
0/0 + 1/1	1.11%	98.09%	0.80%
0/1 + 0/1	15.26%	71.82%	12.92%
0/1 + 1/1	0.38%	56.20%	43.42%
1/1 + 1/1	0.02%	1.39%	98.59%

Table S1: Parent-offspring structural variant (SV) transmission rates, in percentages. Rows and columns denote parent and offspring genotypes, respectively.

basecaller	**bias**	**#flowcells**	E_{total}	E_I	E_D	E_{M_i}	A
Albacore	db	1288	16.3%	3.47%	7.66%	5.20%	82.7%
Albacore	ib	1282	17.9%	8.47%	3.86%	5.52%	73.8%
Guppy	db	1069	14.7%	3.92%	6.31%	4.43%	90.7%
Guppy	ib	985	15.0%	5.57%	4.66%	4.77%	88.8%

Table S2: Alignment error estimates and alignment ratio defined by Eqs. (1)-(5) for both basecallers Albacore and Guppy, using either the deletion biased (db) or insertion biased (ib) basecalling configurations.