Science
AAAS

Supplementary Materials for

Haplotype-resolved diverse human genomes and integrated analysis of structural variation

Peter Ebert*, Peter A. Audano*, Qihui Zhu*, Bernardo Rodriguez-Martin*, David Porubsky, Marc Jan Bonder, Arvis Sulovari, Jana Ebler, Weichen Zhou, Rebecca Serra Mari, Feyza Yilmaz, Xuefang Zhao, PingHsun Hsieh, Joyce Lee, Sushant Kumar, Jiadong Lin, Tobias Rausch, Yu Chen, Jingwen Ren, Martin Santamarina, Wolfram Höps, Hufsah Ashraf, Nelson T. Chuang, Xiaofei Yang, Katherine M. Munson, Alexandra P. Lewis, Susan Fairley, Luke J. Tallon, Wayne E. Clarke, Anna O. Basile, Marta Byrska-Bishop, André Corvelo, Uday S. Evani, Tsung-Yu Lu, Mark J.P. Chaisson, Junjie Chen, Chong Li, Harrison Brand, Aaron M. Wenger, Maryam Ghareghani, William T. Harvey, Benjamin Raeder, Patrick Hasenfeld, Allison A. Regier, Haley J. Abel, Ira M. Hall, Paul Flicek, Oliver Stegle, Mark B. Gerstein, Jose M.C. Tubio, Zepeng Mu, Yang I. Li, Xinghua Shi, Alex R. Hastie, Kai Ye, Zechen Chong, Ashley D. Sanders, Michael C. Zody, Michael E. Talkowski, Ryan E. Mills, Scott E. Devine, Charles Lee, Jan O. Korbel, Tobias Marschall, Evan E. Eichler

Correspondence to:

eee@gs.washington.edu

tobias.marschall@hhu.de

jan.korbel@embl.org

charles.lee@jax.org

This PDF file includes:

Materials and Methods

Figs. S1 to S103

Other Supplementary Materials for this manuscript include the following:

Fig. S52 (full-size PDF version)

Tables S1 to S56

Material and Methods

Table of Contents

DATA PRODUCTION

1 Samples and data availability

(*Contributors: Katy Munson, Qihui Zhu, Scott Devine, Susan Fairley*)

Samples were selected from the 1000 Genomes Project (1000GP) Diversity Panel (*15*) and at least one representative was selected from each of 26 populations (Table S1). Cell lines were obtained from Coriell for data production. Existing sequencing data for NA12878 and NA24385/HG002 from the Genome in a Bottle (GIAB) effort (*90*) were also added to the final dataset. Table S2 provides the sample summary of the project and Table S36 provides a list of samples generated from each technology. The accessions to the raw data generated in this study can be found in the Table S37. The data flowchart can be found in Fig. S1.

Data resources: All data was shared publicly during the lifetime of the project via the International Genome Sample Resource (IGSR), the structure of which is described in Fairley et al. 2020 (*91*).

IGSR collates openly consented human genetic variation data, including data generated from the 1000GP cell lines. By using IGSR, data is shared with the community alongside, and integrated with, existing datasets, thereby improving discoverability. Within IGSR, data is organized into collections in the IGSR website data portal and via an FTP site (which also supports transfer using Globus and Aspera).

Our data is presented in the HGSVC2 data collection at https://www.internationalgenome.org/data-portal/data-collection/hgsvc2, where key datasets are highlighted in IGSR's data portal. This data can also be searched by population, sample, technology, or data collection, as illustrated in Fig. S37.

Within the HGSVC2 page, data can be filtered by technology, with a 'Download the list' button providing the URLs and metadata (such as md5s) needed to download a selected subset of the data as shown in Fig. S38.

On the FTP, data is available at ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/. Consistent with other datasets in IGSR, index files provide listings of the data files used by the project, across technologies. The IGSR FTP further organizes data using working (analysis files released during the project), technical (specific reference materials not generally available), and release (final datasets) directories as appropriate.

IGSR provides extensive FAQs (https://www.internationalgenome.org/faq), documentation on data access (https://www.internationalgenome.org/data#download), and email assistance at info@1000genomes.org.

2 Long-read, whole-genome sequence production

We generated and analyzed PacBio high-fidelity (HiFi)/circular consensus sequencing (CCS) for 9/38 samples and continuous long-read (CLR) data for 30/38 samples across three centers. Among them, three trios were sequenced by both HiFi/CCS and CLR.

Sample quality control (QC) and statistics can be found in Table S6. In addition, existing datasets for five samples sequenced as part of the GIAB or Genome Reference Consortium (GRC) were obtained from publicly available sources (Table S38).

(*Contributors: Katy Munson, Qihui Zhu, Scott Devine, Alex Lewis*)

University of Washington — Genomic DNA (gDNA) was isolated as previously described (*1*) and evaluated for purity and quantity using UV-Vis (Nanodrop 1000, Thermo Fisher) and fluorometric (Qubit, Thermo Fisher) assays. DNA sizing was checked on the FEMTO Pulse (Agilent) using the Genomic DNA 165 kb kit on extended mode. Samples all exhibited mode size above 50 kbp (most above 100 kbp) and were considered good candidates for PacBio sequencing. For HiFi/CCS, three parent–child trios (9 individuals) were selected and DNA was sheared using gTUBEs (Covaris) to a mode size ~15 kbp using 3200 RPM for four passes on an Eppendorf 5424R centrifuge. The sheared material was subjected to SMRTbell library preparation using the Template Prep Kit v1 (PacBio). After checking for size and quantity, the material was size fractionated on the SageELF instrument (Sage Science) using the protocol “0.75% 1-18kb v2”, size-based separation mode, and target value 3400 in well #12. Fractions were checked via fluorometric quantitation (Qubit) and pulse-field sizing (FEMTO Pulse). Additional libraries were prepared by g-TUBE shearing to 20 kbp using 4000 rpm (1500 rcf) in an Eppendorf 5415R centrifuge for six passes, followed by SMRTbell prep using the SMRTbell Express Template Prep Kit 2.0 kit and Enzyme Cleanup Kit (PacBio). These libraries were size fractionated on SageELF using a custom protocol “Waveform 250-100” timed mode, 3-hour run time. All cells were sequenced on a Sequel II instrument (PacBio) using 30-hour movie times. Fractions averaging roughly 11 and 14 kbp were prepared using version 1.0 sequencing chemistry and 2-hour pre-extension. Fractions averaging roughly 18 kbp were prepared with version 2EA sequencing chemistry and 4-hour pre-extension. Fractions averaging roughly 20 kbp were prepared with version 2.0 sequencing chemistry and 4-hour pre-extension. HiFi/CCS analysis was performed using SMRT Link v7.0, v7.1, v8.0, or v9.0, using filters of three full passes and an estimated read-quality value of 0.99. For CLR sequencing, without further shearing, isolated gDNA was SMRTbell library prepped using the Express Kit v2 (PacBio) and subjected to size selection on a BluePippin instrument (Sage Science) with a 40 kbp size cutoff. Libraries were loaded on a Sequel II using v2.0 binding and v2.0 sequencing kits, no pre-extension, and 15-hour movie times.

The Jackson Laboratory — High-molecular-weight DNA was extracted from frozen pelleted cells using the phenol-chloroform approach as previously described (*92*). Purified gDNA was assessed using fluorometric (Qubit, Thermo Fisher) assays for quantity and FEMTO Pulse (Agilent) for quality. For HiFi/CCS sequencing, samples exhibiting a mode size above 50 kbp were considered good candidates. DNA was sheared using gTUBEs (Covaris) to target 12 kbp fragments using the centrifugation settings at 5500 RPM for 60s. The sheared material was subjected to SMRTbell

library preparation using the Template Prep Kit v1 (PacBio). After checking for size and quantity, the material was size fractionated on the SageELF (Sage Science) using the protocol “0.75% 1-18kb v2”, size-based separation mode, and 3400 in Targae. Fractions were checked via fluorometric quantitation (Qubit) and pulse-field sizing (FEMTO Pulse). All libraries were sequenced on a Sequel II (PacBio) using 30-hour movie times. Fractions between 9 to 13 kbp were selected for sequencing using sequencing chemistry v1.0, 2-hour pre-extension and 30-hour movie times. HiFi/CCS analysis was performed using SMRT Link v7.0 or v7.1 by an estimated read-quality cutoff of 0.99. For CLR sequencing, samples exhibiting a mode size above 50 kbp were considered good candidates to proceed for sequencing. gDNA was sheared with passing through a needle, and the SMRTbell library was prepped using the Express Kit v2 (PacBio) and subjected to size selection on a BluePippin (Sage Science) with a 15 and 30 kbp size cutoff. Libraries were loaded on a Sequel II using v1.0/v2.0 binding and v1.0/v2.0 sequencing kits, no pre-extension, and 20-hour movie times.

University of Maryland — High-molecular-weight gDNA was prepared from cell pellets provided by Coriell Institute. The cells were lysed in a buffer containing proteinase K at 50°C overnight, followed by phenol/chloroform extraction and ethanol precipitation. The resulting gDNA typically showed a major peak at 165 kbp on a FEMTO pulsed-field run. For CLR sequencing, DNA samples were purified with SPRIselect beads (Beckman Coulter, Brea, CA) to remove small fragments and impurities prior to library preparation. Prior to sequencing, libraries were bound to polymerase with the Sequel II Binding Kit 2.0 (PacBio, Menlo Park, CA), and then sequenced with the Sequel II Sequencing Kit 2.0 and two 8M SMRT Cells on the Sequel II (PacBio) with 15-hour movie time. For HiFi/CCS sequencing, libraries were constructed using the SMRTbell Express Template Prep Kit 2.0 (PacBio) and fractionated on the SageELF (Sage Science, Beverly, MA) into narrow library fractions. Prior to sequencing, library fractions were bound to polymerase with the Sequel II Binding Kit 2.0 (PacBio), and then sequenced with the Sequel II Sequencing Kit 2.0 and 5-6 8M SMRT Cells on the Sequel II (PacBio) with 30-hour movie times.

3 Strand-seq production

(*Contributors: Ashley Sanders, Benjamin Raeder, Patrick Hasenfeld, Jan Korbel*)

We generated Strand-seq data for 38/38 samples targeted in this study (Table S36). EBV-transformed lymphoblastoid cell lines (Coriell Institute) were cultured in BrdU (100 uM final concentration; Sigma, B5002) for 18 or 24 hours and single isolated nuclei (0.1% NP-40 lysis buffer (*93*)) were sorted into 96-well plates using the BD FACSMelody cell sorter. In each sorted plate, 94 single cells plus one 100-cell positive control and one 0-cell negative control were deposited. Strand-specific DNA sequencing libraries were generated using the previously described Strand-seq protocol (*93*, *94*) and automated on the Beckman Coulter Biomek FX P liquid handling robotic system (*95*). Following 15 rounds of PCR amplification, 288 individually barcoded libraries (amounting to three 96-well plates) were pooled for sequencing on the Illumina NextSeq5000 platform (MID-mode, 75 bp paired-end protocol). The demultiplexed FASTQ files were aligned to the GRCh38 reference assembly (BWA 0.7.15) for standard library selection. Low-quality libraries were excluded from future analyses if they showed low read counts, uneven

coverage, or an excess of 'background reads' yielding noisy single-cell data, as previously described (Fig. S2) (*93*). Selected FASTQ files were used directly to guide the *de novo* reference-free assemblies, and the aligned BAM files were used for structural variant (SV) detection.

4 Illumina sequencing

(*Contributor: Michael Zody*)

We generated and analyzed Illumina whole-genome sequencing (WGS) data for 3,202 samples (*16*). WGS libraries were prepared using the TruSeq DNA PCR-Free Library Preparation Kit (Illumina) in accordance with the manufacturer's instructions. Briefly, 1 ug of DNA was sheared using a Covaris LE220 sonicator (adaptive focused acoustics). DNA fragments underwent bead-based size selection and were subsequently end-repaired, adenylated, and ligated to Illumina sequencing adapters. Final libraries were evaluated using fluorescent-based assays, including qPCR with the Universal KAPA Library Quantification Kit and Fragment Analyzer (Advanced Analytics) or BioAnalyzer (Agilent 2100). WGS libraries were prepared using the TruSeq DNA PCR-free Library Preparation Kit (450 bp). Library sequencing was performed on an Illumina NovaSeq 6000 instrument using 2×150 bp cycles. The average depth of coverage across all samples was 34× (min=28×, max=71×, median=34×) with >91% of the bases sequenced at base quality score of 30 or higher.

5 Bionano production

(*Contributor: Alex Hastie*)

We generated and analyzed Bionano Genomics Optical Mapping data for 33/38 samples. Bionano data were previously generated for NA12878 and HG002, and not available for the three samples HG02818, HG03125, and HG03486 (Table S36). Cell lines were obtained from Coriell and grown in RPMI 1640 media with 15% FBS, supplemented with L-glutamine and penicillin/streptomycin, at 37°C and 5% CO_2. Ultra-high-molecular-weight DNA was extracted according to the Bionano Prep SP Fresh Cells DNA Isolation protocol, revision C (Document #30257), using a Bionano SP Blood & Cell DNA Isolation Kit (catalog #80030). In short, 1.5 million cells were centrifuged and resuspended in a solution containing detergents, proteinase K, and RNase A. DNA was bound to a silica disk, washed, eluted, and homogenized via 1-hour end-over-end rotation at 15 rpm, followed by an overnight rest at room temperature. Isolated DNA was fluorescently tagged at motif CTTAAG by the enzyme DLE-1 and counter-stained using a Bionano Prep™ DNA Labeling Kit – DLS (catalog #8005) according to the Bionano Prep Direct Label and Stain (DLS) Protocol, revision F (Document #30206). Data collection was performed using Saphyr 2nd generation instruments (Part #60325) and Instrument Control Software (ICS) version 4.9.19316.1.

6 Hi-C data generation

(*Contributor: Qihui Zhu*)

We generated and analyzed Hi-C data for 33/38 samples. Hi-C data were not available for the following five samples: NA12878, HG002, HG02818, HG03125, and HG03486 (Table S36). Lymphoblastoid cell lines were obtained from Coriell Cell Repositories and cultured in RPMI 1650 supplemented with 15% fetal bovine serum. Cells were maintained at 37°C in an atmosphere containing 5% carbon dioxide. Hi-C was performed with Phase Genomics Proximo Hi-C kits v3.0 following the manufacturer's protocol using 1.5 M cells as input. Libraries were sequenced at New York Genome Center (NYGC) on an Illumina NovaSeq 6000 in a paired-end 150 bp format.

7 RNA-seq data generation

(*Contributor: Qihui Zhu*)

We generated and analyzed RNA-seq data for 33/38 samples. RNA-seq data were not available for five samples: NA12878, HG002, HG02818, HG03125, and HG03486 (Table S36). Total RNA of cell pellets were isolated simultaneously using QIAGEN RNeasy Mini Kit according to the manufacturer's instructions. Briefly, each cell pellet (1 million cells) was homogenized and lysed in Buffer RLT Plus, supplemented with 1% β-mercaptoethanol. The lysate-containing RNA was purified using an RNeasy spin column, followed by an in-column DNase I treatment by incubating for 10 min at room temperature, and then washed. Finally, total RNA was eluted in 50 uL RNase-free water. RNA-seq libraries were prepared with 300 ng total RNA using KAPA RNA Hyperprep with RiboErase (Roche) according to manufacturer's instruction. First, ribosomal RNA was depleted using RiboErase. Purified RNA was then fragmented at 85°C for 6 mins, targeting fragments ranging 250-300 bp. Fragmented RNA was reverse transcribed with an incubation of 25°C for 10 mins, 42°C for 15 mins, and an inactivation step at 70°C for 15 mins. This was followed by a second strand synthesis and A-tailing at 16°C for 30 mins, 62°C for 10 min. The double-stranded cDNA A-tailed fragments were ligated with Illumina unique dual index adapters. Adapter-ligated cDNA fragments were then purified by washing with AMPure XP beads (Beckman). This was followed by 10 cycles of PCR amplification. The final library was cleaned up using AMPure XP beads. Quantification of libraries was performed using real-time qPCR (Thermo Fisher). Sequencing was performed on an Illumina NovaSeq platform generating paired end reads of 100 bp at The Jackson Laboratory for Genomic Medicine. The RNA-seq QC statistics can be found in Table S39.

VARIANT DISCOVERY

8 Phased long-read genome assembly

(*Contributor: Peter Ebert*)

We applied a recently developed computational pipeline for phased genome assembly using Strand-seq (PGAS, Fig. 1A, (*3*)) with minor modifications to produce fully phased diploid genome assemblies without dependency on parent–child trio data for both CLR and HiFi datasets:

Filtering Strand-seq libraries. Differences in the experimental protocols for preparing Strand-seq libraries between the three family trios (CHS, PUR, and YRI; Strand-seq data available under ENA accession PRJEB12849) and the other 26 individuals in this study required an initial filtering step for the Strand-seq libraries. Based on a previously described methodology (*93*), we excluded control probes, technical dropouts, and libraries of medium quality if the total read count was less than 50,000, which provided little information to guide the assembly clustering process (see Methods in (*3*)). The complete list of excluded libraries is available in Table S40.

Generating squashed and phased assemblies. The initial non-haplotype resolved ("squashed") and the final phased assemblies for PacBio HiFi/CCS reads were generated as previously described (*3*). For the squashed assembly of PacBio CLR reads, we used Flye v2.6 (*69*) or, in cases where Flye v2.6 could not assemble the input reads, Flye v2.7 was applied with otherwise identical parameters. The phased assemblies for all CLR samples were generated with Flye v2.7. Following developer recommendations, we set the parameter "`--asm-coverage`" to 50 to lower Flye's memory consumption during the initial assembly steps; notably, despite this setting, all reads are used for the final assembly. The target genome size was set to the genome size of GRCh38 for the squashed assemblies, and to the total length of all sequences contained in the squashed-assembly cluster for the per-cluster haplotype assembly (see Methods in (*3*)):

```
flye --pacbio-raw {reads} --asm-coverage 50
-g {genome_size} -t {threads} --out-dir {output}
```

Polishing phased assemblies. The final phased assemblies for PacBio HiFi/CCS reads were polished as previously described (*3*). For the PacBio CLR assemblies, we performed a single pass of polishing with gcpp v1.9.0 (SMRT Link v9.0, PacBio) and used pbmm2 v1.1.0 (SMRT Link v8.0, PacBio) to generate the alignments of long reads to the assembled sequence for each cluster (see Methods in (*3*)):

```
pbmm2 align --log-level INFO --sort --sort-memory {sort_memory}M --no-
bai --alignment-threads {align_threads} --sort-threads {sort_threads}
--preset SUBREAD --min-length 5000 --sample {individual}
{input_contigs} {input_reads} {output}
```

```
gcpp --num-threads {threads} --algorithm=arrow --log-level INFO --log-
file {log} --reference {input_contigs} --output {out_fasta},{out_gff}
                        {input_alignments}
```

Generating haplotype read coverage tracks. For each sample, we generated coverage tracks for all three fractions of haplotagged reads (haplotypes 1 and 2, or unassigned) indicating haploid read depth along the human reference GRCh38. We aligned long reads to the human reference with pbmm2 (as above, preset "CCS" for PacBio HiFi) and generated bedGraph coverage tracks with BEDTools v2.29.0 (*96*). Coverage tracks were subsequently converted into the binary bigWig format using bedGraphToBigWig v377 (*97*) to reduce data transfer volumes:

```
          bedtools genomecov -bg -ibam {input_alignments} |
            LC_COLLATE=C sort --buffer-size={sort_buffer}M
        --parallel={threads} -k1,1 -k2,2n > {output_bedgraph}

    bedGraphToBigWig {input_bedgraph} {hg38_chroms} {output_bigwig}
```

8.1 K-mer-based analysis of phased assemblies

(*Contributor: Peter Ebert*)

The basic data structure for the k-mer-based analysis of the phased assemblies is a colored de Bruijn graph constructed with Bifrost (*98*). To benefit from CPU architecture optimizations, we compiled the Bifrost executable from source (git commit hash ab43065). For each phased assembly, a colored de Bruijn graph was constructed as follows:

```
   Bifrost build --input-seq-file {Illumina_reads} --input-ref-file
{phased_assemblies, GRCh38 reference} --output-file {graph} --threads
                {threads} --colors --kmer-length 31
```

The Illumina short reads were quality trimmed with Trim Galore v0.6.5 (*78*) and error corrected with Lighter v1.1.2 (*99*) to ensure that predominant reads consisting of high-quality genomic k-mers were used in the graph construction process:

```
trim_galore --quality 20 --length 51 --trim-n --output_dir {outdir} --
     cores 4 --paired {Illumina.pair-mate1} {Illumina.pair-mate2}

  lighter -r {Illumina.pair-mate1.trim} -r {Illumina.pair-mate2.trim}
      -k 31 3100000000 {alpha} -od {outdir} -t {threads} -zlib 6
```

The parameters for Lighter were set to match the Bifrost graph construction process (k = 31, genome size of GRCh38 3.1 Gbp), and "alpha" was computed following developer recommendations as 7 divided by the coverage of the Illumina short reads relative to the GRCh38 reference. Presence of regulatory sequences in the phased assemblies was assessed by querying the de Bruijn graph with k-mers generated from regions annotated in the Ensembl

Regulatory Build v98 (*100*). Regions shorter than the k-mer size of 31 bp were discarded (167 regions of type "transcription factor binding site" [TFBS] with a combined length of 2600 bp, approximately 0.0002% of all bases in TFBS regions). A query regulatory sequence is counted as present in the haploid assembly if 99% of its constituent k-mers are matched with at most one mismatch or indel error (Bifrost parameter "`--inexact`"):

```
Bifrost query --input-graph-file {graph} --input-color-file
{graph_colors} --input-query-file {input.queries} --output-file
{output} --threads {threads} --inexact --ratio-kmers 0.99
```

8.2 Reference-based analysis of phased assemblies

(*Contributor: Peter Ebert*)

We aligned all phased assembly contigs to the GRCh38 reference genome as previously described (*3*) and examined aligned contig coverage for several annotation data tracks. If not indicated otherwise, alignments were filtered to only retain aligned contigs with the highest alignment quality (MAPQ = 60), and potentially overlapping contig alignments were merged using BEDTools' "`merge`" command (*96*). GRCh38 reference annotation tracks for cytogenetic bands, segmental duplications (SDs), assembly gaps, repeat elements, and functional elements were downloaded from the UCSC Table Browser (*101*–*103*). GRCh38 locations of centromeres and of unresolved sequence issues (type "Gap" or "Unknown") were downloaded from the GRC (*104*) on 2020-07-23. All issue annotations without sequence coordinates or with an indicated fix version (minor GRCh38.p14 or major GRCh39 release) were discarded. Raw and processed versions of the issue annotation are available in the PGAS assembly pipeline repository under "annotation/grch38" (see Code Availability). Illumina short-read genome coverage tracks indicating "Illumina-accessible" regions generated as part of the 1000GP (*15*, *105*) were downloaded for both the "pilot" (lenient) and strict annotation (ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/working/20160622_genome_mask_GRCh38). If applicable, the GRCh38 assembly was modified to define accessible regions of the genome by excluding Giemsa-stained positive or variable regions, roughly corresponding to heterochromatin, and by subtracting N gaps (Table S5). Haploid assembly completeness was furthermore assessed by running QUAST-LG v5.0.2 (*106*) with the option "`--conserved-genes-finding`" to scan for the presence of single-copy genes (BUSCO analysis (*107*)), plus the option "`--features gene:GENEMODEL`" to scan for the presence of genes annotated in the specified gene model (GENCODE v31 (*108*)).

Phased assembly gaps relative to GRCh38. For the analysis of common gaps in the haploid assemblies, we identified candidate regions at several levels of stringency: we selected all regions where none of the assemblies had any contig alignment with minimum mapping quality of 0, 10 and 20, and excluded regions smaller than 10 kbp. The resulting list of common assembly break candidate regions was then subjected to a LOLA v1.12 enrichment analysis (*109*) to reveal underlying sequence features that may cause these assembly breaks. The necessary control set for the LOLA enrichment analysis was derived in a way similar to the candidate region set, but by

selecting all regions where at least one assembly did not have a contig alignment with the required minimum mapping quality. The LOLA analysis was performed with standard parameters using the above listed UCSC annotation tracks.

8.3 Assembly scaffolding

(*Contributors: Peter Ebert, Feyza Yilmaz, David Porubsky*)

Bionano cmaps were aligned to phased assembly haplotypes to identify haplotype 1 and haplotype 2 cmaps using Bionano RefAligner.

```
python2.7 Solve3.5.1_01142020/Pipeline/1.0/runCharacterize.py -t
Solve3.5.1_01142020/RefAligner/1.0/RefAligner -q {querymap} -r
{referencemap} -o {outputfolder} -p Solve3.5.1_01142020/Pipeline/1.0/
-a
Solve3.5.1_01142020/RefAligner/1.0/optArguments_haplotype_DLE1_saphyr_
human.xml -n 64 1>alignmentstatistics.out
```

Bionano genome maps aligning to phased assembly haplotype 1 were assigned as haplotype 1, and genome maps aligning to phased assembly haplotype 2 were assigned as haplotype 2. To obtain scaffolds, phased assemblies were mapped to *de novo* assembly of Bionano cmaps using the Bionano Solve3.5.1_01142020 hybrid scaffolding pipeline.

```
perl Solve3.5.1_01142020/HybridScaffold/12162019/hybridScaffold.pl -n
{inputNGSFASTA} -b {inputBionanocmap} -c
Solve3.5.1_01142020/HybridScaffold/12162019/hybridScaffold_DLE1_config
.xml -r Solve3.5.1_01142020/RefAligner/1.0/RefAligner -o
{outputFolder} -f -B 2 -N 2 -x -y -m {inputBionanoMoleculesbnx} -p
{inputDeNovoAssemblyPipelineDirectory} -q
Solve3.5.1_01142020/RefAligner/1.0/optArguments_nonhaplotype_DLE1_saph
yr_human.xml -e {inputDeNovoAssemblyNoiseParameter}
```

We combined the information from the scaffolded Bionano hybrid assemblies with the contig-level phased assemblies as follows: we established the correspondence between Bionano scaffold and phased assembly cluster, i.e., human chromosome, based on the amount of mapping quality-weighted contig sequence aligned to the GRCh38 reference chromosomes. For cases where either less than 50% of the scaffolded sequence aligned to a single chromosome, or more than 50% pertained to unplaced sequence (“chrUn”), the respective scaffold was marked as entirely “unplaced”. We used the scaffold-to-chromosome assignment to characterize misassemblies detected and corrected in the Bionano hybrid scaffolding process (see Section “Analysis of phased assemblies”). Next, we re-aligned all scaffolded phased assembly contigs to the matched chromosome (chromosomes X and Y for male samples), and all “unplaced” scaffolded sequences to the full GRCh38 reference (chromosomes 1-22, X; plus chromosome Y for male samples). All unsupported sequences, i.e., those parts of our phased assemblies that could not be scaffolded by Bionano, were split into 500 bp reads and aligned to GRCh38 using minimap2’s “-sr” preset.

The resulting read alignments were aggregated in genomic bins of 500 kbp and averaged over all samples before plotting.

For the realigned scaffolded contigs, we derived confidence levels as follows: scaffolds with aligned contigs scattered over at least two chromosomes were assigned low confidence. Scaffolds comprising a single aligned contig were assigned high confidence. For all other cases, a custom script was used to select all alignments within a region of the estimated scaffold length (plus 5% on both ends). The region center was chosen as alignment-weighted midpoint over all contig alignments, thus giving more weight to larger contigs that are part of the scaffold. In case no alignment was selected, and the scaffold comprised at least three contigs, the contig selection was adjusted by discarding the contig with maximal average distance to all other contigs, and subsequently recalculating the region midpoint. This strategy was limited to a single attempt to avoid placing scaffolds based on marginal evidence collected from few contig alignments. If no alignment could be selected to place the scaffold, e.g., because contig alignments were scattered along the entire chromosome, all contigs were assigned low confidence. If all contigs belonging to a scaffold were selected, and their alignment order was identical to the scaffolded order, the contigs were labeled as high confidence, and medium confidence otherwise.

8.4 Analysis of phased assemblies

(*Contributor: Peter Ebert*)

The design of our study enabled us to directly compare the results of our phased genome assembly pipeline for two types of PacBio long reads: less accurate, but on average longer CLR reads, and highly accurate CCS reads (HiFi; Fig. 1B). Notably, HiFi data are more expensive; on average, six SMRT cells per sample were sequenced to produce the HiFi data in this study, compared to only two SMRT cells per sample for CLR data (Table S6). We report statistics from different stages of our phased genome assembly process (Fig. 1A) that suggest varying dependency between pipeline performance and type of the input long reads. The initial step of creating a non-haplotype-resolved (“squashed”) assembly using all input reads resulted in highly contiguous assemblies with a mean contig-level N50 of 30.5 Mbp (Fig. 1D). However, only CLR-based assemblies exceeded N50 values of 35 Mbp in 20% of all cases, presumably due to the longer insert sizes of the sequencing libraries (Fig. S3, Table S6). These differences in assembly contiguity are less relevant for the subsequent step of clustering contigs into roughly chromosome-scale clusters using the Strand-seq data (see Methods in (*3*)). Consequently, on average, 98.4% of the sequence in one cluster aligns to a single GRCh38 reference chromosome (1-22, X). The pipeline step of identifying heterozygous (HET) single-nucleotide variants (SNVs) to obtain local phasing information indicates differences between CLR- and HiFi-based assemblies with an average of 2.32 and 2.66 million HET SNVs, respectively (Fig. S40). These numbers are, however, dominated by the high-diversity African population with a technology-agnostic average number of approximately 2.92 million HET SNVs per assembly, compared to 2.2 million HET SNVs for all other populations combined. These discrepancies seem to have little bearing on variant phasing or haplotagging efficiencies. In general, our assembly pipeline phases more than 99.9% of all HET SNVs irrespective of long-read technology or sample population of origin (Fig. S40). This information is then used to haplotag an average of 84.3% of all input long

reads, with African samples reaching mean haplotagging rates of 89.8% versus 81.7% for non-African samples (Fig. 1C). The assemblies built from the haplotagged read sets reach mean contig-level N50 values of 28.4 Mbp for CLR and 21 Mbp for HiFi reads (Fig. 1D). Notably, 38.3% of all CLR haplotype assemblies exceeded a contig-level N50 of 30 Mbp—with a maximum of 39.4 Mbp—compared to only 3.6% for the HiFi-based haplotype assemblies. We computed base quality (QV) estimates for all polished haplotype assemblies in two orthogonal ways: first, we used Illumina short reads to identify homozygous variants as potential sequence errors (see Methods in (*3*)), which resulted in an average QV estimate of 49.6 (HiFi 53.8, CLR 47.7; Fig. 1E). Second, we used 31-mer counts either unique to the assembly or supported by Illumina short reads or by the GRCh38 reference sequence to calculate an average assembly QV value of 40.4 (HiFi 43.1, CLR 38.8; Fig. 1E). The assembly 31-mer counts supported by Illumina short reads can also serve as an estimate for the amount of sequence that is not present in the current GRCh38 reference. On average, we count 50 million homozygous and 38 million haplotype-specific 31-mers with Illumina support per sample (Fig. S41). The differences in haplotype-specific k-mer counts between CLR and HiFi assemblies are small (37.2 to 39.6 million) when contrasted with the comparison of African to non-African samples (46 to 34.2 million).

We further evaluated the completeness of our phased assemblies relative to various annotations for the GRCh38 genome reference (see Section “Reference-based analysis of phased assemblies” for annotation sources). We plotted the contig coverage along the cytogenetic bands as the mean fraction of covered bases in each chromosomal segment (Fig. S6). This high-level view indicates that the majority of gaps in our assemblies coincides with (peri-) centromeric regions or occurs near the end of chromosomes (see also assembly break analysis in next paragraph). This observation is confirmed by an average 2.3% haploid assembly contig coverage in centromeres, with the exception of chromosome 20, where up to 38.7% of the centromere is covered by aligned contigs for HiFi assemblies (Fig. S7). Overall, we computed a median contig coverage of our haploid assemblies of >95% in accessible regions of the GRCh38 chromosomes 1-22 and X (Table S5). We also observe high-quality alignments (MAPQ 60) on several GRCh38 “ALT” contigs (Fig. S39); at thresholds of 50%, 75%, and 90% of ALT base pairs covered by high-quality alignments, we identify 34 (18,159,658 bp), 23 (12,916,125 bp), and 18 (10,908,115 bp) ALT contigs represented in our assemblies, respectively. In all these cases, the median number of haploid assemblies containing the ALT contigs is between two and three. However, several outliers to that statistic can be easily identified (Fig. S39), indicating that some ALT contigs, e.g., on chromosomes 5, 6, 7 and 19, are present in almost all haploid assemblies.

Next, we performed the same coverage analysis with a set of 98 GRCh38 regions annotated as “unresolved issues” by the GRC (issue type “gap” partially coincides with centromeres). Our phased assemblies have, on average, assembled sequences for 67% of the determined bases in those regions, with no substantial difference between CLR and HiFi assemblies (66% to 68.3% coverage, Fig. S42). Other genomic regions that are prone to coverage dropouts in assemblies are often enriched in SDs. For the UCSC SD annotation track, we observe a clear difference in contig alignment coverage between CLR and HiFi assemblies of 65% to 73.2%. We also examined the presence of known regulatory elements annotated in the Ensembl Regulatory Build v98 (*100*) based on their constituent k-mers (see Section “k-mer based analysis of phased

assemblies"). This analysis did not reveal any sizable differences between CLR or HiFi assemblies: approximately 90% of the sequences annotated as regulatory regions is detectable in both haplotype assemblies, and 3.2% is specific to only one haplotype (Fig. S43).

Besides haploid assembly completeness in terms of regulatory elements, we also find that almost all genes in the GENCODE v31 annotation (*108*) are contained in our assemblies (median complete hits 92.3%, partial hits 3.7%; Table S5). This observation is essentially replicated when restricting the analysis to single-copy genes (BUSCO (*107*)), where we find a median of 96.3% complete and 2% partial hits in our phased assemblies (Table S5).

To systematically characterize regions that we identified as common assembly breaks (see Section "Reference-based analysis of phased assemblies"), we performed enrichment analyses at various thresholds of contig alignment quality to account for potential alignment artifacts. At a minimum size of 10 kbp, we identified between 230 and 236 regions across all haploid assemblies as candidates for common assembly breaks. As expected, the results of our enrichment analysis include regions of unresolved sequence in GRCh38 ("N gaps"), which cannot be aligned and thus provide no mechanistically relevant insight. Besides, enrichment in genomic annotations such as RNA repeats (OR > 7), SDs (OR > 3) and microsatellite repeats (OR > 2) suggests that genome assembly in these sequence contexts is still a challenge (Table S7). While it is expected that difficult-to-assemble regions such as centromeres lead to genuine gaps in the haploid assemblies, GRCh38-relative haploid contig coverage can only approximate the actual completeness of the *de novo* assemblies due to the unresolved N gaps in GRCh38.

We note that we also observe presumably technology-related differences in difficult-to-assemble genomic regions such as chromosome 3q29 (Fig. 4C, Table S46): here, CLR assemblies more often resolve the entire region (35%, 21/60) compared to HiFi assemblies (18%, 5/28). However, given the imbalance between the number of CLR and HiFi assemblies in our study, and other confounding factors of unclear importance such as the respective genome assembler, these observations are of anecdotal quality.

As an orthogonal method for assessing assembly completeness that does not rely on aligning contigs to a reference, we scaffolded our contig-level phased assemblies for 32 out of the 35 individuals for which Bionano optical maps were available (see Section "Assembly scaffolding"). This enabled us to detect and characterize various types of misassemblies (Fig. S8), the most frequent one being a contig break due to missing support from Bionano. Despite these assembly breaks (median of 65 misassemblies per haploid assembly), the median concordance between our phased assemblies and the Bionano maps is >97% (CLR median 98.2%, HiFi median 92.0%; Table S8). We aligned the unscaffolded sequence back to GRCh38 and observed a tendency for elevated read coverage in 500 kbp genomic bins around centromeres and in acrocentric regions, suggesting that at least some of the unscaffolded sequence may originate from those regions (Fig. S9).

Next, we realigned the scaffolded contigs to GRCh38 and classified contig alignments as high, medium, or low confidence. Across all assemblies, a median of 69.01% of the aligned sequence

is part of a high-confidence scaffold, with CLR samples showing an overall lower percentage (median 65.69%) compared to HiFi (median 79.03%). The majority of the remaining sequence pertains to contig alignments within medium-confidence scaffolds (median CLR 31.74%; HiFi 20.37%). For the remaining scaffolds, the majority of the contig alignments are scattered across one or more chromosomes, making it impossible to reasonably infer an approximate genomic location for the respective scaffold. Consequently, we consider the remaining contig alignments as low confidence (median CLR 0.62%; HiFi 0.3%), reflecting the view that the scaffolded assembled sequence cannot be adequately represented via GRCh38-relative alignments.

Finally, we combined the alignment information from contig-level and scaffolded assemblies to estimate the amount of sequence that can be interrogated with our phased assemblies. As a standard of comparison, we selected Illumina coverage tracks generated by the 1000GP (see Section “Reference-based analysis of phased assemblies”) that define between ~2.29 Gbp (strict) and ~2.75 Gbp (lenient) of the human genome as “accessible” to short-read-based analyses (excluding N gaps in the GRCh38 reference, restricted to chromosomes 1-22, X and Y). We defined an analogous set of regions in two ways: first, we selected all regions where at least one phased assembly had a MAPQ 60 contig alignment; three samples (HG03486, HG02818, HG03125) were excluded from this procedure because no corresponding Bionano data were available for hybrid scaffolding of these samples. Second, we selected all regions with contig alignments that were part of at least one high- or medium-confidence scaffold, i.e., the second set of regions does not enforce a threshold on the contig alignment MAPQ, thus being more permissive in that regard. Notably, we do not consider our scaffold-derived definition of “assembly accessible” regions as necessarily more lenient than the MAPQ-based regions because we included orthogonal information from Bionano supporting contig alignments that may be below a stringent MAPQ 60 threshold. Both resulting region sets cover approximately 2.87 Gbp of sequence, with as few as 250 and 243 regions, respectively (Fig. S44). The median region size is ~397 kbp (N50 ~58 Mbp) and ~480 kbp (N50 ~61 Mbp), compared to 79 bp (N50 ~2 kbp) and 240 bp (N50 ~20 kbp) for the strict and lenient Illumina region sets. Our definitions include ~582 Mbp more sequence relative to the strict definition of Illumina-accessible regions, and ~123 Mbp relative to the lenient definition. However, we also find that approximately 848 kbp and 121 kbp of the sequence in the Illumina strict set is not covered by our MAPQ60 or scaffold-derived coverage tracks, respectively. The lower amount of sequence that is missing from the scaffold-derived region set suggests that thresholding on the alignment quality excludes many regions that may be difficult to align to. We thus first removed all Illumina-exclusive strict regions overlapping centromeres (~2 kbp), which we already know not to be accessible by our assemblies (see above), and then annotated all remaining Illumina strict regions with their distance to the closest SD. As expected, we observe a median distance of 40 kbp (mean: 53.5 kbp) between the Illumina region and the closest SD for regions missing from our MAPQ60 coverage track, and a median distance of 0 kbp (mean: 5.9 kbp), i.e., overlapping, for Illumina regions missing from our scaffold-derived coverage track. We observed a similar behavior when checking how many of our SV calls based on the phased contig-level assemblies are located outside of our region sets. For our MAPQ60 assembly coverage track, we found 1,222 SV calls to be uncovered. This number was reduced to 113 for the coverage track based on the scaffolded assemblies, which in turn does not

include regions covered by contig alignments as part of low-confidence scaffolds, irrespective of the MAPQ value of the individual contig alignments.

9 Reference and global merging strategy

(*Contributor: Peter Audano*)

9.1 Genome references

GRCh38 No-ALT reference. Long reads and contigs were aligned to the GRCh38 primary assembly only, which includes chromosome scaffolds, unplaced contigs, and unlocalized contigs. No ALTs, patches, or decoys were included, which were constructed for short-read mapping and would confound long-read and assembly analysis. This reference was used by multiple variant discovery pipelines including PAV and PBSV (see below). The No-ALT reference is available via the HGSVC2 data portal (see Section 1 “Samples and data availability”).

9.2 Variant merging strategies

Variant merging was done for several analyses in this study, including merging haplotypes into a diploid callset, merging calls into a nonredundant set, and comparing SVs with other callsets. Merging variants by 50% reciprocal overlap (RO) has been used for numerous SV papers (*1*, *4*, *19*), but it often under-merges small SVs and indels.

To address this problem, we adopt a three-step approach for SVs and indels. First, exact match variants are intersected (same size and location for insertions and deletions). Second, the RO is applied to variant calls for variants with intersecting reference coordinates. For insertions, the end position is the sum of the start position and variant length. Since an RO-only approach often misses smaller variants, we finally merge variants within 200 bp and 50% overlap by size (i.e., maximum RO if variants were shifted) (Fig. S45). For the 200 bp distance, we take the minimum of the start and end position offsets. SNVs are only intersected by exact match (same position and alternate base). In all cases, only the same variant classes are considered for matches (INS with INS, DEL with DEL, etc.). During this merging process, a variant in a new sample could support a call if it intersects the lead variant of a merged set (i.e., it is not matched against other supporting variant calls). The code implementing this approach is available by the HGSVC as SV-Pop (https://github.com/EichlerLab/svpop, project version released via (*88*)).

10 Phased assembly variant (PAV) discovery

(*Contributor: Peter Audano*)

Variants including SVs were called by directly comparing the haplotype-resolved assemblies to the human genome reference, GRCh38. This process is implemented in the Phased Assembly Variant (PAV) discovery tool described in this section (available at https://github.com/EichlerLab/pav and released as part of this manuscript via (*88*)).

10.1 Contig alignment and trimming

Contig alignment. For each haplotype, contigs were aligned to the GRCh38 No-ALT reference with minimap2 2.17 with parameters "-x asm20 -m 10000 -z 10000,50 -r 50000 --end-bonus=100 --secondary=no -a -t 20 --eqx -Y -O 5,56 -E 4,1 -B 5" (*110*). The callset was generated using minimap2 alignments. For validations, we also ran the pipeline with LRA (*76*) alignments using parameters "-CONTIG -p s -t" (PAV-LRA).

Alignment trimming per haplotype. Minimap often makes redundant alignments where the same part of a contig is aligned to more than one location. Alternatively, a location of the reference may be covered by more than one alignment record, and these overlaps often occur around SVs that truncate alignments (i.e., one contig in multiple alignment records). For example, for a large deletion flanked by repeats, the single contig copy of the repeat is often mapped to both reference copies. This obscures the size of the SV and introduces artifacts in the flanking repeats that appear to be SNVs, indels, and small SVs. A similar anomaly was observed for tandem duplications where multiple contig copies were aligned to the same reference copy. To address these multiple-mapping issues, we trimmed alignment records to resolve multiply-mapped contig bases followed by a second round to trim multiply-mapped reference bases. For a pair of alignment records, trimming is performed with a dynamic programming algorithm that attempts to maximize the number of variant events removed, conditioned on zero overlap of trimmed alignment. One variant event is defined as a single SNV, insertion, or deletion regardless of size. In short, it takes one contig and finds the cut-site if only that contig were trimmed. It then traverses the CIGAR operations of both alignment records using the zero-overlap and maximum-event conditions to guide its progress without considering more cut-site combinations than necessary (Fig. S46). The trimmed bases are soft-clipped. Lastly, whole alignment records less than 1 kbp before or after trimming, or if it is completely contained within another alignment record. One haplotype can be processed within approximately 1-2 core minutes. After alignment trimming, each assembly base maps to, at most, one reference base and vice versa. The alignment-truncating variant caller in PAV (see below) then uses these refined breakpoints to make an SV call.

The variants in Fig. S46 were aligned by LRA without fragmenting assemblies, and the SVs were called directly from the CIGAR string. Both minimap2+alignment trimming and LRA place breakpoints at nearly identical locations and call an SV of a similar size.

10.2 Inter-alignment and alignment truncating variants

Inter-alignment variant calling per haplotype. Most variant calls are contained within alignment records and can be obtained by examining the alignment CIGAR string. For this, PAV requires alignments with "=" and "X" CIGAR operations (an error is produced if there are any "M" CIGAR operations). These variant calls rely on the aligner for correct placement.

Alignment-truncating variants. SVs often cause contig alignments to break, leaving more than one alignment record for the same contig. Because alignments are trimmed, the insertion and deletion breakpoints are mapped at the ends of the alignment records and can be identified if they leave

an excess of reference bases (deletion) or contig bases (insertion). Both alignment records must be in the same orientation or SV discovery is not attempted. The variants are detected by finding alignment records matching the same chromosome and contig. The reference and contig size gap between the two alignment records is compared to the minimum alignment length of two records (lesser of the number aligned bases in each alignment). If the minimum alignment length is greater than the contig gap size or greater than three times the reference gap size, the pair are ignored and no SV call is attempted.

10.3 Inversion detection

Flagging inversion signatures. Inversions in assemblies either fragment alignments into alternating forward- and reverse-oriented records, or they cause aberrant variant patterns if contigs are aligned through them without inverting. We define these as inter-alignment and intra-alignment inversions, respectively. Intra-alignment inversions often result in matched insertion and deletion events of a similar size, and they are often accompanied by clusters of false SNVs and indels (Fig. S47-A). We take advantage of this by using matched SVs and indels to flag regions that may contain an inversion. Inter-alignment inversions are flagged by alignment-truncating events (Fig. S48-A).

Inversion detection with k-mer density. Each intra- and inter-alignment flagged site is a reference region (contig, start position, and end position). It is initially expanded by 4 kbp because flagged sites are often much smaller than the inversion. K-mers of size 31 (31-mers) are extracted from the reference region and counted. If a reference 31-mer appears more than 100 times in that region or if the region contains N bases, inversion discovery terminates to avoid excessive CPU time spent on unresolvable loci.

Using the alignments, the reference region is lifted to a contig region. If the lift fails or the reference and contig region sizes are not within 60%, inversion detection is terminated. The contig region is extracted and translated into a list of k-mers in the order they appear in the contig region. Reference k-mers are reverse complemented if contig alignment records suggest that the contig is in the opposite orientation of the reference. Using the reference k-mers, each contig k-mer is assigned to one of three states: FWD (reference k-mer set has the k-mer), REV (reference k-mer set has the reverse-complement of the k-mer), or FWDREV (reference k-mer set has both FWD and REV k-mers). Contig k-mers with no representation in the reference k-mer set are dropped. If any state (FWD, REV, FWDREV) has fewer than 20 k-mers, those k-mers are also dropped because they cause large false spikes in density calculations that confound analysis. If this results in fewer than 2,000 k-mers, inversion detection is terminated.

A density function is constructed to guide inversion detection (Fig. S47-B and Fig. S48-B). First, the remaining informative k-mers in FWD, REV, and FWDREV states are re-indexed consecutively from 1 (1 ... nth k-mer), which reduces biases in the density function if an inversion contains variants (e.g., Alu insertion), but the original index is retained so coordinates can be accurately translated back to the contig region. A Gaussian-kernel density function is then computed over these indices. Density bandwidth is computed using Scott's rule with one-

dimensional data using all informative k-mers, and the density function is constructed with Scipy 1.2.1 (module “scipy.stats.gaussian_kde”). One density function for each state is then computed. When the density function is applied to a k-mer index, it is subsequently scaled by the number of k-mers with the same state (FWD, REV, FWDREV) so that density-smoothed values for each state approach 1 regardless of the number of k-mers in each state, which is needed for maximum state computations.

Density calculations are initially computed on one of every 20 informative k-mers and the blocks are subsequently filled in by interpolation (to save CPU cycles) or full density computation. If there are no state changes within the block and the density value between each side of the block changes by less than 0.005, the block is filled in by interpolating with the density values on each side; otherwise, a full density computation is performed for each index in the block.

Finally, each k-mer is assigned a new state using the maximum density, which yields smoothed runs of FWD, REV, and FWDREV states. An inversion is detected if the contig region begins and ends with reference oriented k-mers and contains inverted k-mers. The outer breakpoints of the inversion are placed at the first and last non-FWD states in the state list, and these outer breakpoints are the reported breakpoints for the inversion event. Often, k-mers are flanked by inverted duplications, so a typical inversion state progression is “FWD, FWDREV, REV, FWDREV, FWD”. In this case, inner breakpoints are placed at the first and last REV state(s) and reported as an annotation with the inversion call. Finally, the breakpoints identified in contig coordinates are lifted to the reference (Fig. S47-C and Fig. S48-C) and a variant call is constructed with inner- and outer-breakpoint annotations in contig and reference coordinates.

If inverted k-mers are identified in the region but density calculations did not yield smoothed states indicating an inversion (FWD states on flanks with REV states between), then the region is expanded by 50% and the density is recomputed. If one flank has FWD k-mer states, then the expansion is biased to add more bases on the non-FWD-state end. If only FWD k-mers are found after three expansions (including the initial expansion), inversion detection terminates. There is currently no upper limit, so expansion could continue until alignment records for the contig are exhausted.

10.4 Callset finishing

In-PAV filtering. Variant calls intersecting inversions are removed. Variants occurring within inter-alignment deletions are also removed. We observed that incomplete phasing generated false contigs that fit into large heterozygous deletions, which was likely created by reads from the other haplotype. Ignoring variants inside deletions on the same haplotype removes false calls from these phantom contigs and from misalignments.

Merging haplotypes. An independent callset is first generated for each haplotype and they are merged using the three-step strategy (see Section 9.2). As before, SNVs are intersected only if they match exactly (position and alternate base). PAV merges SVs and indels as one large set of variants and then splits them into SV/indel variant classes post-merge. For all homozygous

variants, PAV chooses the variant representation from haplotype 1 as a default by the merging process.

10.5 Post-PAV callset filters

SDA. Many high-identity repeats are still not resolvable by assemblers and result in assembly collapses, which carry a mix of paralog-specific variants (PSVs) from multiple sites and may or may not align to the best reference copy. We identified collapsed regions with Segmental Duplication Assembler (SDA, (*74*)) and filtered PAV calls if they intersect a collapse on either haplotype. To avoid over-filtering inversion and deletion SVs, we require at least 50% of the SV to intersect a collapse before removing. Indels and SNVs are removed if they intersect collapses by 1 bp or more (Table S41).

Misclustered contigs. Using cluster IDs generated by the assembler, which were embedded in the contig names, PAV assigns the best reference chromosome for each cluster if 85% or more of the mapped bases within contigs 1 Mbp or greater are assigned to the same reference chromosome. For chromosomes where a max was defined, variants called on contigs belonging to a different cluster are filtered out. This prevents misaligned contigs (or parts of contigs) from inflating the callset.

Pericentromeric filter. A centromere and pericentromeric region filter (*1*, *4*) is applied to address regions that are difficult to replicate among callsets.

10.6 PAV versus dipcall and svim-asm

We compared PAV to both dipcall (*111*) and svim-asm (*112*) for three child samples (HG00733, HG00514, and NA19240) by searching for SV support within 1 kbp (Fig. S49, Table S56). PAV supports 95% of dipcall SVs and 89% of svim-asm SVs. Considering only PAV calls that passed QC with orthogonal support, we find that on average these callers miss 1,915 and 1,201 validated SVs for dipcall and svim-asm, respectively. Compared to the unfiltered PAV callset, we estimate 4.7% FDR for dipcall and 11.1% FDR for svim-asm and note that 3.6% (dipcall) and 7.5% (svim-asm) of the PAV-unsupported calls are in regions trimmed by PAV using a dynamic programming approach to remove redundantly mapped reference and contig bases before making variant calls.

Given these results and that the svim-asm callset contains between 28,000 and 32,000 SVs (about 5,000 more than expected), we assert dipcall is more specific but less sensitive while svim-asm is more sensitive but less specific when compared with PAV. While dipcall and svim-asm are themselves useful tools, PAV is best suited for this experiment because of its ability to remove spurious alignments, which boosts specificity without sacrificing sensitivity. This method is now reported and released as part of this manuscript (https://github.com/EichlerLab/pav (*88*)).

10.7 Comparing variant discovery from HiFi and CLR

Application of both HiFi and CLR long-read technologies on the same samples allowed us to

compare their performance. While HiFi sequencing was more expensive (on average 6 HiFi vs. 2 SMRT cells in this study), HiFi-based assemblies are more accurate. CLR-based assemblies are more contiguous and resolve the complex 3q29 locus (Fig. 4C) more often (35% or 21/60) than HiFi-based assemblies (18% or 5/28). As sequencing technology and assembly algorithms continue to evolve (*113*, *114*), HiFi sequencing has been predicted to predominate because accuracy and higher depth will afford access to even more previously inaccessible regions (*113*, *114*). Nevertheless, orthogonal technologies such as optical maps and Strand-seq data are still required. Strand-seq, for example, was critical to achieve chromosome-length phasing in the absence of parental sequencing data (*115*). Given the challenges of broadly obtaining parental material from populations as well as patients, trio-free haplotype-resolved assembly will be a major asset for further expanding human genome diversity and the discovery of more complex forms of pathogenic variation (*116*, *117*).

11 Variant discovery with aligned reads

To increase sensitivity, we also applied methods that call variants based on mapping of the underlying raw long-read data against the human reference genome (GRCh38). These methods have been optimized for particular SV classes (e.g., PBSV for SVs and DeepVariant for indels and SNVs).

11.1 PBSV

(*Contributors: Aaron Wenger and Peter Audano*)

Reads were aligned to GRCh38-NoALT (see Section 9.1 “Genome references”) using pbmm2 v1.2.1 (SMRT Link v9.0, PacBio) with "--sort --preset CCS -L 0.1 -c 0" for CCS and “--sort --median-filter” for CLR. Variants were called using PBSV v2.3.0 (SMRT Link v9.0, PacBio). The PBSV workflow was executed separately per sample (not jointly) and per chromosome. First, SV signatures were discovered with "pbsv discover --tandem-repeats <SR.bed> -r <CHROM>" where SR.bed is the GRCh38 UCSC simpleRepeats track with regions within 200 bp merged ("bedtools merge -d 200"). Next, variants were called with "pbsv call" with "--ccs -O 2 -P 20 -m 10" for CCS and "-m 10" for CLR. For each sample, the per-chromosome calls were concatenated, sorted, and compressed with BCFtools 1.9 (*118*).

11.2 DeepVariant

(*Contributor: William Harvey*)

CCS reads were aligned to the human reference genome GRCh38 with pbmm2 v1.2.1 (SMRT Link v9.0, PacBio) with “--preset CCS --min-length 5000”. DeepVariant v0.9.0 (*67*) uses a deep neural network with a pre-trained model (--model_type=PACBIO) specifically for PacBio CCS reads. The pipeline consists of three separate steps: make_examples, call_variants, and postprocess_variants. The first step utilizes the existing alignments to make a six-channel (read base, quality score, mapping quality score, read strand, read allele support, read match reference)

TensorFlow object, which is then processed by the neural network using the pre-trained model to make variant calls. The variant calls are processed and combined into a VCF.

11.3 DeBreak

(*Contributors: Zechen Chong, Yu Chen*)

Reads were aligned to human reference genome GRCh38 with minimap2 (*119*) version 2.15-r905 with the option "--secondary=no" for both CLR and CCS samples. DeBreak version v1.0.2 (https://github.com/ChongLab/DeBreak, (*88*)) was executed to call variants from read-to-reference alignments for each sample independently with option "--poa --ref hg38.fa". For each sample, DeBreak scanned all read alignments for both intra- and inter-alignment SV signals. Smaller indels can be detected within a single alignment. For larger indels, inversions, duplications, and translocations (potential mobile element insertions or MEIs), DeBreak utilized split-read information to infer SV breakpoint positions and lengths. All raw SV calls were sorted and clustered with density-based clustering to generate SV candidates. Auto-adaptive filters were applied to remove noise and artifacts. Only variants that passed all filters were retained, while variants with "HighCov" or "LowMapQ" filters were discarded to ensure accuracy of SV callsets. Partial order alignment (*120*) was carried out using variant-supporting reads for each variant to refine SV breakpoints. SVs located in ALT contigs were also removed from the callset (Table S42). The DeBreak SV callset was compared with the PAV callset for each sample with 50% RO (Fig. S50). For CLR data, the numbers of DeBreak deletions overlapping with PAV ranged from 6,882 to 8,738, while insertions ranged from 10,595 to 12,913. For CCS data, the numbers of DeBreak deletions overlapping with PAV ranged from 7,076 to 9,040, while insertions ranged from 10,907 to 13,181. DeBreak also identified unique calls. For CLR data, the numbers of unique DeBreak deletions ranged from 991 to 1,495, while unique insertions ranged from 2,385 to 3,560. For CCS data, the numbers of unique DeBreak deletions ranged from 1,307 to 2,057, while unique insertions ranged from 1,691 to 3,501.

12 Illumina variant calling

12.1 Alignment of Illumina reads and SNV/indel calling

(*Contributors: Uday S. Evani and Marta Byrska-Bishop*)

Alignment of reads to the reference genome GRCh38, duplicate marking, and Base Quality Score Recalibration (BQSR) were performed as described in the functional equivalence pipeline standard (*121*) developed for the Centers for Common Disease Genomics (CCDG) project (*6*). SNV and indel calling were performed using the GATK version 3.5 (*122*). Briefly, the HaplotypeCaller tool (*123*) was used for variant discovery in the GVCF mode, with sex-aware ploidy settings applied to chromosomes X and Y. Joint genotyping across the 3,202 samples was performed using GenotypeGVCFs, followed by Variant Quality Score Recalibration (VQSR) with truth sensitivity level of 99.8% for SNVs and 99.0% for indels.

12.2 Haplotype phasing of SNVs and indels in the Illumina data

(*Contributors: Anna O. Basile and Marta Byrska-Bishop*)

Haplotype phasing of SNVs and indels on autosomes was performed using statistical phasing with pedigree-based correction using SHAPEIT2-duohmm (*124*, *125*). Phasing was limited to non-singleton, high-quality sites that meet the following QC criteria: 1) VQSR PASS; 2) GT missingness <5%; 3) Hardy-Weinberg equilibrium p-value > 1e-10 (HWE exact test) in at least one of the five superpopulations; and 4) Mendelian error rate <5%. Prior to phasing, the multiallelic sites were split into separate lines and indels were left-aligned and normalized using BCFtools norm *(118)*. Phasing of variants on chromosome X was performed using statistical phasing, as implemented in the Eagle2 software (*126*), which supports heterogeneity in genotype ploidy.

12.3 SV discovery from individual algorithms

(Contributors: see following sections)

SV discovery from Illumina short-read whole-genome sequences (srWGS) on the 34 samples in this study was done in two batches: 15 samples (HG00096, HG00171, HG00513, HG00731, HG00732, HG00864, HG01596, HG03009, NA12878, NA18534, NA18939, NA19238, NA19239, NA20509, NA20847) were processed in the first batch that included 2,504 genomes, which we refer to as 'batch 1' in the following text. The remaining 19 samples (HG00512, HG00514, HG00733, HG01114, HG01505, HG02011, HG02492, HG02587, HG02818, HG03065, HG03125, HG03371, HG03486, HG03683, HG03732, NA12329, NA19240, NA19650, NA19983) were processed together with another 679 samples in 'batch 2'. SV discovery methods were mostly the same on the two batches with slight differences in technical details.

12.3.1 Manta

(*Contributors: Allison Regier and Xuefang Zhao*)

The Manta VCFs for batch 1 samples were produced in a docker container with the image halllab/manta_samtools@sha256:6c8dfccfd3124ebf902ac6f0303e6f09b02a15e2c09963354620740788c407d0 using Manta v1.4.0 (*127*) with default parameters.

Manta V1.5.0 (*127*) were applied to batch 2 samples in single-sample mode on Terra (previously FireCloud) with default parameters. SVs were only discovered from canonical chromosomes (chromosome 1 to 22, X and Y) and mitochondrial DNA.

12.3.2 Wham

(*Contributors: William Harvey and Xuefang Zhao*)

Wham v1.7 (*128*) was applied to the 15 samples in batch 1 in single-sample mode with default parameters and a mapping quality filter of 15. All SVs were required to be called as a 'PASS' by whamg genotyping and have a confidence of variant mapping >0.2. Additionally, DELs and INVs

smaller than 50 bp and larger than 1 Mbp were removed. For DUPs, the size threshold was 300 bp and 1 Mbp, respectively. Finally, calls inside of SDs, tandem repeats, and centromere/peri-centromere regions were removed.

Wham v1.7 (*128*) was applied to batch 2 samples in single-sample mode on Terra with default parameters. A customized whitelist that excluded N-masked genomic regions to keep the computing cost under reasonable scale (a copy of the whitelist file “wham_whitelist.bed” is available at the HGSVC2 data portal).

12.3.3 MELT

(*Contributor: Scott Devine*)

MELT v. 2.1.5 (*33*) was used to discover MEIs (Alu, L1, SVA, HERV-K) from all samples in both batches using the MELT_Split mode. Trace size (-t) was set to 150 bp, coverage (-c) was set to 30X, and human reference genome GRCh38 was used as the reference in this analysis.

12.3.4 LUMPY

(*Contributor: Allison Regier*)

LUMPY (v0.2.13) (*129*) was run on all batch 1 and 2 samples using the Smoove wrapper (v0.2.3) in a docker container with the image brentp/smoove@sha256:c839ed223462a1c1ae26e7acc27f28f0f67b4581d80a06823895f295ad2bdaf4. The command `smoove call` was run with the `--noextrafilters` and `--genotype` parameters set. The `--exclude` parameter was set to blacklist regions in the file “exclude.cnvnator_100bp.GRCh38.20170403.bed” (a copy of the blacklist file is available at the HGSVC2 data portal).

12.3.5 CNVnator

(*Contributor: Allison Regier*)

CNVnator (v0.3.3) (*130*) was run on all batch 1 samples using the wrapper script cnvnator_wrapper.py contained in the docker image halllab/cnvnator@sha256:8bf4fa64a288c5647a9a6b1ea90d14e76f48a3e16c5bf98c63419bb7d81c8938. Depth histograms were generated with a window size of 100 bp.

12.3.6 DELLY

(*Contributor: Tobias Rausch*)

DELLY v0.8.3 (*131*) was used to discover SVs (deletions, insertions, tandem duplications, and inversions) in all 3,202 samples with the high-coverage, short-read data mapped to GRCh38. To generate a population callset, we first discovered SVs separately in each sample, then merged the detected SV sites using the DELLY merge subcommand and finally re-genotyped the merged

SV site list in all 3,202 samples. Using the DELLY filter subcommand in germline mode, we successfully genotyped 110,153 nonredundant SVs across the 3,202 samples.

12.4 SV integration

12.4.1 FusorSV

(*Contributor: Qihui Zhu*)

Existing SV detection algorithms are typically optimized for the discovery of specific types of SVs, thus their performance varies by type, size, and genomic location of SVs. Therefore, it is impossible to achieve comprehensive discovery over the entire SV spectrum with a single algorithm. To mitigate this issue, we applied FusorSV (*132*), a unique data-mining method, to integrate SV calls from six different algorithms including CNVnator (*130*), DELLY (*131*), Manta (*127*), LUMPY (*129*), Wham (*128*) and MELT (*33*). Tool versions were described above in the section 12.3.

FusorSV takes SV calls from different algorithms and integrates them in a manner that minimizes false positives and maximizes discovery. FusorSV includes two phases: the training and the discovery phase. In the training phase, we use the Illumina SV calls from NA19238, NA19239, HG00731, HG00732 and HG00513 reported by Chaisson et al. (*1*) as the ground truth to train a FusorSV model. The performance of each of the six algorithms was compared to the ground truth, in addition to calculating the pairwise performances across all algorithms. Combinations of algorithms that work more comprehensively will be promoted. Then the score for every possible combination of algorithms will be calculated and the performance value will be determined to a FusorSV fusion model (*132*). In the discovery phase, the FusorSV fusion model was used for comprehensive SV detection in all 33 test samples. This fusion model integrated individual algorithm performances with pairwise similarity of algorithms across SV types, which allowed FusorSV to select subsets of algorithms for each SV type that are more comprehensive, with maximum discovery and minimal subset false positives.

FusorSV provides a unique data-mining method that intelligently takes SV calls from different algorithms and combines them in a manner that minimizes false positives and maximizes discovery. Using per-algorithm performance information and similarity between algorithms, the smallest set of SV callers can be selected using the concept of mutual exclusion, which makes our method both more accurate and comprehensive than other approaches merging SV calls based on consensus or other heuristics.

12.4.2 GATK-SV

(*Contributor: Xuefang Zhao*)

GATK-SV is a multi-module SV discovery and refinement pipeline for srWGS data. This method was previously adopted by the Genome Aggregation Database (gnomAD) for SV discovery, and the technical details have been described by Collins et al. (*5*). In this study, GATK-SV was applied

to all 3,202 samples from both batches for SV discovery, genotyping, complex resolution, and final refinement. Moreover, we applied a machine-learning method (*133*) to further refine the callset. Most parts of the method were the same as in Collins et al. (*134*), while processes that were unique to this study are described here. An overview of this method is represented in Fig. S51-A.

Sample batching. The samples were clustered into batches of ~200 samples for raw algorithm processing, then merged for complex resolution and filtering. The batching strategy depends on the gender, dosage bias score (*∂*), and median read coverage of each sample. *∂* and median read coverage were calculated to determine genomes with an unusual distribution of reads across segments of the genome using the same method as in Collins et al. (*5*). Samples were first separated by gender to form groups of 1,599 male samples and 1,603 female samples. Each group was then divided into four subgroups by the rank of median coverage, and each subgroup was further split into four groups by the rank of *∂*. The resulting 32 batches of approximately 100 samples of unique gender were matched by their rank of median coverage and *∂* to form 16 batches of 200 samples with equal numbers of male and female samples. A brief summary of the batching schema is represented in Fig. S51-B.

SV discovery from raw algorithms. SVs discovered from Manta (*127*), Wham (*128*), MELT (*33*), cn.MOPS (*135*) and GATK-gCNV were integrated in this pipeline. Tool versions and run settings of Manta, Wham and MELT have been described above. cn.MOPS was executed in a custom implementation on Terra for all samples, in ~200-sample batches. For each batch, the read depth (RD) per 100 bp bin across each genome was calculated and merged to form an RD metric. We composed RD matrices across all samples at 300 bp and 1 kbp resolution, excluding any samples with a median bin coverage of zero per contig, then ran cn.MOPS with R v3.3.3 (*136*), split raw calls per sample, segregated calls into deletions (copy number < 2) and duplications (copy number > 2), merged the 300 bp and 1 kbp resolution calls per sample per CNV type using BEDTools (*96*) merge, and subtracted any N-masked bases from all CNV calls using BEDTools subtract. GATK-gCNV was implemented on Terra and applied to all samples in each batch with default parameters (see (*137*) for full details).

Integration, quality refinement and re-genotyping of SVs. Module 01-05 of GATK-SV has been described in Collins et al. (*5*); please refer to supplementary pages 45-50 for technical details.

SV refinement. This refinement module corrects false positive variants caused by the ambiguous alignment of short reads. The refinements were applied in two layers: per-sample level and per-SV-site level. For SVs genotyped as non-reference in each sample, multiple features were collected, including averaged RD of each SV region, averaged RD of the 1 kbp flanking region of each SV, count of aberrant pair ends within 150 bp of each SV, count of split reads within 100 bp of each breakpoints, variant type, allele fraction across all samples, count of children that carry this SV as *de novo*, genomic location of the SV (split into simple repeats, SDs, and the rest relatively unique genomic sequences), genotype prediction and quality from previous modules, and whether they were discovered by the original algorithms. These features were integrated into a lightGBM model using these settings: *objective = "binary", metric = "auc", num_class = 1L, learning_rate = 0.1, min_data_in_leaf = 1L, min_sum_hessian_in_leaf = 1.0*. SVs in three

samples, i.e., HG00514, HG00733 and NA19240, which were overlapped by PacBio SVs reported by Chaisson et al. (*1*) and validated by VaPoR (*138*) on the high-coverage PacBio sequences were selected as a truth set, while SVs that neither have overlap in PacBio nor supported by VaPoR were considered false positives. A unique score, referred to as 'boost score' (BS) in this manuscript, would be assigned to each SV by the lightGBM model. To keep the false discovery rate (FDR) of SVs under 5%, we considered SVs with BS lower than -0.404 as low quality and removed from the callset.

For each SV site, a boost model ratio (BR) is defined as the proportion of samples that passed the lightGBM model among all samples that had non-reference genotypes and used to infer quality of the SV locus. To keep the *de novo* rate of SVs under 5% in the 602 trios, which represents a combination of false positives in children and false negatives in parents, we set BR at 0.509 and labelled SVs with lower BR as low-quality loci.

It is noticed that VaPoR has decreased power when evaluating large CNVs located within SDs, so we also applied the VCF refinement method described by Collins et al. (*5*) (pages 50-51 of the supplemental methods) and included the large CNVs (>5 kbp) that passed this method into the final GATK-SV callset.

Comparison of SVs between short-read Illumina sequences and long-read PacBio sequences. SVs detected from srWGS were compared against SVs from long-read PacBio sequencing with matching samples on the matched sample (Table S30). SVs were considered concordant if they met these following criteria:

1. For insertions: a pair of SV loci were considered concordant if their reported insertion points were within 100 bp from each other, and the length of their inserted sequences was within 10 times the size of each other.

2. For deletions, duplications, CNVs, inversions, and complex SVs >5 kbp in size, 50% RO and match of variant types were required to be considered concordant; while for variants between 50 bp and 5 kbp, 10% RO and match of variant types were required. Complex SVs were considered concordant if overlapped by inversions, and CNVs were considered concordant if they overlap either deletions or duplications.

12.4.3 Absinthe

(*Contributor: André Corvelo*)

On a per-sample basis, insertions with a minimum length of 100 bp were discovered through *de novo* assembly of unmapped and discordant read pairs using Absinthe (*88*), and then genotyped using Paragraph v2.4a (*139*), with the maximum allowed read count for a variant (-M) set to 20 times the mean of the sample depth and respecting sex-specific ploidies. Insertion calls from all 3,202 samples that were positively genotyped with a PASS filter flag were then clustered by genomic location and aligned using MAFFT v7.407 (*140*). For each locus, the most consensual allele was selected. Variants from the resulting merged callset were then re-genotyped using Paragraph v2.4a (*139*) on all 3,202 individuals. To produce the final callset only variants with 1) genotyping PASS filter rate ≥ 80%; 2) Mendelian Error Rate ≤ 5% for complete trio calls; and 3) HWE p-value > 10^{-6} (chi-squared test) in at least one of the five superpopulations were kept.

12.4.4 svtools

(*Contributors: Haley Abel and Allison Regier*)

The svtools (*141*) software toolkit and workflow are designed for generation of large-scale SV callsets. The workflow combines per-sample variant discovery with LUMPY (49) and Manta (46) with resolution-aware cross-sample merging. The set of merged variants is then genotyped with svtyper, followed by copy number annotation with CNVnator (50) and reclassification of variants based on concordance of read depth with breakpoint orientation. All parameter settings and versions are as implemented in the wdl-based workflow (https://github.com/hall-lab/sv-pipeline, commit c49150e7d6c7acb01b099ee2f7f8def0c170a997, (*88*)).

13 Bionano Genomics discovery

(*Contributors: Alex Hastie, Joyce Lee, Feyza Yilmaz*)

13.1 Bionano Genomics de novo assembly and structural variant calling

De novo assemblies of the samples were obtained using the Bionano's *De Novo* Assembly Pipeline (Bionano Solve v3.5) with haplotype-aware arguments (optArguments_haplotype_DLE1_saphyr_human_downSampleLongestMole.xml; a copy of the configuration file is available at the HGSVC2 data portal). With the Overlap-Layout-Consensus paradigm, pairwise comparison of DNA molecules, which are at least 250 kbp in length and contribute to a coverage of 250X, was generated to create a layout overlap graph and produce initial consensus genome maps. By realigning molecules to the genome maps (alignment confidence cutoff of Bionano p-value < 10^{-12}; Bionano p-value derived from methodology described in (*142*)) and by using only the best match molecules, we applied a refinement step to label positions on the genome maps and to remove chimeric joins. Next, during an extension step, the software aligned molecules to genome maps (Bionano p-value < 10^{-12}) and extended the maps based on the molecules aligning past the map ends. Overlapping genome maps were then merged (Bionano p-value < 10^{-16}). These extension and merge steps were repeated five times before a final refinement was applied to "finish" all genome maps.

To identify all alleles, clusters of molecules that are aligned to genome maps with unaligned ends >30 kbp in the extension step were re-assembled to identify potential alternate alleles. To identify alternate alleles with smaller size differences from the assembled allele, we also searched for clusters of molecules that aligned to genome maps with internal alignment gaps of size <50 kbp, in which case, the genome maps were converted into two haplotype maps. To improve accuracy, large repetitive regions, such as SDs, are identified in a *de novo* fashion during the last round of merging. Maps sharing more than 140 kbp in common but diverging on both sides of the non-unique region generate splits within alignment maps. Molecule clustering and SD filtering generate more accurate diploid assemblies. We called and annotated SVs using Bionano's *De Novo* Assembly Pipeline (Bionano Solve v3.5), in which the final genome maps were aligned (Bionano p-value <10^{-12}) to the reference genome (GRCh38).

13.2 Bionano Genomics discovery of large, complex structural variants

Using Bionano genome maps, we identified 19,821 insertions and deletions that are at least 5 kbp in size mapping to high-confidence regions where there were no more than two aligned genome maps at breakpoints (Table S26). We clustered calls based on 80% size concordance and 80% reciprocal overlap to generate a nonredundant set of 2,017 insertion and 1,978 deletion clusters (Table S43), of which 516 insertion and 331 deletion clusters were localized to regions where we identified at least five different SVs contributed by at least seven samples. About 75% of these multi-site complex polymorphic variants overlapped SDs, and 2-4% of them overlapped gaps in the current human reference genome suggesting that these have been particularly problematic during sequence and assembly. We plotted these large complex SV sites as well as the underlying set of coordinates in GRCh38 (Fig. S52). In addition to insertions and deletions, we also identified 162 nonredundant inversions and 192 duplications across all samples, of which 53% and 58% of them overlapped SDs, respectively.

We compared Bionano Genomics SV calls to those discovered by PAV, requiring 50% concordance with respect to size and breakpoints within 1-5 kbp (Fig. S53). With such criteria, 72% of the 19,821 Bionano calls overlapped SVs detected by the phased assembly of PacBio data (Table S26). The comparison revealed 3,453 Bionano unique insertions and 2,137 Bionano unique deletions, corresponding to 1,657 clusters (Table S27). A cluster might have PacBio overlapped calls in one sample but not in other samples, and we found that there were 697 insertion clusters and 478 deletion clusters (Table S28) that had never been found in any of the PacBio assembly based callsets. Bionano-unique deletion and insertion clusters showed substantial overlap with genomic regions that are potentially difficult to assemble or to align to, e.g., SDs or GRC-defined reference issues (Section 8.2, Fig. S54), as well as genes (Table S44).

We highlight several examples of complex structural polymorphisms corresponding to gene-rich regions that are not yet fully resolved at the sequence level within our phased assemblies. For example, we identified a 150 kbp region enriched in SDs mapping to chromosome 1 (chr1: 108.3–108.45 Mbp), which overlaps *NBPF5P* and *NBPF6*. We also identified a 75 kbp inversion in 24 out of the 30 samples, and a 74 kbp deletion was found in the alternate allele in two of the samples. The same deletion was also found in three samples that do not have the inversion allele, and another 74 kbp deletion and 130 kbp inversion were found in two different samples (Fig. S55). In addition, we identified duplications of up to 10 copies on chromosome 5 (21.1–21.7 Mbp), which overlaps with *GUSBP1* (Fig. S56). Each one of these complex regions were characterized and manually evaluated by checking single-molecule support for each sample. Critical labels, which were located in unique regions, were identified to detect the number of molecules spanning regions of interest.

Similarly, we identified 18 haplotypes at 3q29 (chr3:195.6-196Mbp), which overlaps *SDHAP2* genic region (Fig. S57, Table S45), by manual curation. Single molecules that were anchored to unique regions confirmed each haplotype (QC data [optical mapping molecule support] and haplotype segment coordinates are available at the HGSVC2 data portal). Haplotype 1 (H1) is the most common with 12 haplotypes in five superpopulations. GRCh38 reference assembly

haplotypes is the third most common haplotype observed in all populations except for Admixed American. A summary of the assembly contig coverage for 3q29 can be found in Table S46.

14 Strand-seq Inversion detection and genotyping

(*Contributors: Ashley Sanders, David Porubsky, Wolfram Höps, Hufsah Ashraf, Maryam Ghareghani, Tobias Marschall, Jan Korbel*)

To detect inversions using Strand-seq data, directional composite files were generated for each sample as previously described (*1*, *9*). To automate composite file generation, a merging protocol was implemented using the breakpointR 'synchronizeReadDir' function (*143*), which locates Watson-Watson (WW) and Crick-Crick (CC) regions in each chromosome and for each cell before building these into the sample-specific composite files (Fig. S58). Segmental changes in composite file orientation, suggestive of an inverted allele, were identified using breakpointR. To detect both larger and smaller strand-state changes, we used breakpointR in two settings—applying either a window size length of 5 kbp or 20 reads per bin. In both cases, we scaled an initial bin size by multiples of 2, 3, 4, 5, 10 and 20. This resulted in a redundant dataset with putative inversions detected per sample.

To construct a nonredundant set of Strand-seq inversions, we merged and filtered all detected strand-state changes (putative inversions) in multiple stages as follows: First, we cropped regions that overlap with highly identical SDs (≥98% identity) or gaps defined in GRCh38 from each inversion breakpoint. Second, we iteratively merged ranges with ≥50% RO. Such collapsed ranges were then subjected to regenotyping using 'genotypeRegions' of the primatR package (*38*). Each region in each sample was assigned a genotype: 'HET' - approximately equal mixture of plus and minus reads, 'HOM' - majority of minus reads, 'REF' - majority of plus (reference) reads and 'lowReads' - less than 20 reads in a region. Ranges that genotype only as a reference ('REF') orientation or has less than 20 ('lowReads') reads across all samples were filtered out. Next, we collapsed ranges that share the same genotype across all samples and are embedded with respect to one another. Lastly, from regions that genotyped only as a 'HET' or 'lowReads', we retained only those that do not overlap with regions where Strand-seq inversion call was already made. The same procedure was repeated for windows defined by the readcount (20 reads per bin), which allowed adding smaller inversions missed by the larger bin size to the final Strand-seq inversion callset.

Results for inversion detection and genotyping. The strategies for inversion discovery described in the previous sections operate on the basis of individual samples, leaving us with the need to unify these inversion calls across samples. Additionally, we saw an opportunity to integrate information about inversion loci across samples to further improve the individual callsets.

To this end, we developed a new method, ArbiGent, which determines inversion genotype likelihoods for chosen genomic loci of 5 kbp or larger in size based on Strand-seq data. We built ArbiGent utilizing a statistical framework previously devised for discovering subclonal structural variation in cancer (*95*), which we extended to allow estimating germline SV genotype likelihoods

for DNA segments of choice using strand-specific reads. In particular, on the basis of a Bayesian probability framework that models strand- and haplotype-specific read counts using Negative Binomial distributions, ArbiGent computes inversion genotype likelihoods for inversions and copy number changes. SV genotype likelihoods derived from individual cells from the same sample are concatenated by summing up log-likelihoods across cells, to result in a combined genotype likelihood estimate per sample and genomic locus of interest. The code of ArbiGent is available open source at https://github.com/friendsofstrandseq/pipeline/tree/arbitrary-segments (project version released via (*88*)).

Utilizing inversion discovery techniques based on Strand-seq, Bionano, and aligned long reads, we found 394 genomic loci with evidence for an inversion across the 64 haplotypes studied (see "Nonredundant callsets, filtering, and properties" below). We used ArbiGent to assign genotype likelihoods to these loci across all samples and filtered them for high-quality sites, which led to the removal of 60 sites for which ArbiGent did not confirm any inverted haplotype across the samples considered ('suspected false positive'). We additionally tested the inversion genotypes for Mendelian consistency, utilizing trio-based Strand-seq data previously generated in Chaisson et al. (*1*), which resulted in another eight rejected inversion sites. Sites that were consistently predicted as complex events (6) and those for which short-read mappability was low (71) were flagged, but kept in the final callset. Accordingly, we report 316 high-quality inversion sites with genotypes, with a median of 130 detected inversion events (54 in HOM state, 54 HET and 23 complex events) comprising 24.8 Mbp of inverted DNA per diploid human genome.

Analysis of 16p12 regions. Detailed analysis of 16p12 regions is based only on Strand-seq inversion calls detected based on the distribution of direct and inverted Strand-seq reads in composite files as stated above (Fig. S25A). Each inversion was phased by separating Strand-seq reads per haplotype as reported before (*38*). Inversions that fail to reliably phase in any given sample (due to a low read coverage) were considered as a reference orientation. Next, we enumerated the frequency of unique combinations of inverted alleles along the 16p12 region per superpopulation. We evaluated the genetic background of each inversion by exploring SNVs that lie ±500 kbp from each inversion breakpoint. SNVs from within the inversion and those that overlap with SDs (≥98% identity) were removed from the analysis. For each inversion, we constructed a neighbor-joining tree based on the flanking SNVs. Inversions that occurred at seemingly different genetic backgrounds (flanking SNVs) were considered as likely recurrent or toggling (Fig. S26). Last, we phased genome assembly alignments to GRCh38 of 16p12 regions have been analyzed. We report only phased contig alignments with mapping quality 60. We summarized alignment gaps in respect to GRCh38 as a coverage of all gaps (Fig. S25-B).

15 MEI discovery and integration

(*Contributors: Scott Devine, Nelson Chuang, Weichen Zhou, Ryan Mills, Bernardo Rodriguez-Martin, Martin Santamarina*)

Mobile element insertions (MEIs), including long interspersed element-1 (L1), Alu, and SVA (SINE-VNTR-Alu) retrotransposons, comprise approximately 46% of the human genome and

represent ~25% of all SVs in human genomes (*144*, *145*). They have been shown to play an important role in human development, population diversity, and genomic disease (*29*, *30*, *146*–*148*). Various strategies have been developed to identify candidate polymorphic MEIs from srWGS data (*23*, *33*), though they struggle in regions where reads can map equally to multiple alternative genomic positions. Long-read sequencing technology provides a better resolution in such regions by directly sequencing long stretches of contiguous DNA that enable the discovery of potentially overlooked MEIs (*1*, *4*, *75*, *149*). By leveraging the PacBio assemblies, especially from HiFi reads, we are able to discover the largest set of non-reference MEIs with *bona fide* characteristics among diverse samples, allowing the investigations of their impacts on shaping human genome and population diversity.

MEIs were primarily discovered using PAV, which were then annotated using MEIGA. This procedure led to the identification of 9,453 fully sequence-resolved non-reference MEIs, including: 7,738 Alus, 1,175 L1s, and 540 SVAs. In addition, Illumina and PacBio alignments were processed using MELT and PALMER, respectively, in order to increase sensitivity for MEI discovery. Overall, MELT reported 6,935 Alus, 977 L1Hs, 502 SVAs, and 32 HERV-Ks; and PALMER reported 7,607 Alus, 1,226 L1Hs, and 478 SVAs among diverse samples (Fig. S59-A). This procedure recovered an additional 1,932 putative MEIs (although not all were fully sequence resolved). Afterwards, we applied an integration strategy across these different platforms to obtain a better comprehensive landscape of MEIs among the diverse populations (see “MEI integration”).

15.1 Callsets across different platforms

15.1.1 MELT

(*Contributors: Scott Devine and Nelson Chuang*)

MEIs were discovered in the same 33 genomes that were sequenced by PacBio sequencing in this study. High-coverage (~30-40X) Illumina WGS were obtained from the 1000GP bucket at Amazon. MEIs (Alu, L1, SVA, and HERV-K elements) were discovered using the mobile element locator tool (MELT, ver. 2.1.5 (*23*, *33*)) using MELT-SPLIT, which is a stepwise version of MELT that uses discordant read pairs and split reads to perform MEI discovery across all samples in a given population. This improves the modeling at each MEI site, and improves genotyping, compared to analyzing the samples as individuals. We only included calls in the final VCF that were classified as “PASS” and “lc” (low complexity), and all calls with alternative filters were removed. Build GRCh38 of the reference human genome was used to perform MEI discovery. 3' transduction analysis was conducted using the 3' transduction finder in MELT (*33*). The MELT software package, along with documentation, can be downloaded from: https://melt.igs.umaryland.edu/.

15.1.2 PALMER

(*Contributors: Weichen Zhou and Ryan Mills*)

We developed an enhanced version of PALMER (Pre-mAsking Long reads for Mobile Element inseRtion, (*75*)) to detect MEIs across the long-read-sequenced genomes (). Reference-aligned BAM files from long-read technology are used as input. Known reference repetitive sequences (L1s, Alus or SVAs in reference) are used to pre-mask the portions of individual reads that align to these repeats. After the pre-masking process, PALMER searches subreads against a library of mobile element sequences within the remaining unmasked sequences and identifies reads with a putative insertion sequence (including 5′ inverted L1 sequence, if available) as candidate supporting reads. PALMER opens the bins in 5′ upstream and 3′ downstream of insertion sequence for each read and then identifies candidate TSD motifs, transductions, and poly(A) tract sequence. All supporting reads are then clustered at each locus and those with a minimum number of supporting events are reported as putative insertions. To improve the accuracy of non-reference MEI sequences derived from individual subreads, which have lower per-read base-pair accuracy, we used local sequence alignments and error-correction strategies. Error correction was conducted by applying CANU v1.8 (*150*) to subreads that contain the PALMER MEI sequence, allowing the generation of error-corrected reads that served as inputs for local realignment using minimap2. A second pass of the PALMER pipeline was then executed using these locally aligned error-corrected reads to generate a high-confidence callset of germline non-reference MEIs. Eventually, we used CAP3 to assemble all the MEI sequences reported by the second pass of the PALMER pipeline and obtained a high-confidence consensus contig for each non-reference MEI event.

15.1.3 MEIGA-PAV

(*Contributors: Bernardo Rodriguez-Martin and Martin Santamarina*)

Phased assembly SV calls (PAV), in the form of a multisample VCF, were processed using a custom pipeline derived from MEIGA algorithm (available at https://github.com/MEIGA-tk/MEIGA-PAV and project version released via (*89*)) to detect non-reference Alu, L1 and SVA insertions. First, candidate retrotransposition events were identified by searching for poly(T) and poly(A) tails at the 5' and 3' ends of assembled inserted sequences for PAV insertion calls. Poly(A/T) tails were required to be at least 10 bp in size, have a minimum purity of 80%, and be at a maximum distance of 30 bp to the insert end. Then, inserted sequences for candidates were realigned using minimap2 (*119*) v2.10 into a database of consensus L1, Alu, SVA and ERV sequences. Sequence alignments are chained based on complementarity in order to identify the minimum set of nonoverlapping alignments that contribute to resolve the maximum percentage of the inserted sequence. Based on the alignment chain, MEIs are called and multiple insertion features are inferred, which includes orientation, insertion length, 5' inversion, and truncation lengths. Unresolved sequence 3' ends are interrogated for Poly(A/T) tracts using the same criteria described above. MEI calls harboring a single tract are classified as “solo” insertions, while events with multiple tracts are considered potential 3' transduction events with the transduced sequences in between. Candidate 3' transduced sequences together with unresolved 5' ends for each MEI

are aligned into the reference genome using BWA-mem 0.7.17 (*119*, *151*) to search for 5' and 3' partnered transductions. 5' and 3' transductions calls are made if at least 75% of the target sequence aligns on the reference with a mapping quality over 30, respectively. Candidate “solo” and partnered transductions with less than 60% of their sequence resolved are filtered out to generate a high-quality set of non-reference insertions. Subfamilies for L1, Alu and SVA inserts were inferred using two distinct strategies. For L1 events, subfamily assignment was performed through the identification of subfamily diagnostic nucleotide positions on their 3' end (*38*, *152*). L1 integrations bearing the diagnostic “ACG” or “ACA” triplet at 5,929-5,931 position were classified as “pre-Ta” and “Ta”, respectively. Ta elements were subclassified into “Ta-0” or “Ta-1” according to diagnostic bases at 5,535 and 5,538 positions (Ta-0: G and C; Ta-1: T and G). Elements that did not display any of these diagnostic profiles could not be assigned to a particular category, and their subfamily status remained undetermined. For Alu and SVA, the assembled inserts were processed with RepeatMasker v4.0.7 to determine the subfamily. If multiple RepeatMasker hits were obtained, the one with the highest Smith-Waterman score was selected as representative. In addition, the coding potential for each L1 sequence was assessed in every of the six potential reading frames. Codon equivalencies followed the Standard Genetic Code currently adopted by NCBI, considering AUG, UUG and CUG as starting codons, and UAA, UAG and UGA as terminators (*153–155*). Every candidate open reading frame (ORF) identified through this strategy was compared with reference L1 ORF1 and L1 ORF2 in terms of 1) length and 2) sequence identity, before being catalogued as one of the two main components of the L1 coding system (*156*). Reference amino acid sequences encoded by L1Hs ORF1 and ORF2 proteins were obtained from Uniprot (*156*, *157*). To end, a multi-sample VCF containing MEI calls together with all collected pieces of annotation was generated as output.

15.2 MEI integration and PAV annotation

(*Contributors: Weichen Zhou and Ryan Mills*)

To gain a more comprehensive landscape of MEIs in human genomes and better understand the assembly-based method (PAV) on MEI calling, we carried out a integration analysis for MEIs across different platforms/pipelines, including: MELT (independent caller for Illumina), PALMER (independent caller for PacBio mapping-based), and MEIGA_PAV (pre-filtered PAV assembly-based callset with MEIGA annotation).

MEI calls were pre-merged at the caller level before MEI integration. We then stratified the information of MEIs from four callers into three tiers: Tier 1, the MEI type/family, the insertion orientation/strand, the insertion site, the caller; Tier 2, the insertion length, the structures (e.g., 5' inverted sequence, 3' transduction, intact ORFs), the consensus contig; and Tier 3, detailed information at every single discovery sample for each call (e.g., VNTR variation in SVAs). We first merged the Tier 1 information with the exact MEI family, insertion strand, and an overlap of insertion sites in bin sizes of ±50 bp. The CIPOS (Confidence Interval POSition) information is provided showing the difference between the original insertion sites from callers and the merged site. Callers and the number of callers were also recorded. The consensus contigs in Tier 2 are integrated by priorities as follows: Assembly HiFi, Assembly CLR, CLR with local assembled

strategy. Structures/characteristics of MEIs are then identified based on the consensus contigs. Tier 3 information was not included in the merged VCF but kept in the sample-level VCFs.

In the integrated callset, there are 11,882 MEIs, including 9,516 Alus, 1,646 L1Hs, 688 SVAs, and 32 HERV-Ks from three different methods. 5,499 (57.8%) Alus are called by at least three callers, 1,889 (19.9%) are called by two callers, 2,128 (22.4%) are called by one single caller, and 1,215 out of 2,128 are singleton events. 653 (39.7%) L1Hs are called by three callers, 445 (27.0%) are by two callers, 548 (33.3%) are only called by one, and 308 out of 548 L1Hs are singletons. There are 337 (49.0%) SVAs called by three callers, 174 (25.3%) called by two callers, 177 (25.7%) SVAs are called by one, and 95 out of 178 SVAs are singletons (Fig. S59-A).

To annotate the final PAV callset, the similar strategies (for the Tier 1 and partial Tier 2) were used to compare with the prior integrated callset. Eventually, 9,950 non-reference MEIs, including 8,110 Alus, 1,248 L1Hs, 589 SVAs, and 3 HERV-Ks were annotated in the final PAV PacBio assembly-based callset. Considering the intersection with two independent callers (MELT and PALMER), there are 6,615 (66.5%) calls that are supported by two callers, including 5,587 Alus, 671 L1Hs, and 357 SVAs; and 2,424 (24.4%) calls are supported by one caller, including 1,843 Alus, 422 L1Hs, 156 SVAs, and 3 HERV-Ks. Overall, 90.8% of the PAV calls have evidence from at least one or more of independent callers in the prior MEI integrated callset (Fig. S59-B).

15.3 Characteristics of MEIs and analysis

15.3.1 Phylogenetic analysis and age estimation for active sequence-resolved L1s

(*Contributor: Martin Santamarina*)

DNA sequences for 143 human-specific (L1Hs) and L1Pt [GenBank: KF661301.1] full-length L1 elements were aligned using MUSCLE (*158*) v3.8 with default number of iterations. Manual inspection of the multiple alignment was performed with Jalview (*159*) v2.11 in order to remove short upstream and downstream spurious sequences. L1 phylogenies were built using the following R packages: Ape (*160*) v5.4 and Phangorn (*160*, *161*) v2.5. Nucleotide substitution model selection was performed with phangorn::modelTest. GTR + G + I was set as the best model for tree inference according to both AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) scores. Maximum likelihood tree was estimated with phangorn::optim.pml function, starting from an initial distance-based tree. Tree topology was adjusted via nearest neighbor interchange (*162*). After maximum likelihood optimization, a midpoint rooting strategy was followed, rendering L1Pt as the expected tree outgroup. Non-parametric bootstrap analysis with 1000 replicates was done in order to determine the node consistency along the inferred tree. For display purposes, only values informative of high support (≥80%) were labeled to the tree nodes.

The average of L1 elements could be estimated by measuring the level of sequence divergence from each subfamily consensus sequence (*163*, *164*). We used DECIPHER v2.14 to generate a consensus sequence for each L1Hs subfamily (pre-Ta, Ta1 and Ta0) from our set of active elements. For each element, we identified the number of nucleotide substitutions acquired since

the ancestral state represented by this consensus. Using an L1-specific mutation rate of 0.25% per million years (*152*), and assuming constant and neutral rate of evolution, we obtained rough estimates about the age of individual L1 elements.

15.3.2 VNTR distributions in SVAs

(*Contributors: Weichen Zhou and Ryan Mills*)

An SVA is a composite noncoding retrotransposon, which likely uses the L1 machinery for its mobilization into human genomes (*165*). It consists of a hexameric CCCTCT repeat on the 5' end, followed by two antisense Alu-like fragments, a variable number of GC-rich tandem repeats (VNTR), a SINE-R sequence, a canonical polyadenylation signal AATAAA, and a poly(A) tract (*165*). Few studies have the privilege to investigate the variation of VNTR regions in SVAs before the advent of long-read sequencing technologies, given it is a high-repetitive region containing multiple copies of a 35-50 bp repeat (*166*).

To define the copy numbers of VNTRs in both reference and non-reference SVAs, at the individual level, we used the sequences directly from the assemblies of discovery samples (27 CLR samples and 11 HiFi samples). We are able to resolve their expansions (duplications and deletions) by unequal homologous recombination at individual and population level as well. The locations of reference SVAs are obtained from RepeatMasker track from UCSC, and the insertion sites of non-reference SVAs are based on PacBio assembly callset (PAV). All SVA sequences were obtained from the consensus contig of each individual assembly, stratified into subfamilies (SVA_F, SVA_E, SVA_D, etc.) by using the diagnostic nucleotides, and then divided into different populations to conduct the population level analysis. The sample frequency of SVAs is calculated based on the number of discovery samples for each event.

We eventually obtained the assembly sequences that can be used for downstream analysis for 271 reference SVAs and 435 non-reference SVAs among the discovery samples (27 CLR samples and 11 HiFi samples). In the Fig. S60, we observed a) increased variable length of VNTRs in non-reference SVAs as compared to reference SVAs (p-value < 1x10-5, student's t-test, two-sided), and b) reference SVAs with a higher sample frequency than non-reference SVAs (89.1% vs. 17.0%, p-value < 1x10-5, student's t-test, two-sided). We also observed that in the non-reference SVAs, the youngest non-reference SVAs (SVA_F) show the most variation in VNTR copy numbers compared to the non-reference SVA_Es, than the non-reference SVA_Ds. In addition, we observed the VNTR length distributions in non-reference SVAs across five superpopulations and found that the African population had more length variability (the difference between maximum and minimum: 257.8 bp ± 338.2) comparing to that seen in the other superpopulations (210.0 bp ± 225.0). Our hypotheses for the different length variations of the VNTR in SVAs, between reference and non-reference are: a) SVA elements harbor copy number variations of VNTR in their interior regions that have been associated with changes in local gene expression (*8*), thus those copy numbers in VNTR tend to be conservative by time under the pressure of the purifying selection and fewer copy number variations are left in the reference (older) SVAs than those in the non-reference (younger) SVAs, and b) in the VNTR region of SVAs, there is a sequence, GGGGGGTCAGCCCCCC, that can lead to a structure of imperfect

palindrome, which may result in VNTR deletion or its copy number variation during DNA replication (*8*, *165*). Because of the burden of nucleotide substitution by time, the reference SVA would have less possibility to have an intact segment of this palindrome sequence, which may have lower chance of the DNA replication problem caused by the palindrome and thus less copy number variations happening in the VNTR region of the reference SVAs than that in the non-reference SVAs.

15.3.3 Distributions of poly(A) tract and endonuclease cleavage site sequences

(*Contributors: Weichen Zhou and Ryan Mills*)

The length of the poly(A) tracts can represent the mobilization activities of non-reference retrotransposons (*167–169*). With the help of assembled sequences, we are able to take a close look at the exact poly(A) tract length without worrying about the ambiguous mapping of the short reads. Meanwhile, we are able to identify the sequence distribution of endonuclease cleavage sites for the first time in such large scale population-level samples for different subfamilies of retrotransposons (*170–172*), where L1 retrotransposition machinery occurs.

We divided the non-reference MEI into subfamilies: Alu (AluYb8, AluYa5, AluY, and AluS), L1Hs (L1Ta, L1Ta-1, L1Ta-0, and L1PreTa), and SVAs (SVA_F, SVA_E, and SVA_D) (Fig. S61). We compared in a pairwise manner, the overall length of the poly(A) tracts for each of the Alu, L1H, and SVA subfamilies. In the L1H set, we observed that the poly(A) length in the non-reference calls of the L1Ta subfamily are significantly longer than the ones in L1Ta-0 or L1PreTa (p-value < 0.05, student's t-test, two-sided). We made a similar observation when we compared the poly(A) tract length in the non-reference calls of the AluYb8 or AluYa5 subfamilies with those belonging to the AluS subfamily (p-value < 0.05, student's t-test, two-sided). We did not find anything significant in the non-reference SVA subfamilies. The results that younger subfamilies of non-reference Alus and L1Hs have longer poly(A) tracts indicate that they would have a better chance to be retrotransposed by L1 machinery and tend to be more active in the genomes. We further investigated the endonuclease cleavage sites of all subfamilies. By building the sequence logos for all different subfamilies among diverse populations, we observed a preference for the endonuclease cleavage sites 5'-TTTT/AA-3', as seen in previous studies (*170–172*). Meanwhile, we also observed that the signals of this pattern are getting weaker as the subfamily of MEIs are getting older, indicating the degeneration of this pattern in the retrotransposition history or the slightly different endonuclease cleavage preferences for older subfamilies.

16 Nonredundant callsets, filtering, and properties

(*Contributor: Peter Audano*)

16.1 Nonredundant callset merging

Merging independent PAV samples. We produced PAV calls independently for 32 genomes (11 HiFi and 21 CLR), excluding replicate CLR and child samples, and applied QC steps per sample (see Section 10 and Table S4). We applied the three-step method (see Section 9) to merge them

in a stepwise process similar to previous studies (*4*) where samples are iteratively added to the nonredundant set. To seed the merge, the first sample is taken as the nonredundant set. The next sample is intersected with the nonredundant set using the three-step approach for SVs and indels or exact matches for SNVs (see Section 9.2). Variants that intersect the callset are added as annotations to the existing calls and variants that do not intersect are appended to the callset. This process is repeated until all samples have been merged.

The end product is a set of nonredundant variants that may be supported from one or more samples (merged distribution shown in Fig. 2C). The sample and variant ID are retained so the record can be linked back to the supporting calls without ambiguity. The lead variant, which represents the whole set, is from the first sample with the variant. Since merging order matters, we added HiFi trio parents first (CHS, PUR, then YRI), then we added supplementary HiFi samples ordered by contig N50 (best assembly N50 first), and finally CLR samples ordered by contig N50.

The average callset growth per haplotype is estimated by using the number of singleton variants that would be added to the nonredundant callset by each of the 64 haplotypes if it was merged last. Compared to out-of-Africa haplotypes, African haplotypes are adding 2.21× SVs, 3.70× indels, and 2.97× SNVs (Table S11).

The average callset growth rate per sample by new homozygous variants was estimated by examining callset growth from homozygous variants in the final sample if each sample was added last. Compared to out-of-Africa haplotypes, African samples are growing the callset with new homozygous variants at a rate of 1.80× SVs, 2.40× indels, and 2.74× SNVs per sample (Table S11).

Compared to previous long-read studies (*4*, *19*–*21*, *173*), we added 30,149 insertions and 15,734 deletions that were not previously sequence resolved (Fig. S10).

16.2 Post-merge filtering

Filtering nonredundant SVs. For SVs, we required support from one or more of PBSV, Bionano, DeepVariant, PAV run with LRA alignments (PAV-LRA), LRA-based alignments with CIGAR string variant discovery (LRA Assembly), Subseq, or 61-mer support over breakpoints (details above for each of these methods). For 61-mers, we counted support if non-reference k-mers were identified at variant breakpoints and the k-mers were supported by Illumina reads. If the lead variant passed one or more of these filters, it was accepted into the final callset. By manually inspecting SVs ≥50 kbp (n = 96 SVs), we identified 54 assembly errors that resulted in large deletions where the largest 9 (690 kbp - 17.3 Mbp) were all false calls.

Filtering nonredundant indels and SNVs. For indels, we required support from two or more of DeepVariant, GATK, PAV-LRA, or 61-mer support over breakpoints (same 61-mer approach as for SVs). For SNVs, we required support from two or more of DeepVariant, Longshot, GATK, PAV-LRA, or 61-mer support over breakpoints (same 61-mer approach as for SVs).

Assessing orthogonal callset support. To assess support from other callers and to compare HiFi and CLR, we intersected variant calls from multiple sources for HG00733 (Fig. S63). There is a higher error rate among CLR samples, which is seen in the PAV callset and previous studies (*1*, *4*). This CLR error is reflected in low validation rates by subseq for calls only seen by PBSV CLR. Outside annotated tandem repeats (UCSC track), the trends are more stratified with very little subseq support for any single caller, including variants called by PAV in both HiFi and CLR. Many of these variants will not make it through the filtering stages (Fig. S62).

16.3 Callset quality estimates

SV variant support. Including orthogonal callsets used for filtering, we added support from DeBreak, Inspector, and an LRA-based validation method (*76*) to estimate merged callset FDR (Table S47). Single-method FDR estimates ranged from 35.3% (DeBreak) to 7.38% (Inspector). It is interesting to note that PAV-LRA alone estimates 25.8% FDR, which highlights the importance of the aligner for variant discovery and likely improvements yet to be made to LRA. The callset was filtered with some of these methods (excluding the ones added for QC) and required support from two or more sources, we estimate 5.27% FDR. Inspector alone predicts an FDR of 6.73%.

HG002/NA24385 GIAB callset support. We compared our callset to a recently released SV callset from GIAB benchmark sample HG0002 (NA24385) (*149*), and hence taking advantage of the manual curation work done by GIAB. Following GIAB guidelines, we applied Truvari (Spiral Genetics) to the high-confidence regions for this sample (2.66 Gbp, 85% of the human genome was analyzed in this comparison) yielding an FDR of 3.93%.

Relaxed Mendelian consistency. We investigated evidence of Mendelian inconsistency. Using the three parent–child trios, where both HiFi and CLR data were generated, we assessed Mendelian consistency by intersecting variants with parental variants and allowing up to 1 kbp distance. Based on this analysis, we estimate an FDR of 1.72% for HiFi and 3.26% for CLR. As a control for this analysis, we also investigated each child with respect to incorrect parent pairs yielding a much greater FDR of 30.0% for HiFi and 30.2% for CLR. This shows that our FDR estimates of 1.7 and 3.6% are unlikely to occur by chance.

ONT validation. To compare to an independent sequencing platform, we additionally measured concordance between PAV calls and SVs detected from raw reads from Oxford Nanopore Technologies (ONT). SVs were discovered from read-level alignments by LRA and minimap2 and compared to SVs discovered by PAV from assemblies. Calls from PAV that do not have corresponding SVs discovered in reads were annotated as false discoveries. The FDR for NA24385 using ONT reads was 8% and 12% for deletion and insertion SVs, respectively, for an overall FDR of 10%. Using the same concordance measure with HiFi reads yields an FDR of 4.2% and 3.9% for deletion and insertion SVs, respectively, and a combined FDR of 4.0%. The upper bound of FDR of 10% reflects a technology bias for sensitivity of discovering SVs; using comprehensive comparisons of alignment algorithms and SV-calling methods, ONT data has a

3-5% lower recall than HiFi data (*76*) on the GIAB high-quality SV set.

Technical note: SV discovery using ONT data is highly dependent on how the base calling was performed. When developing the LRA aligner, we measured the accuracy of SV detection methods using ONT called with Guppy version 2.3.5 and Guppy 4.0.1 (*76*). The Truvari precision and recall using Guppy-2.3.5-called reads from GIAB sample NA24385 (HG002) aligned by LRA and called using Sniffles was 0.77 and 0.69, respectively. When applying the same analysis to reads recalled using Guppy 4.0.1, the Truvari precision and recall was 0.95 and 0.94. The recalled reads for NA24385 are available at the Sequence Read Archive under accession PRJNA678534. Notably, the coverage for the ONT dataset is also lower than the original HiFi coverage, which leads to a lower number of PAV variants that have support relative to the HiFi support.

16.4 Variant frequency distributions

(*Contributors: Peter Audano, Jan Korbel, Tobias Rausch*)

In previous work, we observed a linear relationship in the distribution of SV variant allele frequencies on a log scale (*23*). Before applying quality filters to the merged callset, we observed a visible increase in low allele frequency variants. Although the allele frequency was not considered while making callset filtering decisions, low-frequency variants were filtered more heavily (Fig. S64), and the distribution shifts closer to a linear pattern.

16.5 Identify SV clusters - hotspot analysis

(*Contributor: David Porubsky*)

Hotspot definition. To identify SV clusters, we selected the middle position of each SV and submitted it to the 'hotspotter' function from the primatR package (*77*) (parameters: bw=200000, num.trial=1000). This function searches for regions of increased density of SV midpoints around the genome by using the density function to perform a KDE (kernel density estimation). A p-value was calculated by comparing the density profile of the genomic events with the density profile of a randomly subsampled set of genomic events (bootstrap support).

Hotspot analysis. Using the total number of SVs (≥50 bp) detected across all autosomes and chromosome X (n=105,327), we found SVs significantly (p-value < 0.001, permutation test) clustered at the terminal 5 Mbp region of each chromosome with ~4-fold enrichment (Fig. S14-A). To alleviate this bias, we removed SVs that reside at the terminal 5 Mbp (remaining SVs n=73,105) of each chromosome and detected a total of 221 SV hotspots (Fig. S14-B). As expected, we found these hotpots significantly enriched for SDs (p-value < 0.001, ~6.6-fold enrichment, permutation test) (Fig. S14-B left). Of these, 112 hotspots have been previously identified (*23*) while the rest are novel (n=109) (Fig. S13-A,C). As expected, a number of hotspots overlaid with regions of lower mapping quality from our phased contigs against GRCh38. (Fig. S13-B). In total, we report 278 SV hotspots sites including 57 terminal hotpots (Fig. 2F, Table S14) covering ~279 Mbp of the genome (Fig. 2F inset, Fig. S65).

We searched the GENCODE database (version 36) to detect genes overlapping our SV hotspots. We selected only protein-coding genes and removed duplicated gene names (n=24) to get a total of 19,938 protein-coding genes. We tested enrichment of protein-coding genes overlapping our SV hotspots using the R package regioneR. To determine how many immune genes reside in our hotspots, we selected genes reported in innateDB (version 5.4). Before defining the proportion of overlapping genes with our hotspots, we removed duplicated gene names (n=46) to get a total of 4,677 immune genes.

We find the detected SV hotspots (n=278) to be enriched for protein-coding genes (mean enrichment=1.4; p-value=0.001; z-score=6.044, permutation test). This enrichment becomes less significant after removing hotspots mapping to the terminal chromosomal regions (mean enrichment=1.2; p-value=0.012; z-score=2.405, permutation test). In total, we find 3,211 protein-coding genes overlapping at least one SV hotspot and SV hotspots are enriched for genes associated with sensory perception, immunity, and cell differentiation. These genes show a wide variance in mutational tolerance as projected by the number of very high and low pLI values for genes overlapping detected hotpots.

16.6 HLA analysis

(*Contributor: David Porubsky*)

We focused specifically on the major histocompatibility complex (MHC on chr6:28,510,120-33,480,577) to assess the level of completion and variation. This region is sequence resolved (MAPQ60 contig alignments) with a median of two assembled contigs per haploid assembly covering 98.36% of the locus. The assembly with the lowest amount of coverage still resolves 96.35% of the sequence in a single assembled contig, corresponding to missing sequence of ~181 kbp (Table S15). The high completeness of this biologically relevant region allows us to compare the level of divergence across all phased genomes (Fig. S66). At the SNV level, we observed three clusters of high sequence divergence that are also sites of previously detected SV hotspots (Fig. S67) and there is an indication of a population-specific variation (Fig. S68A, arrowhead). We further investigated the effect of this variation on exons. As expected, we found exons of HLA genes carrying more alternative alleles, which resulted in an elevated number of non-synonymous amino-acid changes (Fig. S68). At the SV levels, we found the HLA region to cluster in about seven distinct haplotypes (Fig. S15) with a relatively even distribution of the haplotypes across populations (Fig. 2G, Fig. S69).

16.7 Shared variants

(*Contributor: Peter Audano*)

We defined a set of shared variants that were discovered in all haplotypes. For this analysis, we excluded variants that were not callable in at least one haplotype (i.e., AN = 64 and AF = 1). In total, 1,573 SVs were shared, 1,310 (83.3%) were genotyped with 100% allele frequency in population samples by PanGenie (see Section 19.2), 1,548 (98.4%) were supported by at least one call in our previous work, and 1,118 (71.1%) matched a known shared variant (*4*). Of the

2,238 shared SVs reported by (*4*), 1,118 (50.0%) were also annotated as shared in this study. These variants represent a refined set of reference errors or extreme minor alleles (*4*).

16.8 Distribution of VNTR allele lengths

(*Contributors: Tsung-Yu Lu and Mark Chaisson*)

The boundaries of orthologous VNTR loci were detected in each assembly in order to assess the distribution of VNTR lengths in a manner that is independent of SV-calling algorithms. Because the alignments near VNTR regions are often diffuse and inconsistent between pairs of assemblies, we applied a method previously developed to harmonize VNTR boundaries across assemblies (*174*) to haplotype-resolved assemblies in this study.

VNTR annotations were detected using Tandem Repeats Finder v4.09 (TRF option: 2 7 7 80 10 50 500 -f -d -h) (*175*) to roughly annotate the VNTR regions of five PacBio assemblies (AK1, HG00514, HG00733, NA19240, NA24385). We selected the set of VNTR loci with a motif size greater than 6 bp and a total length greater than 150 bp and less than 10 kbp. For each haplotype, the VNTR coordinates were mapped to GRCh38 to identify homologous VNTR loci. To maintain data quality, VNTR loci that could not be assigned homology were removed from datasets, leaving an initial set of 84,111 VNTRs. Base-level orthology mapping of VNTRs across the 70 haplotypes were annotated using the boundary expansion algorithm in danbing-tk (*174*) by maintaining the following invariant for each VNTR: the flanking sequence in any of the haplotypes does not share k-mers with the VNTR regions from all haplotypes. VNTR boundaries in each haplotype are iteratively expanded until the invariant is true or if expansion exceeds 10 kbp in either 5' or 3' direction. The size of the flanking regions was chosen to be 700 bp, which is approximately the upper bound of the insert size of typical short reads. The following QC step removes a haplotype if its VNTR annotation is within 700 bp to breakpoints or if the orthology mapping location to GRCh38 is different from the majority of haplotypes. A VNTR locus with the number of supporting haplotypes <90% of the total number of haplotypes is also removed. Adjacent VNTR loci within 700 bp to each other in any of the haplotypes will induce a merging step over all haplotypes. Haplotypes with distance between adjacent loci inconsistent with the majority of haplotypes are removed. Finally, VNTR loci with the number of supporting haplotypes <80% of the total number of haplotypes are removed, leaving 79,762 of the initial 84,411 loci (Figs. S102, S103).

As a resource we include a table of the top 1% of VNTR alleles ranked by variance of allele length that are also within 10 kbp of a gene in Table S48.

VALIDATION

17 SV validation by reads and contigs

17.1 Subseq validations with raw reads

(*Contributor: Peter Audano*)

As a QC metric, we assessed raw-read support for SVs with the minimap2 alignments used for PBSV. For each SV, a window around the SV is defined by extending the breakpoints upstream and downstream of the SV (padding) with a larger pad for larger SVs (SVLEN < 100: 20 bp, 100 ≤ SVLEN < 200: 25 bp, 200 ≤ SVLEN < 500: 40, 500 ≤ SVLEN < 1 kbp: 100 bp, 1 kbp ≤ SVLEN < 2 kbp: 200 bp, 2 kbp ≤ SVLEN < 4 kbp: 250 bp, SVLEN ≥ 4 kbp: 300 bp). We found these pad values to work well for each variant class based on expected distributions and agreement with other QC metrics, such as orthogonal caller support.

For each padded window, all reads traversing the window were identified and the length of the reads mapping to the window (distance in read space between the bases mapped to the first and last base of the window) using custom code (subseqfa, https://github.com/EichlerLab/seqtools, project version released via (*88*)) (Fig. S70). If there were fewer than four reads spanning the window, we did not attempt to make a validation call. If four or more reads spanned the window and at least two supported the SV call allowing up to a 50% shift in the expected size of the window, then the variant call was validated by this method. Although we have PBSV calls from the same alignments, this method allows SV calls to be fragmented into multiple insertion or deletion events relative to the SV call.

There is a clear difference between HiFi and CLR validations with HiFi showing a much cleaner pattern (Fig. S71). CLR is much more difficult to validate and we expect a higher validation error rate with CLR with more calls falsely invalidated.

17.2 LRA read and contig alignment support

(***Contributors: Jingwen Ren, Mark Chaisson***)

In order to filter out calls due to misassembly or misalignment, SVs were detected from raw reads. A combination of two alignment methods was used: LRA v1.0 (*76*) and minimap version 2.17-r941 (note: we capitalized the name “lra” throughout this manuscript to ease readability). An aligned read was considered to be supporting the variant if an SV was detected in the read with length between 50% and 200% of the length of the assembly-based call (e.g., 50% length overlap) and with breakpoints within 1 kbp. Calls are considered validated if at least four reads support the call in either dataset (Fig. S72). By including both alignment approaches, additional support is given for SV support.

As expected, the application of two alignment methods can provide support for more variants than one alignment method alone. Considering SV calls from the CLR assembly of the Puerto Rican individual, HG00733, 89.3% of calls had at least four reads supporting an SV from either method, compared to 85.1% of calls having similar support by only LRA and 86.7% of calls are supported by minimap2 alignments. An example of a call that is not supported by either alignment method is shown in Fig. S73.

The average percentage of variants detected from HiFi-based assemblies supported by either method was 92.3%, with 86.7% of variants supported by LRA and 88.6% of variants supported by minimap2. The average percentage of variants detected from CLR-based assemblies was lower: 85.0% of variants had support from at least one alignment method (80.1% LRA and 82.3% minimap2). The full per-assembly rate of supported variants is shown in Fig. S74 and given in Table S49 for normal, wide, and broad support. The fraction of variants that are not supported generally increases as the coverage decreases; however, this is a combination of the effect of assembly quality and number of reads available to support a call. The genome-wide SV distribution that is not supported by reads is shown in Figs. S75 and S76.

17.3 Comparison of phase 1 and phase 2 callsets

(*Contributors: Jingwen Ren, Mark Chaisson*)

Variant calls are made from different alignment approaches between HGSVC phase 1 (*1*) and phase 2 (this study). Various stringencies of relative length and proximity to SVs were used to measure the number of unique calls in the HGSVC phase 1 versus phase 2 callsets. The SV callsets from phase 1 generated only by local assembly (phase1-asm), as well as the multi-technology, merged, nonredundant (phase1-mnr) callsets were compared to PAV calls made from HiFi and CLR assemblies. The phase 1 callsets were filtered to ensure at least four reads supporting the call, similar to the calculated rate for PAV calls. The counts of the calls that are compared are in Table S50.

Under the most lenient overlap, where a call of the same type is made within 1 kbp, an average of 1,279 variants are missing from the PAV CCS assembly-based callsets, and 1,274 variants are missing from the PAV CLR assembly-based callsets when compared to the phase1-asm callsets. Also, 3,618 and 3,584 calls are missing from the CCS/CLR callsets, respectively, when compared to the mnr callset (Table S51). Requiring one call to be within 25% length of the other, the number of calls that are unique to phase 1 increases such that 2,141 and 2,144 calls are missing from the CCS/CLR callsets compared to the phase1-asm callset, and 4,795/4,775 calls are missing from the CCS/CLR callsets compared to the mnr callset. The number of calls unique to phase 1 increases as the length of call increases (Table S51). This indicates a difference in the representation of calls, rather than missed calls, and so subsequent analyses could use the metric of any proximal SV (within 1 kbp) of the same type, without requiring a specific length.

We categorized each variant with the following genomic features: overlapping SDs, tandem repeats, and overlapping assembled contigs to determine if variants were more likely to be missed

if they were in that genomic feature (variants will be de-facto missed if they do not overlap an assembly contig). Across all classes of comparisons, 493–837 variants were missed that overlap SDs relative to the fraction of calls made in SDs in the phase 1 calls. This represents 22–44% of calls missed in phase 2, and a 2.7- to 5.8-fold enrichment relative to all missed calls. Tandem repeats, while representing a greater number of variants that are missing from PAV assembly calls (914–2,587 variants/callset), also represent the majority of calls made from long-read sequencing and are not overrepresented as a fraction of missed calls (enrichment 0.71–1.09).

The majority of variants in the phase 1 callsets that are missing from the assembly-based callsets are annotated as overlapping assemblies. Each assembly was mapped using LRA and a base was considered overlapping by an assembly if an alignment from either haplotype overlapped. Depending on the callset, between 92–96% of variants overlapped an assembly. This indicates that missing calls are due to either filtering parameters from assembly-based alignments or incomplete representation of haplotypes in the assemblies.

17.4 Variant concordance by reads, contigs, and assemblies (Inspector)

(*Contributors: Zechen Chong, Yu Chen*)

To further validate SV calls with raw reads, Inspector version v1.0.1 (https://github.com/ChongLab/Inspector, (*88*)) was applied on all CLR and CCS samples. Raw reads were aligned against all assembled and phased contigs with minimap2 version 2.15-r905. Read alignments at each SV breakpoint were used for evaluating the confidence of SV calls. A highly confident SV call should have reasonable depth (half of the sequencing depth due to haplotype-resolved assembly) and no indel or clipped alignments. Strong indel signals at the SV breakpoint often suggest the presence of assembly errors. Extremely high or low read-alignment depth suggests assembly collapse or expansion, respectively. Variants that passed both filters were classified as “PASS”, while variants with indels or clipped alignment near breakpoints were marked as “NoisyAlign”. Variants were marked as “HighDepth” or “LowDepth” when read alignment depth at SV breakpoints were higher than two times of the mean sequencing depth or ≤3, respectively (Fig. S77).

17.5 Raw read variant and breakpoint concordance

(*Contributors: Kai Ye, Jiadong Lin, Xiaofei Yang*)

To examine the breakpoint accuracy of the SV callset, MUMMer v4.0 (*176*) and NucDiff v2.0.3 (*177*) based raw reads realignment were applied to all CCS and CLR samples. Since insertions only have one breakpoint, we used the reported breakpoint location plus the insertion length as another breakpoint for further evaluation. All SV spanning reads were extracted from the pbmm2-aligned BAM file with SAMtools (*178*). If no reads were fetched, this call was marked as “NoRawReads”. On the contrary, PAV calls with spanning reads were realigned with MUMMer and their corresponding breakpoint type was classified by NucDiff. We first examined whether the breakpoints of each call could be supported by the realigned breakpoint. If none of the breakpoint differences between the PAV call and the realigned breakpoint are smaller than the allowed

distance (allowed_dist=500), this PAV call was marked as “NoRealignSupport”. For the rest of the PAV calls, we then grouped the breakpoints from each realigned read based on their position (100 bp difference allowed) and type to make realigned calls. For a PAV call and its corresponding realigned call, if the position difference of the two breakpoints are both smaller than maximum breakpoint shift (allowed_shift), this PAV call is considered as “Precise”, otherwise it is considered as “Imprecise”. As a result, PAV calls were marked as “NoRawReads”, “NoRealignSupport”, “Precise” and “Imprecise” (Fig. S78).

For the HiFi samples, the realignment validation rate for both DEL and INS is around 75% for each sample (Fig. S79-A). We observed approximately 1.5% of the calls as “NoRawReads”, indicating the alignment issue of HiFi reads in these regions. But we found some regions could be aligned with NGMLR V0.2.6 (Fig. S80-A). By setting the allowed_dist = 500 bp, ~9% and ~7% of the DELs and INSs, respectively, became labeled as “NoRealignSupport”, with ~80% of the calls being smaller than 100 bp (Fig. S80-B). We further annotated these SVs with RepeatMasker and TRF (Fig. S80-C), leading to ~65% calls inside the repeat regions and more than half of these calls (58%) found in STR regions (repeat unit ≤7 bp). For the CLR samples, we used the same parameters (Fig. S79-C) and noticed the percentage of precise calls to be highly variable among CLR samples, whereas HiFi-based calls showed consistently better performance for both DEL and INS (~75%). One possible explanation for this observation is that the segment chaining algorithm used by MUMMer and increased sequencing errors by CLR lead to inconsistent performances for the CLR dataset.

17.6 Breakpoint analysis

(*Contributors: Sushant Kumar and Mark Gerstein*)

We further characterized the underlying mechanism for a subset of our SV (deletion and insertion) callsets after identifying their precise breakpoints. For the mechanism analysis, we applied the mechanism detection module of our previously published BreakSeq method. Briefly, the mechanism module of BreakSeq utilizes distinct sequence-based signatures within and around the breakpoint junction of a given SV to detect its underlying mechanism. For instance, it uses the RepeatMasker program to detect extensive coverage of tandem repeats and low-complexity regions within a given SV to classify them as a variable number of tandem repeats (VNTRs). Similarly, SVs belonging to the nonallelic homologous recombination (NAHR) class show extensive homology at their breakpoint junction (consisting of flanking sequences of a given SV). SVs that align to known interspersed MEIs in the genome are considered transposable element insertions (TEIs). It further subclassifies TEIs based on whether a given SV is aligned to a single (STEI) or multiple (MTEI) transposable element insertion. Furthermore, the SV mechanism pipeline also identifies processed pseudogenes that originate due to TEI-associated mechanism. Finally, SVs for which BreakSeq extracts the flanking sequence, but lacks any of the signatures mentioned above, were classified to the nonhomologous recombination (NHR) mechanism class.

In this work, we applied BreakSeq’s mechanism pipeline in two distinct settings by varying the homology length cutoff (≥50 bp & ≥200 bp) and flanking sequence length (200 bp & 1 kbp) for

detecting NAHR events (Table S17 A-B). Furthermore, in the first setting (homology length ≥50 bp), we grouped VNTR events with repeat length ≥50 bp and NAHR events together to define a new homology-associated event class. In contrast, VNTRs (with repeat length ≥50 bp and covering 50% of SV regions) constituted the newly defined homology-associated category in our second analysis set. Finally, deletions and insertions that overlap with the assembly-resolved MEIs (generated as part of the current work) were classified as TEIs. We performed these analyses on the subset of SVs, for which precise breakpoint definition was available. Overall, we applied our mechanism analysis pipeline on 97,162 SVs (33,531 DELs and 63,631 INSs). Furthermore, we also quantified mechanism distribution differences for a subset of SVs with precise breakpoints that overlapped with the short-read-based SV callset (Table S17 C-D).

Our mechanism characterization of the long-read-based SVs indicated different mechanism distribution for deletions and insertions (Fig. S81). For instance, while we observed that a large fraction of deletions and insertions belonged to the homology-associated mechanism class, the homology-associated mechanism had, on average, a relatively higher contribution to deletions (mean fraction value of 0.628) than insertions (mean fraction value of 0.512). In contrast, a somewhat smaller fraction of deletions (mean value of 0.14) and insertions (mean value of 0.18) belonged to the NHR mechanism class. We observed a distinct mechanism composition pattern for our SV dataset with a more stringent definition of NAHR events (homology length ≥200 bp) (Fig. S82). While, the homology-associated and NAHR mechanism class together, on average (mean value of 0.187), had a slightly higher contribution toward describing deletions compared to the NHR class (mean value of 0.172), a large fraction of insertions belonged to the NHR mechanism category (mean value of 0.266).

Additionally, we compared the underlying mechanism distribution for deletions and insertions of different lengths (Fig. S83). We classified deletions and insertions into three distinct categories: 1) length below 200 bp (lt200bp), 2) length between 200 and 1 kbp (lt1K), and 3) length above 1 kbp (ge1K). Overall, we found that SVs with length below 200 bp primarily belonged to homology-associated class (mean fraction value of 0.8). In contrast, a relatively larger fraction of SVs belonging to lt1K and ge1K categories corresponded to TEI (mean value of 0.377) and NHR (mean value of 0.416) mechanism classes. We observed similar mechanism contribution pattern differences for distinct length categories of insertions and deletions (Fig. S83). However, with a more stringent definition for the NAHR mechanism category, we observed few differences in length distribution pattern for deletions and insertions (Fig. S84). We observed that a relatively higher fraction of deletions belonging to the lt1K category corresponded to the NHR mechanism class (mean value of 0.42). In contrast, the NHR mechanism contributed similarly to insertions corresponding to the lt200bp (mean fraction value of 0.302) and lt1K (mean fraction value of 0.305) categories.

Finally, we also quantified relative contributions of different mechanism categories for SVs overlapping with distinct functional elements in the genome (Fig. S85). Overall, we found a large fraction of functionally relevant SVs belonged to the homology-associated and NHR classes. Interestingly, we found that homology-associated mechanisms explained most of the functionally pertinent deletions and insertions. In contrast, the NHR mechanism had a more considerable

contribution to the emergence of functionally relevant deletions. We also observed subtle variability in the contribution of distinct mechanisms toward deletions overlapping with different functional elements. For instance, a relatively larger fraction of deletions overlapping with DNase I hypersensitive sites (DHS, mean value of 0.439) and regulatory elements belonged to NHR mechanism class (mean value of 0.431). In contrast, a relatively larger fraction of deletions overlapping with ncRNA (mean value of 0.49) and untranslated regions (UTRs; mean value of 0.59 for 3' UTR and 0.66 for 5' UTR) belonged to the homology-associated class. Similarly, a larger fraction of insertions overlapping with intronic regions (mean value of 0.51), 5' UTRs (mean value of 0.527), and CpG islands (mean value of 0.53) regions belonged to the homology-associated mechanism category. For a stringent definition of the NAHR mechanism class, we saw a completely distinct contribution distribution for deletions and insertions overlapping with different functional elements (Fig. S86). In particular, we found that a relatively large fraction of functionally relevant deletions belonged to the NHR class compared to the homology-mediated mechanism class. Deletions overlapping with DHS (mean fraction value of 0.465) and CCRE elements (mean fraction value of 0.467) primarily belonged to the NHR mechanism category. We observed a relatively higher contribution from homology-associated mechanism category toward functionally relevant insertions compared to deletions. However, there was a sharp decline in the NAHR mechanism's contribution toward explaining insertions overlapping with various functional elements in the genome.

18 Analyzing non-reference k-mers using 3,202 genomes

(*Contributor: Tobias Rausch*)

Using 3,202 Illumina sequenced (30X) samples, including 2,504 unrelated samples from the 1000GP and an additional 698 related samples, we created a database of non-reference k-mers absent in the GRCh37 reference of the 1000GP (hs37d5). For each sample, we first error corrected the raw sequencing reads with Lighter v1.1.2 (*99*) using a k-mer length of 23 and then computed a compacted de Bruijn graph with BCALM 2 v2.3.0 (*179*) using a k-mer length of 61. We then used dicey v0.1.6 from the GEAR genomics framework (*180*) to compute k-mer hash sums and to extract non-reference k-mers for each sample by means of depleting the reference k-mers present in hs37d5. Each sample contained ~270–370 million non-reference k-mers, with African samples showing an increase of non-reference k-mers (Fig. S87).

We then aggregated all non-reference k-mers in a database, separately for the set of 2,504 unrelated samples from the 1000GP and the additional 698 related samples.

We then utilized this resource to evaluate the assemblies and assigned each non-reference k-mer from the assembly its occurrence count in the 1000GP. We interpreted a count of zero, a so-called missing, non-reference k-mer, as a likely assembly error or as a k-mer that cannot be sequenced by Illumina due to very high or low GC content. Notably, the CLR assemblies showed, on average, 8.3% missing, non-reference k-mers compared to 3.7% for the CCS assemblies, reflecting the higher sequencing accuracy of CCS compared to CLR.

We also used this non-reference k-mer resource to assess whether assembly-based SV calls are supported or unsupported in the Illumina data using a simple approach that interrogates SV breakpoint k-mers of length 61 for being present or absent in the non-reference k-mer database of the 3,202 Illumina-sequenced genomes.

GENOTYPING AND ASSOCIATION

19 Genotyping Illumina genomes

19.1 Paragraph genotyping

(*Contributors: Wayne Clarke and Michael Zody*)

Using Paragraph (*139*), a sequence graph-based genotyping tool, we genotyped 107,590 SVs using a per-sample approach. Paragraph requires, as input, a) a VCF file of variants, b) a reference sequence FASTA, c) a manifest containing the path to the sample BAM file, and d) additional information including the read length, average read depth, and the sample's sex. Paragraph outputs a genotype for each of the input variant sites, which we merged into a multi-sample VCF. Variants were evaluated for Mendelian consistency and Hardy-Weinberg equilibrium (HWE), with sites violating Mendelian consistency and sites with extreme deviations from HWE being removed, resulting in 96,145 SVs carried forward.

19.2 PanGenie genotyping and graph construction

(*Contributor: Jana Ebler*)

PanGenie. We used our recently developed method PanGenie (*42*) to showcase the utility of our haplotype resources for improving short-read-based genotyping. This method uses a panel of known population haplotypes and sequencing reads in order to genotype a new sample. In a first step, PanGenie constructs an acyclic and directed graph that represents the genetic variation across the input panel haplotypes. Variants are represented as bubble structures and each haplotype constitutes one path through this graph. In the next step, k-mers uniquely characterizing variant alleles are determined and counted in the reads. These counts provide information about the presence or absence of variant alleles in the sample to be genotyped. However, SVs might be located in genomic regions poorly covered by unique k-mers. Therefore, our model additionally leverages the global haplotype structure provided by the input haplotype panel. This enables genotype imputation, which is especially useful in such difficult-to-access regions.

PanGenie is based on a Hidden Markov model (HMM), which integrates information from k-mer counts and known haplotypes. It aims at constructing the unknown sample haplotypes such that they best explain the observed read k-mer counts and, at the same time, are mosaics of the panel haplotypes. Genotype likelihoods are computed based on the Forward-Backward algorithm. This model has been demonstrated to provide ultra-fast genotyping by bypassing the expensive read

alignment step, making it especially well suited for genotyping large sets of samples. Leveraging the provided haplotype structure enables one to access regions of the genome otherwise hard to interrogate from short reads only. Since the base version of PanGenie becomes rather slow for panels with more than 15 individuals (30 haplotypes), we extended it to efficiently handle larger panels for genotyping. Per default, this extended version of PanGenie randomly divides the panel samples into groups of 15 haplotypes and computes genotype likelihoods based on an HMM, constructed only on these paths. The likelihoods obtained for all subsets in this way are later summed up and normalized in order to obtain the final genotype likelihoods from which to make a genotype prediction.

Using this approach, we genotyped all 3,202 sequenced samples in order to demonstrate the utility of our method for large, short-read-based studies. For each sample, we provided PanGenie with short-read sequencing reads (in FASTQ format) as well as a multisample VCF file containing phased variant calls from the 64 assemblies. This input VCF file is derived from the merged PAV callsets produced for SNVs, indels, and SVs. At first, we removed all positions at which more than 20% of the panel haplotypes carried a missing allele. Furthermore, we kept only variants located on chromosomes 1-22 and chromosome X for genotyping. We then created a multiallelic VCF representation in which overlapping variants are combined into multiallelic positions. Some variants in the input PAV callsets were overlapping on the same haplotype (e.g., an SNV inside of a deletion). We removed such conflicts by setting the corresponding alleles to missing ("."). In this way, the VCF represents an acyclic and directed variation graph representing the panel genomes. Table S29 provides an overview of the number of variants obtained. As an output, PanGenie generates a VCF file that contains a genotype for all these variants. We merged all 3,202 sample VCFs with genotypes into a single multisample VCF, and additionally produced a bi-allelic version of this VCF.

We started with a pilot set consisting of 300 individuals selected from the 3,202 samples. This subset was constructed by randomly choosing 20 trios from each of the five superpopulations (AFR, AMR, EAS, EUR, SAS). We then ran PanGenie in order to genotype all 15.5 M SNVs, 1.03 M indels, and 96.1 K SVs across these 300 samples. For comparison, we additionally ran Paragraph to derive genotypes for all SVs. As a QC measure, we computed allele frequencies across all 200 unrelated samples from the PanGenie and Paragraph genotypes and compared them to the allele frequencies derived from the PAV calls for all 64 assembly haplotypes. For Paragraph, we observed allele frequency correlations (Pearson correlation) of 0.61 and 0.54 for deletions and insertions. For PanGenie, these values are 0.85 and 0.86. Fig. S30 indicates that Paragraph tends to genotype variants as heterozygous and especially struggles with genotyping insertions. We concluded that PanGenie seems to be more suitable for genotyping our SV calls and proceeded with genotyping all SNVs, indels, and SVs across the 3,202 samples using PanGenie. We determined the number of heterozygous SVs for each population from these genotypes (Fig. S88) and observed higher numbers for the African populations, reflecting their increased genetic diversity. We also computed allele frequency correlations from the allele frequencies derived from the PanGenie genotypes for all 2,504 unrelated samples and the PAV calls. We observed correlations of 0.98, 0.95, and 0.85 for SNVs, indels, and SVs, respectively. However, these numbers indicate that there are variants for which PanGenie and PAV allele

frequencies differ significantly. In order to filter out such potentially wrong genotyped calls, we defined a strict subset of variants based on statistics we computed from the genotypes of the 3,202 samples. We defined the five filters listed below:

- **ac0_fail**: A variant was genotyped as absent (genotype 0/0) by PanGenie in all 3,202 samples (i.e., the allele frequency is zero).
- **mendel_fail**: There are 602 trios among the 3,202 genotyped samples. For each variant, we counted the number of trios with Mendelian-consistent genotypes. Here, we only take trios with at least two different genotypes into consideration, meaning we skip trios in which all samples were typed as 0/0, 0/1 or 1/1 respectively. A variant fails this filter if the Mendelian consistency is below 90%.
- **gq_fail**: At least 200 low-quality genotypes were reported by PanGenie.
- **nonref_fail**: All panel samples were genotyped as homozygous reference.
- **loo_fail**: In addition to genotyping the 3,202 samples, we conducted a leave-one-out experiment, in which we repeatedly take out one of the panel samples from the input and use PanGenie to genotype it based on the remaining samples in the panel. We then compare the predicted genotype to the left-out, ground-truth genotype of the sample. This enables us to compute the genotype concordance across all panel samples at each variant position. This filter fails if the genotype concordance of the panel samples is below 80%.

We applied these filters to all SNVs, indels, and SVs genotyped by PanGenie. For SVs, 16,343 out of 60,238 insertions (27%) failed the "ac0_fail" filter and were genotyped with an allele frequency of zero across all 3,202 samples. For deletions, 8,948 out of 35,862 (25%) were genotyped as homozygous reference in all samples. About 57% of these variants are rare and were carried by only a single haplotype in the input panel. Such variants are, in particular, difficult to genotype by a panel-based approach like PanGenie, especially if the k-mer counts show no strong indication for the presence of an allele in a sample. To obtain a filtered callset, we removed all variants for which at least one of the five filters failed. This leads to a rather stringent, but high-quality set of genotypes that serves as a basis for further analysis. Our filtered set contains 12,283,650 SNVs (79%), 705,893 indels (68%), and 24,107 SVs (25%).

We provide callset statistics for this strict set in Fig. 5B and Fig. S89. Fig. 5B shows the allele frequencies obtained for SVs across the PanGenie genotypes, as well as the corresponding allele frequencies in the input panel haplotypes. The allele frequencies for PanGenie were computed based on all 2,504 unrelated individuals as part of the 3,202 samples. For both variant types, insertions and deletions, the allele frequencies match well with very few outliers, indicating that the genotypes are of good quality. We obtained an allele frequency correlation of 0.99 (0.98 for deletions, 0.99 for insertions). Likewise, allele frequency correlations for SNVs and indels in this filtered set are 0.99 and 0.99, respectively. We also investigated the relationship between variant length and allele frequencies across the PanGenie genotypes. The peaks (Fig. S89) show a clear tendency of Alu insertions towards lower allele frequencies, while it seems to be the opposite for Alu deletions. We suspect that this behavior is caused by Alu insertions present in the reference genome. Figs. S27 and S28 show the relationship between Fst values of the five superpopulations

(AFR, AMR, EAS, EUR, SAS) and the length of the SVs. For each superpopulation, Fst was computed between individuals that are part of it and the union of the remaining populations. We observed higher Fst values for African and East Asian populations indicating higher degrees of differentiation among these populations.

Defining a lenient set. For SVs, our filtered set contains only 25% of all input variants. In addition to this strict set, we also defined a larger, more lenient set of SVs using a machine-learning approach based on support vector regression. Our model is designed to assign scores close to -1 to poorly genotyped SVs and scores around 1 to those passing all our filters. For training, we used our strict SV set as "true positives". The "true negatives" were defined as all variants genotyped with an allele frequency larger than zero ("ac0_fail" did not fail) that failed at least three of the remaining filters. The regression is based on 70 features collected from the PanGenie genotypes, the leave-one-out experiments, and concordances with k-mer-based presence/absence genotyping (see Section 18). We predicted regression scores for all yet-unlabeled SVs (with an allele frequency > 0). We then created more lenient callsets by adding those variants to the strict set for which the scores were above a certain threshold. We investigated three different cutoffs: -0.5, 0.0, and 0.5. Table S52 contains corresponding callset statistics. As expected, numbers improve as we increase the cutoff that we use to define a lenient callset. While allele frequency correlations are 0.86 and 0.85 for insertions and deletions, respectively, in the unfiltered set, they reach levels between 0.95 and 0.98 when applying the different cutoffs. Likewise, average genotype concordances of the PanGenie genotypes with the PAV calls for the assembly samples increase, reaching levels above 94% for cutoff 0.5. While containing >50% more variants, statistics for the lenient set with cutoff 0.5 are close to the strict set for which we observed correlations of 0.99 for insertions and deletions, as well as genotype concordances of 96.9% and 96.1%, respectively.

20 RNA-seq analysis

20.1 Read QC and mapping

(*Contributors: Arvis Sulovari, Marc Jan Bonder, Jan Korbel*)

To gain insight into the molecular impact of indels (insertions and deletions <50 bp), SVs (≥50 bp) and SNVs, we jointly analyzed two lymphoblastoid cell line (LCL) RNA-seq datasets, the GEUVADIS study (n=462) as well as 33 newly generated, deeply sequenced (>230M output/sample) 1000GP LCL RNA-seq samples (d1KG). To minimize batch effects, we re-analyzed the raw data.

First, low-quality reads were removed and both adapters and low-quality bases were trimmed using Trim Galore! (*78*) (v0.6.5), a wrapper around FastQC (*181*), and cutadapt (*182*). Next, trimmed and QC reads were aligned using STAR (*79*) (v2.7.5a) with a custom two-pass alignment. All alignments were performed using the GRCh38.p12 reference genome and NCBI Homo sapiens gene annotation version R109.20190125. The first pass was done on all samples,

using the default parameters as proposed by ENCODE (c.f. STAR manual) with the exception of the two-pass alignment flag. Based on this mapping and an initial expression quantification based on featureCounts (*80*) (subread 2.0.1), we performed sample QC.

We chose to remove samples with: 1) high multimapping percentage (>20%), 2) low % Picard usable bases (<85%), and 3) samples failing per_base_n_content ("fail") from FastQC. Afterwards, we performed a PCA on the edgeR (*183*) normalized initial featureCount quantification and removed outliers based on the first two principal components (per dataset), keeping in 411 GEUVADIS samples and all of the deeply sequenced 1000GP samples. Using VerifyBamID (*184*) (v 1.1.3) we verified the links between the genotype and the RNA-seq data, using the PanGenie (*42*) single-nucleotide polymorphism (SNP) genotypes (Section 19) as a reference. First, we verified low FREEMIX scores for the passed QC samples (<0.1) and filtered links between genotype and expression data (CHIPMIX <0.02). This removed an additional 14 GEUVADIS samples, leaving 397 samples. For the d1KG samples, all RNA-seq data matched to the expected 1000GP samples. No sample swaps were observed in the GEUVADIS cohort.

After sample QC, we obtained 430 high-quality samples, which underwent the second-pass mapping in STAR. In this second pass, we used the same settings with two additions: 1) WASP output mode, and 2) aggregated splice junction database from the first round of alignments. Specifically, we aggregated the junction databases of the high-quality samples and included junctions that: 1) were observed in at least 40 RNA-seq runs, 2) had support of at least 45 uniquely mapping reads, and 3) had spanning reads support of at least 20 bases before and after the splice junction. In total, we included 135,356 high-quality *de novo* LCL-based splice junctions to the STAR junction database. To enable the WASP mode, which requires SNP information, we chose to use a custom (single) VCF with all SNP variants observed in the strict PanGenie genotyping (Section 19) with a minor allele frequency (MAF) of 1% and above in the samples with RNA-seq data. After mapping, we filtered the reads based on the WASP filters, allowing reads without a WASP flag or the flag "vW:i:1".

20.2 Expression and splicing quantification and normalization

Based on the filtered alignments, we quantified both gene-expression levels and splicing levels. The gene-expression levels are based on featureCounts, using "-p -B -C --primary -p", to count only primary alignments, not count chimeric reads, and count only reads where both ends of paired-end sequencing data are mapped. The featureCount read counts were transformed to transcripts per million and normalized for library size using edgeR, before log transformation. Last, to quantify splicing we used leafCutter. We took the WASP-filtered expression data and used regTools v0.6.0 (*185*) to extract junction reads, i.e., reads spanning splicing junctions, and then ran leafCutter v0.2.9 (*81*) to cluster these splicing junctions into splice clusters, clusters with overlapping exon start/end points, and finally quantified the splice-ratios for the splice junction clusters.

21 QTL and GWAS

21.1 eQTL analyses

(*Contributors: Marc Jan Bonder, Arvis Sulovari, Yang Li, Jan Korbel*)

To assess the impact of SVs, indels, and SNVs on gene expression and splicing, we performed expression quantitative trait locus (eQTL) as well as splicing quantitative trait locus (sQTL) mapping using the high-quality genotypes from the PanGenie strict-set (Section 19). The eQTL and sQTL analyses were carried out using a linear mixed-model approach implemented in LIMIX V2 (*82*–*84*) and employing an IBD matrix to account for population structure. The IBD matrix was calculated with PLINK (*186*) (v1.0.7), using pruned SNP variants (MAF>5%) from the PanGenie strict-set. The QTL mappings were run in a mega-analysis setting, analyzing the 430 expression samples (397 GEUVADIS and 33 d1KG), originating from 427 unique donors, all in one single analysis.

To select the number of expression covariates to include in the eQTL mapping, we initially performed a focused eQTL mapping analysis for chromosome 2 on SNPs with at least an MAF of 10% and within a window of 10 kbp around the genes. By varying the number of hidden factors taken along in the eQTL mapping (including 0-100 principal components [PCs] by steps of 5), we observed that using 60 PCs maximizes the fraction of genes with an eQTL (eGenes) discovered. Therefore, we selected 60 PCs to correct our eQTL map (Fig. S90).

Next, we performed the full *cis*-eQTL mapping analysis linking gene expression levels to all PanGenie-derived high-quality genotypes, within a window of 1 Mbp centered around the gene body and testing all variants with an MAF of 1% and higher and an HWE (HWE exact test p-value ≥ 0.0001. We simultaneously considered all variants genotyped using PanGenie (i.e., SNVs, indels, and SVs). For SVs, we made sure to include all variants intersecting with the test window, in order not to miss potential SV associations. To correct for multiple testing, we used an approach related to FastQTL (*187*), first correcting for genetic variants based on genotype permutations and using a beta approximation to increase precision, followed by Storey's Q-value correction (*188*), correcting for the number of tested genes. By mapping the permuted p-value that corresponds to gene level FDR of 5%, we derived all eQTLs (for further details on this approach, see (*84*)).

In the eQTL mapping analysis we considered 23,953 genes, which were expressed (i.e., non-zero count) in at least 20% of the samples and which showed an average count (transcripts per million) of at least 0.01 in the 430 post-QC RNA-seq samples. Using this setup, we identified 9,413 lead cis-eQTLs using an FDR of 5% and 850,479 cis-eQTLs in total. Of the lead eQTLs, 8,488 were SNV-eQTLs (8,141 unique SNVs, 90.2% of the total cis-eQTL signals), 887 are indel-eQTLs (851 unique indels, length ≤49 bp, 9.4% of total) and 38 are SV-eQTLs (36 unique SVs, length ≥50 bp, 0.40%) (Table S31). If there was full linkage disequilibrium (LD) between variants, we treat the longest genetic variant as the lead. The set of cis-eQTL hits was enriched for indels

(Fisher's exact test [FET] p-value = 8.8e-113, OR = 1.19 [95% CI: 1.17-1.2]) and SVs (FET p-value = 1.02e-06, OR = 1.21 [1.12 - 1.3]), and depleted for SNPs (FET p-value = 1.8e-118, OR = 0.84 [0.82-0.85]) (Table S33). The SV enrichments are mostly driven by deletions (SV-deletion-specific FET p-value = 2.038e-08, OR: 1.41 [1.25-1.61]), whereas SV insertions are enriched but not significantly (SV insertion FET p-value = 0.1, OR = 1.09 [0.98-1.21]). The insertion indel class is more significantly enriched (FET p-value = 1.38e-108, OR = 1.29 [1.26-1.33]) as compared to the indel deletion class (FET p-value = 5.4e-26, OR = 1.11 [1.09-1.12]). All association tests were conducted on nonredundant counts of variants, using the expected proportions of SNVs, indels, and SVs from the full set of tested high-quality genotypes. Of note for the deletion/insertion labels are all seen as compared to GRCh38, and here we do not distinguish between different genomic regions that get deleted or inserted. The average indel-eQTL and SV-eQTL lengths were 4.1 bp and 836 bp, respectively. Additionally, eight of the SV-eQTLs (24%) were Alu insertions with lengths ranging from 120 bp to 327 bp.

As an analogous measure of the importance of SVs on gene expression variation, we estimated the fraction of expression heritability that is explained by SVs. Specifically, to do so we used a linear mixed model implemented in LIMIX (V2) (*82*–*84*), where we use a fix effect to reflect the SV and a random effect based on the top 1,000 most significant SNVs and indels in the cis-region to reflect non-SV genetic variation. This approach is similar to several prior studies (*45*, *189*–*191*). We ran these heritability estimations for all genes with at least one significant SV-eQTL. In line with previous research, we estimated the heritability of SVs on gene expression finding that, on average, the top SV for eGenes explains 6.6% of expression variability versus 24.2% for the top 1,000 SNVs and indels (Fig. S91). These estimates are also in line with previous work (*45*). For eGenes with at least one significant SV-eQTL (1,505), we find that for 347 (23.1%) genes the explained variation is higher for the SV as compared to the top 1,000 SNVs and indels, suggesting that the SV is more likely causal than the SNVs and indels. Within these 347 SV-linked-eQTLs, we observe a strong enrichment for genic variants (FET OR: 5.7, p-value <2.2e-16).

Finally, we dissected the eQTL signals based on the distance between the eQTL gene (eGene) and the eQTL variant (eVariant). We split the eVariants based on the genetic variant type they represent and calculated the distance between the eVariant and the eGene. For SVs, we took the midpoint of the SV as its genomic location. We observe that the distance distribution for each of the variant classes is very comparable (Fig. S92). Though for SV-eQTLs they seem to be closer to the eGene; however, given only 2-3% of the eQTLs are driven by SVs, this difference was not significant. We observe both positive and negative effect sizes for the all eVariant types considered (SNV, indel [1 base], indel-DEL, indel-INS, SV-DEL, SV-INS) (Fig. S93). The strongest effect sizes are observed for SNVs, but on average the absolute effect sizes per class are highly similar (SV: 0.33, indel: 0.32, SNV: 0.32). We observe for all deletions (INS-DEL & SV-DEL) that the median effect size is negative (-0.98 lead based, -0.09 all eVariants). However, when looking at only SV-DEL eVariants, we observe that the median effect is positive (0.45 lead based, 0.11 all eVariants). This is most likely driven by the low numbers of eVariants in these groups and the different genomic locations that are associated with these eQTLs.

21.2 sQTL analysis

(*Contributors: Marc Jan Bonder and Yang Li*)

The sQTL mapping procedure is matched to the eQTL mapping. In short, we selected the number of covariates (PCs) to correct sQTL mapping for in a similar way as described for eQTL mapping above. We focused on splice junctions on chromosome 2 and SNPs (MAF>10%); we varied the number of factors between 0 and 10 and a step-size of 1. We included the first five PCs for our sQTL map (Fig. S94). Next, we performed the full cis-sQTL mapping, again in line with the eQTL mapping described above. For the sQTLs we linked all splice junctions in a splice cluster to the same genetic variants, for this we selected the outermost positions of the splice cluster and extended this with a window of 1 Mbp centered around the splice junction cluster. All variants in this window with an MAF of 1% and higher and at HWE (HWE exact test p-value ≥ 0.0001) were tested for sQTL. To correct for multiple testing we used an approach in line with the eQTL mapping; however, given that splicing ratios within a cluster are correlated, we do multiple testing correction over all splice junctions within a cluster at once. We do so by taking the top p-values for each permutation over the cluster members and fitting the beta approximation once overall cluster members. Afterwards, we use Storey's Q-value procedure over the top splice junction per cluster, correcting for the number of clusters tested (for further details on this approach, see (*188*)).

In the sQTL mapping analysis we considered 141,565 junctions from 36,100 junction clusters linked to 11,277 genes. We identified 14,324 lead cis-sQTLs using an FDR of 5% and a total of 850,479 cis-sQTLs. Of the lead sQTLs, 13,040 were SNV-sQTLs (9,408 unique SNVs, 86.9% of the total cis-sQTL signals), 1,219 are indel-sQTLs (901 unique indels, length ≤49 bp, 18.6% of total) and 65 are SV-eQTLs (47 unique SVs, length ≥50 bp, 0.99%) (Table S32). As with the eQTLs, we find an enrichment for SVs linked to splicing ratios (FET OR = 1.20, p-value <1.6-4), as well as indels (FET OR = 1.18, p-value <2.2-16), and a depletion for SNPs (FET OR = 0.84, p-value <2.2-16) (Table S33). We observe that SVs can have a big impact on splicing, by for instance deleting exons. For example, we identified two large SVs (2.7 kbp and 6 kbp) each overlapping two exons in *GBP3* (Fig. S95) and *ZNF718*, respectively. As predicted, transcripts in homozygotes for the deletion allele fully skip the two deleted protein-coding exons, providing a simple yet clear mechanism for inter-individual variation in splicing. In another example, we identified a polymorphic AluY element in the intron of TMEM156 as associated with RNA splicing of the same intron. We also identified SV-sQTLs for SVs that were not previously characterized, including a 98 bp SV that is associated with RNA splicing of HEIH. In all four examples, the SV-sQTLs were also associated with the expression levels of the same gene, consistent with the idea that polymorphic SVs play an important role in inter-individual variation in gene regulation.

As for the expression we assessed the heritability of RNA splicing attributed to SVs relative to the top 1,000 SNVs and indels per splice junction for junctions with at least one SV-sQTL (see above, Fig. S96). We find that for splicing, on average the top SV for splicing variation explains 5.5% of expression variability versus 29.6% for the top 1,000 SNVs and indels (Fig. S96). For splicing we

again see that the heritability-based approach highlights more SVs to be potentially causal as compared to the number of lead SV-sQTLs (159 vs. 65)—for these the heritability explained by the top SV is higher as compared to the top 1,000 SNVs/indels (out of 999 assessed, 15.9%). We observe that SVs seem to impact expression on average more as compared to splicing. The average heritability explained by SVs for gene expression is higher (6.6% gene expression vs. 5.5% splicing), and SVs are more often explaining a larger fraction of heritability in gene expression as compared to the top 1,000 SNVs and indels (23% of the eGenes vs. 15.9% for sQTL junctions).

21.3 GWAS intersection and enrichment analysis

(*Contributors: Junjie Chen, Chong Li, Xinghua Shi*)

We examined if PanGenie genotypes, eQTLs, and sQTLs were associated with human phenotypes or traits, by intersecting SNV, indel, and SV eQTLs/sQTLs with SNPs previously reported in genome-wide association studies (GWAS). We first aggregated GWAS summary statistics and generated a union GWAS set that contains known GWAS results (threshold p-value ≤ 1.0e-6) extracted from the GWAS Catalog (*49*, *192*), Pan-UKB project and UK Biobank (UKBB, Pan-UKB team, https://pan.ukbb.broadinstitute.org. 2020.)), and PhenoScanner V2 (*193*, *194*). Since GWAS SNPs in UKBB and PhenoScanner are based on GRCh37, we lifted their coordinates over to GRCh38 and mapped them to corresponding rsIDs using dbSNP version 151 (*195*). In total, the union GWAS set contains 1,887,371 autosomal SNPs associated with 6,528 traits (Fig. S97). For SNVs, we counted those variants that are in the union GWAS dataset, while for indels and SVs, we counted those indels/SVs that have at least 1 bp overlapping with any GWAS SNP in the union set. We found that 1,132,074 SNVs, 2,339 indels, and 3,564 SVs overlapped or intersected with GWAS signals. For eQTLs, we found 288,311 SNV eQTLs, 743 indel eQTLs, and 465 SV eQTLs overlapped or intersected with GWAS signals, while 417,562 SNV sQTLs, 1,014 indel sQTLs, and 667 SV sQTLs overlapped or intersected with GWAS signals (Table S53 and Fig. S98). Next, we assessed the LD between any SNV/indel/SV with known GWAS SNPs within a 1 Mbp window using Plink v1.90b6.10 (*186*). We identified 536,992 SNV eQTLs, 7,905 indel eQTLs, and 133 SV eQTLs that are in high LD ($r^2 \geq 0.8$) with GWAS SNPs (Table S53). For sQTLs, we identified 354,932 SNV sQTLs, 4,895 indel sQTLs, and 198 SV sQTLs that are in high LD ($r^2 \geq 0.8$) with GWAS SNPs (Table S53).

We observed that SV eQTLs and SV sQTLs are respectively enriched for GWAS signals compared with random subsets of SVs that were fed into the eQTL or sQTL analysis pipeline (permutation test, p-value < 1.0e-4, Fig. S99). We performed 10,000 random permutations, with each permutation sampling 2,099 random SVs, with the same number as in the SV-eQTL and -sQTL set, from the 146,426 SVs that were evaluated in our eQTL/sQTL analysis. When generating these random SVs to match the distributions of SV eQTLs/sQTLs, we stratified the sampling process so that we controlled for the chromosome sources, SV types, SV sizes, and distances to transcription start sites (TSS) and transcription end sites (TES), respectively. In particular, SV sizes were matched after logarithm transformation to the base 2 and rounded down, while the distances to TSS/TES were matched after dividing the distance by 1,000 and rounding

down. In each permutation, we counted the number of random SVs that overlapped with at least one GWAS SNP in the union GWAS set. We then estimated the statistical significance of the observed overlap of SV eQTL/sQTL with the union GWAS set, compared with the overlaps from these random permutations.

21.4 GWAS and QTL co-localization analysis

(*Contributors*: *Junjie Chen, Chong Li, Xinghua Shi*)

We conducted a GWAS and QTL co-localization analysis to identify human trait associations driven by QTLs that may indicate a molecular mechanism induced by genetic variants. We performed this analysis using a Summary-data-based Mendelian Randomization (SMR) test (*60*) based on summary statistics from our eQTL/sQTL analysis, respectively, combined with GWAS summary statistics extracted from the union GWAS set. More specifically, the SMR approach used genetic variants as instrumental variables to calculate two-step-at-least-square estimates and test for the causative effect of the expression level of a gene on a trait without confounding from non-genetic factors. Consequently, Chi-square tests were performed to evaluate the significance of the estimated causal effect of gene expression or splicing transcripts on an associated trait, and the Benjamini-Hochberg correction (*196*) was then used to control the multi-test correction. Those gene-trait pairs passed the SMR test implied a potential causal link between the eQTL/sQTL, gene, and the associated trait.

To perform the SMR test, we first mapped SNV/indel/SV eQTLs to GWAS SNPs if the eQTLs have any overlap with GWAS SNPs regarding their chromosomal locations. For each of the 5,305 eGenes with GWAS associations, we collected the summary statistics of SNV/indel/SV eQTLs and those of GWAS SNPs that are located within a 2 Mbp window centered around the midpoint position of the tested gene. We identified 5,296 eGenes that are associated with 3,976 traits that passed the SMR test (5% FDR; Table S54). Among these eGenes, 1,178 of them have SV eQTLs and 4,494 of them have indel eQTLs. Similarly, we performed an SMR test on sQTL results to assess GWAS associations mediated by sQTL that suggest another molecular mechanism for SNP-trait associations. With FDR of 5%, we identified 2,826 genes associated with 2,834 traits that passed the SMR test for sQTLs (Table S55). Among these genes, 568 of them have SV sQTLs associated with 1,534 traits, and 1,960 of them have indel sQTLs associated with 2,829 traits. These observations indicated that SV and indel eQTLs and sQTLs potentially drive the changes of gene expression that lead to variations in the associated traits.

22 Ancestry analysis

(*Contributors: Rebecca Serra Mari and PingHsun Hsieh*)

22.1 Local ancestry

Local ancestry inference using haplotype-phased assembly, RFMix, and an HMM. We leveraged the haplotype-phased, sequence-resolved assemblies to infer local ancestry along chromosomes using RFMix (*47*) and an HMM. Ancestry information is based on haplotypes from a predetermined set of reference populations. For our analysis, a reference population panel was assembled using published phased genomes from the 1000GP (*15*). To control for potential biases due to the inclusion of admixed individuals in the reference panel, we used ADMIXTURE (*197*) and chose less admixed samples from African (LWK, MSL, GWD, YRI, and ESN; n=472), European (CEU, GBR, FIN, IBS, and TSI; n=381), East Asian (CHB, CHS, JPT; n=185), and South Asian (ITU and STU; n=186) 1000GP samples. In addition, as part of our reference panel, we also included 19 Native American samples from the Simons Genome Diversity Project (*48*, *197*) that show little European ancestry. Note that for the inference of local ancestry in Puerto Rican individuals, we also specifically explored and set the reference panel to be African, European, and Native American populations given the recent demographic history of this population (*198*). To avoid inaccurate ancestry calls, we removed SNVs in known gaps, SDs, and heterochromatin, telomeric, and centromeric sequences (GRCh38) from this analysis.

In short, RFMix segments haplotypes from a sample into windows of SNVs and uses a model of random forests based on conditional random fields to determine the ancestry of each genomic window. Analysis was performed with the following flags: `-G 15 -e 5 -w 0.4 -n 5 -c 0.2 -s 0.2 --rf-minimum-snps=100 --reanalyze-reference`. Note that we chose a node size of 5 to reduce bias in random forests resulting from unbalanced reference panel sizes as suggested in previous studies. We also developed an HMM-based method to infer the local ancestry of each genomic region of the haplotype-resolved assemblies (Fig. S100; https://github.com/rebeccaserramari/kyoshi (*88*)). Briefly, the computation of the most likely ancestral population for each genomic position is inferred through an HMM that models the reference haplotypes from a reference panel as hidden state sequences and the target haplotype as the observed sequence and computes a standard forward-backward algorithm on this HMM. At each variant position, the HMM penalizes two types of events: 1) discrepancies between target and reference haplotype and 2) switches to a different reference haplotype happening between two variant positions. As a consequence, the computed probability will be high for reference haplotypes that largely coincide with the target. The outcome of this method is a set of probabilities for underlying ancestries at each variant position on a haplotype. The probabilities of reference samples that belong to the same superpopulation are then added up to get a likelihood for every superpopulation at each position.

We restricted our downstream analysis to a set of high-confidence ancestry markers where the HMM gives evidence for a single ancestry with a probability >90%. To construct the final ancestry callsets for each individual haplotype, we first computed the concordant calls between the RFMix

and HMM results. Because the HMM results provide a per-variant ancestry resolution, we re-called the ancestry of a discordant region if there are more than 10 consecutive high-confidence ancestry markers.

To explore the quality of our ancestry inference, we performed the analysis on three trio families where we have both HiFi and CLR data using the parameters described above. In general, our results show a high concordance for ancestry calls between the two datasets (>90%), indicating that the input data type has little impact on the inferred local ancestries (Fig. S32). Our analysis showed that while the ancestry distribution of European and East Asian as well as most of African samples are relatively homogenous, there are clear admixture signals in other continental population samples (Figs. S33 and S34). Consistent with the history of slave trade from Africa to North America, the two African American samples (HG02011 and NA19983) carry large segments with European ancestry (>17%). In addition, the four Admixed American samples all carry long-stretch ancestry blocks from African (2-5%), American (6-13%), and European (83-89%) populations, reflecting the colonial history of these populations in America (*49*). Notably, comparing with the Strand-seq inferred recombination breakpoints (*1*) in the parental chromosomes in the trios, we found that ancestry blocks on the paternal and maternal haplotypes of the child switch at the locations of the crossing-over events in the parental chromosomes (Fig. 6B). This highlights the importance of these haplotype-resolved assemblies as a resource for population genetics and evolutionary analyses.

22.2 Variant age estimations

We estimate the age of an SV of interest using the software Relate (*199*). In short, Relate reconstructs the local genealogy of the region of interest using a scalable computation, which guarantees the inferred genealogy exactly producing the observed data. We include SNVs from the 500 kbp sequences flanking individual candidate SVs and remove/mask sites that have missing genotypes. In addition to the SV of interest, to approximate the age of the SV of interest, we also select a focal SNV that is in complete LD (r^2=1) to estimate variant age. The mutations are then mapped onto the branches of the resulting local tree using Relate to estimate mutation age.

22.3 Ancestral state determination from primate assemblies

We categorize SVs as ancestral if they were matched to calls in two nonhuman primates (9,829) (*200*, *201*), non-ancestral if they were not seen in both (72,641), and unknown if they could not be confidently assigned (25,120) because they were uncallable (14,865) or differed (10,255) with the latter, dominated by tandem repeats (91%). The allele frequency differed significantly with an average of 0.40 for ancestral and 0.12 for non-ancestral SVs (p-value < 1e-15, t-test). These polymorphic ancestral variants may be a useful resource for studying recurrence. We observe a significant difference between the numbers of insertions and deletions between ancestral and non-ancestral states with more ancestral deletions and non-ancestral insertions (p-value < 1e-15, Fisher's exact test [FET]). Reference collapses are known to enrich for fixed insertions (*4*), and we find SV insertions are more likely to be fixed (AF = 1 by PanGenie) than deletions (23% vs.

4.6%, p-value = 5×10^{-165}, FET), but only 0.88% of insertions and 0.33% of deletions are fixed, which is significantly lower compared to ancestral SVs (p-value < 1e-15, FET). In total, 1,301 SVs are homozygous for the alternative allele in all 3,202 samples and also exhibit AF = 1 in the assembly-based callset, indicating cases where GRCh38 contains errors or includes alleles with very low AF, with the expected bias toward insertions where the reference assembly likely collapsed (1,152 INS, 149 DEL). This is an increase from our previous work where we reported 507 such events (*4*).

22.4 Population stratification and population branch statistics

To identify variants stratified by population, we computed Fst values for each superpopulation (superpopulation vs. all other samples) and each sample (sample vs. all other samples). We assigned a variant to the population with the maximum Fst value if the value was at least 0.2 and greater than 0.1 from all other populations (Table S54). Similarly, we assigned the remaining variants to superpopulations with the maximum Fst value if the value was at least 0.2 and 0.1 higher than all other superpopulations (Fig. 6C).

To identify variants in regions subject to selection, we computed population branch statistics (PBS) (*51*) for combinations of populations for all SVs within 5 kbp of a gene. For each variant, we found the maximum PBS for each population using all possible combinations of an ingroup (same superpopulation) and an outgroup (different superpopulation) (Table S34). We focused specifically on the top 10 hits per superpopulation (Fig. 6D). As expected, there are distinct differences among population groups attributable to bottlenecks and expansions in their recent evolutionary history (Fig. S101).

23 Functional annotations

(*Contributor: Peter Audano*)

23.1 Functional variant annotations

Variant calls were intersected with RefSeq annotations (retrieved from UCSC on 2020-06-29) and counted the CDS, UTR, and intron sequence intersected by each variant. Variant calls were also intersected with *cis*-regulatory element (cCRE) annotations against GRCh38 from ENCODE (*22*) (Table S19). We categorized SVs as intersecting promoters (PLS), proximal and distal enhancer-like signatures (pELS and dELS), and CTCF (CTCF-only).

23.2 Triplet repeat expansions

We ran TRF (*175*) on SV sequences and searched for perfect repeat copies by comparing the repeat motif to the SV sequence and required that perfect repeats span at least 95% of the SV. We identified 286 repeat expansions and contractions by sample (382 alleles by haplotype) within these 106 sites (Tables S20 and S21).

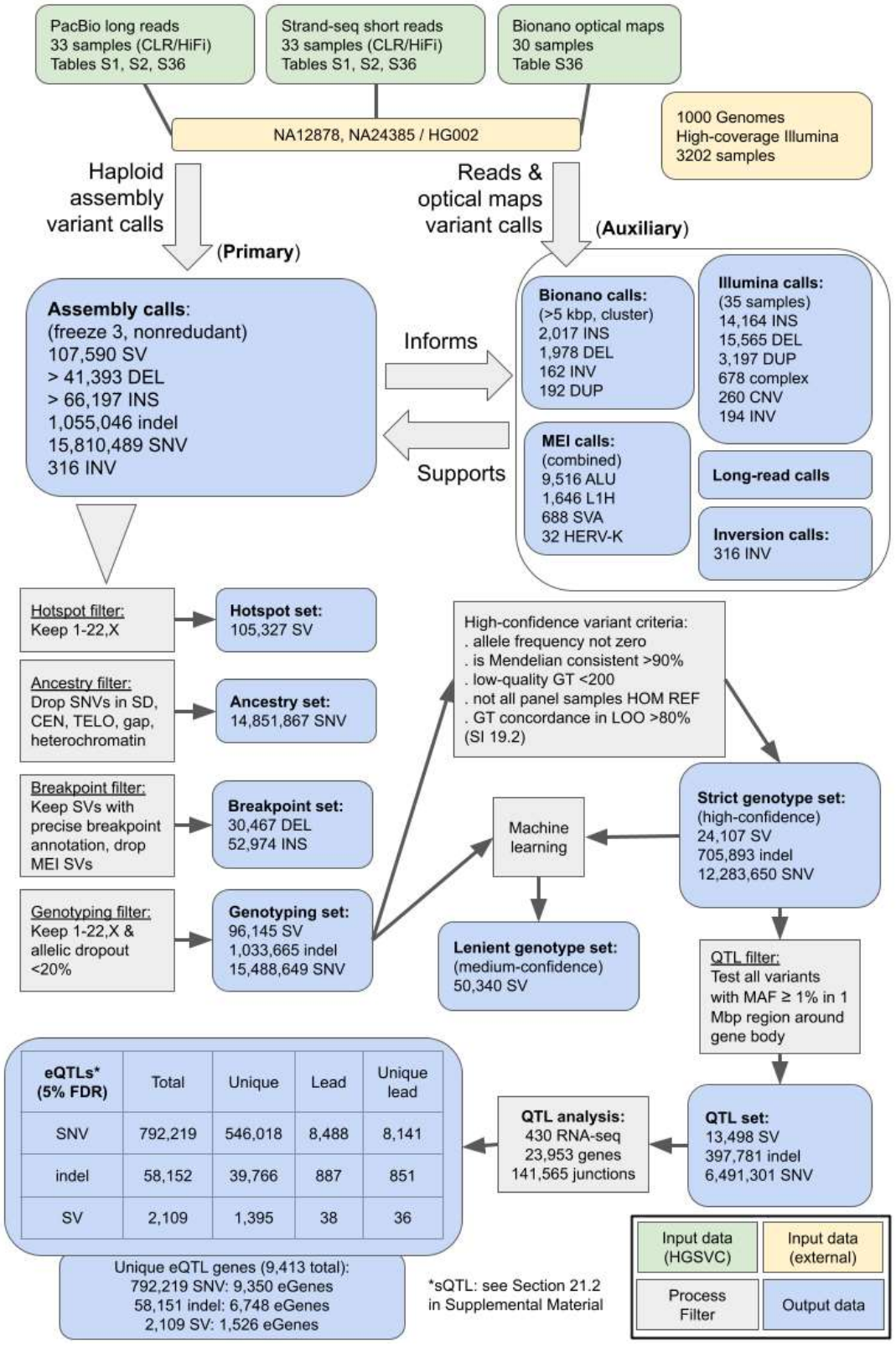

PacBio long reads
33 samples (CLR/HiFi)
Tables S1, S2, S36
Strand-seq short reads
33 samples (CLR/HiFi)
Tables S1, S2, S36
Bionano optical maps
30 samples
Table S36
NA12878, NA24385 / HG002
1000 Genomes
High-coverage Illumina
3202 samples
Haploid assembly variant calls
(Primary)
Reads & optical maps variant calls
(Auxiliary)
Assembly calls:
(freeze 3, nonredudant)
107,590 SV
> 41,393 DEL
> 66,197 INS
1,055,046 indel
15,810,489 SNV
316 INV
Informs
Supports
Bionano calls:
(>5 kbp, cluster)
2,017 INS
1,978 DEL
162 INV
192 DUP
Illumina calls:
(35 samples)
14,164 INS
15,565 DEL
3,197 DUP
678 complex
260 CNV
194 INV
MEI calls:
(combined)
9,516 ALU
1,646 L1H
688 SVA
32 HERV-K
Long-read calls
Inversion calls:
316 INV
Hotspot filter:
Keep 1-22,X
Hotspot set:
105,327 SV
Ancestry filter:
Drop SNVs in SD, CEN, TELO, gap, heterochromatin
Ancestry set:
14,851,867 SNV
Breakpoint filter:
Keep SVs with precise breakpoint annotation, drop MEI SVs
Breakpoint set:
30,467 DEL
52,974 INS
Genotyping filter:
Keep 1-22,X & allelic dropout <20%
Genotyping set:
96,145 SV
1,033,665 indel
15,488,649 SNV
High-confidence variant criteria:
. allele frequency not zero
. is Mendelian consistent >90%
. low-quality GT <200
. not all panel samples HOM REF
. GT concordance in LOO >80%
(SI 19.2)
Machine learning
Strict genotype set:
(high-confidence)
24,107 SV
705,893 indel
12,283,650 SNV
Lenient genotype set:
(medium-confidence)
50,340 SV
QTL filter:
Test all variants with MAF ≥ 1% in 1 Mbp region around gene body
QTL set:
13,498 SV
397,781 indel
6,491,301 SNV
QTL analysis:
430 RNA-seq
23,953 genes
141,565 junctions
eQTLs* (5% FDR) | Total | Unique | Lead | Unique lead
SNV | 792,219 | 546,018 | 8,488 | 8,141
indel | 58,152 | 39,766 | 887 | 851
SV | 2,109 | 1,395 | 38 | 36
Unique eQTL genes (9,413 total):
792,219 SNV: 9,350 eGenes
58,151 indel: 6,748 eGenes
2,109 SV: 1,526 eGenes
*sQTL: see Section 21.2 in Supplemental Material
Input data (HGSVC)
Input data (external)
Process Filter
Output data

Fig. S1 Data flowchart (caption next page)

Fig. S1. Data flowchart

Schematic of the data flow in our study focusing on variant callsets (blue boxes) and necessary processing and filtering steps (gray boxes) to derive reduced callsets for the various analyses downstream of the haploid assembly variant calling (blue box “Assembly calls”). Green boxes: input datasets created as part of this study; yellow boxes: external datasets. Results for sQTL analysis omitted for layout reasons, please see Section 21.2 in the Supplemental Material.

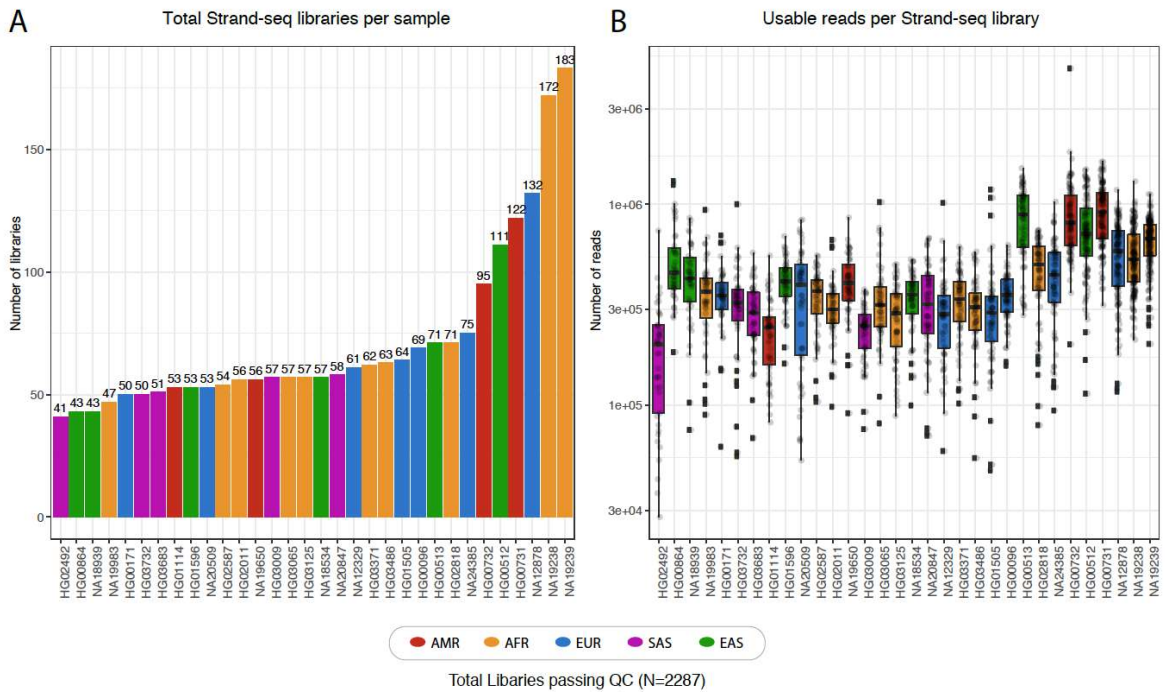


Fig. S2. Summary of Strand-seq libraries used in this study

(A) Bars indicate the total number of high-quality Strand-seq libraries (N=2,287) selected for analysis, per sample. (B) Box plots show the range of usable sequencing fragments per high-quality library. Graphs are colored by superpopulation; asterisks mark published datasets included in this analysis.

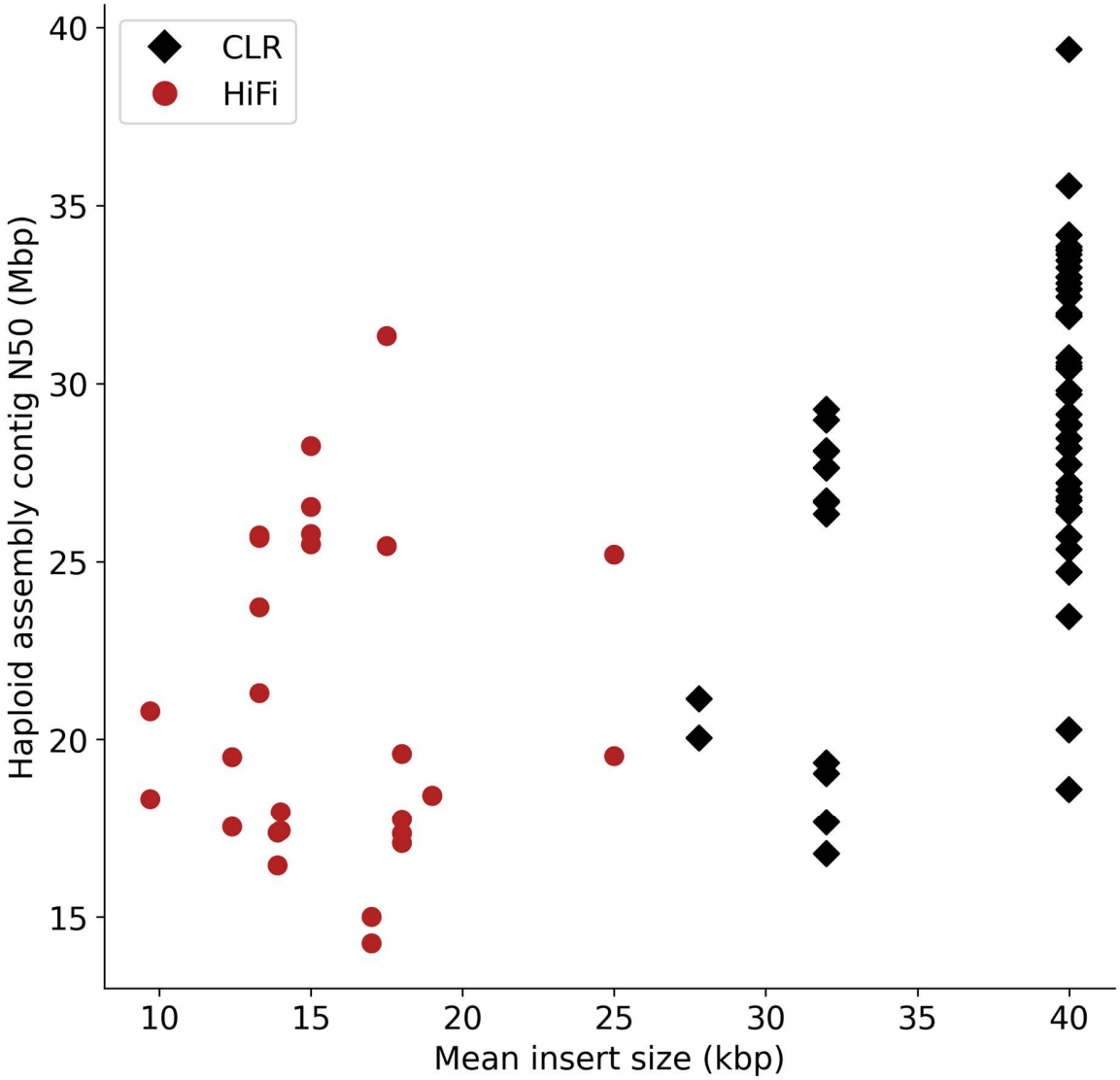


Fig. S3. Haploid assembly contig N50 as a function of library insert size

Relation between long-read sequencing library insert size averaged over all SMRT cells per sample (x-axis) and haploid assembly contig N50 (y-axis) plotted for CLR (black diamonds) and HiFi (red circles) samples.

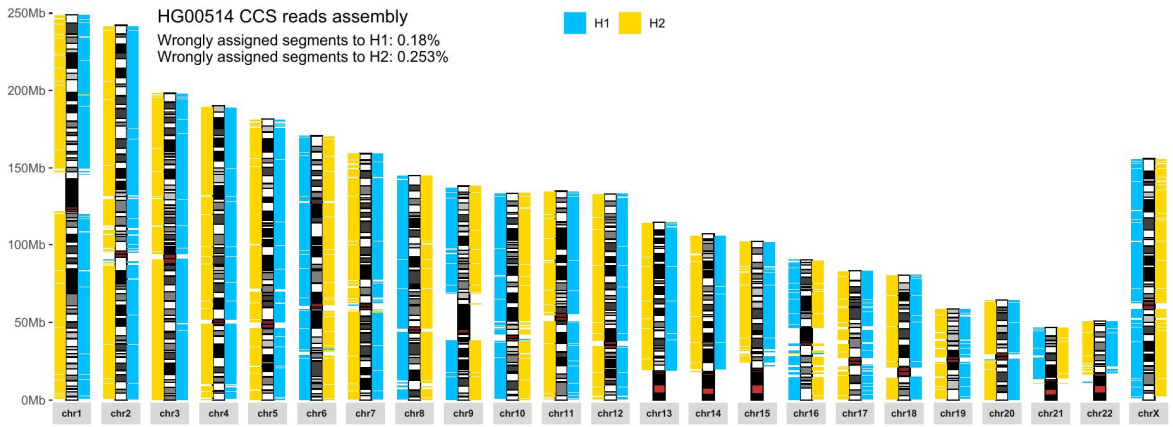


Fig. S4. Phasing accuracy for child HG00514 (HiFi)

The phased HiFi assembly for sample HG00514 is divided into 1 Mbp blocks sequence, which are aligned to GRCh38 and colored by parental haplotype (H1 blue, H2 yellow) based on a trio-phased set of reference SNVs for this individual. The fraction of blocks assigned to the wrong haplotype is indicated at the top.

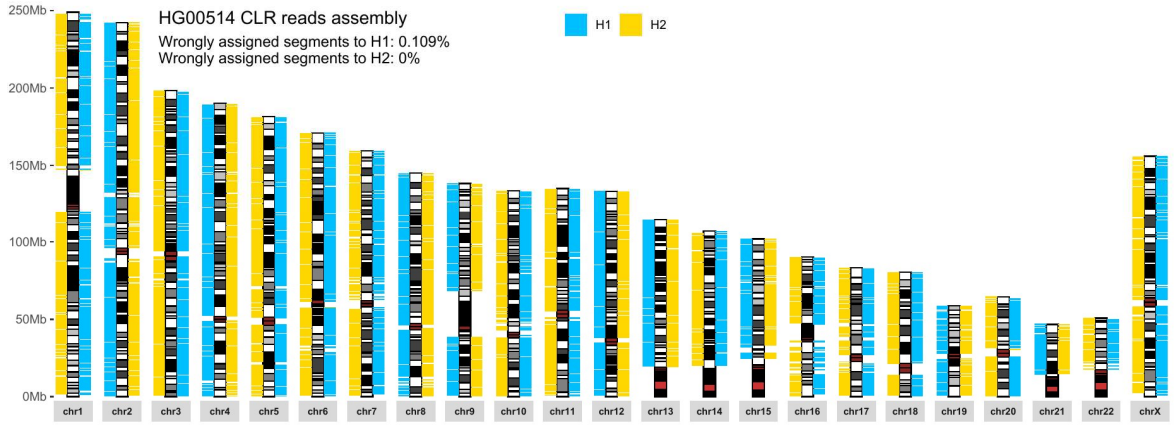


Fig. S5. Phasing accuracy for child HG00514 (CLR)

The phased CLR assembly for sample HG00514 is divided into 1 Mbp blocks, which are aligned to GRCh38 and colored by parental haplotype (H1 blue, H2 yellow) based on a trio-phased set of reference SNVs for this individual. The fraction of blocks assigned to the wrong haplotype is indicated at the top.

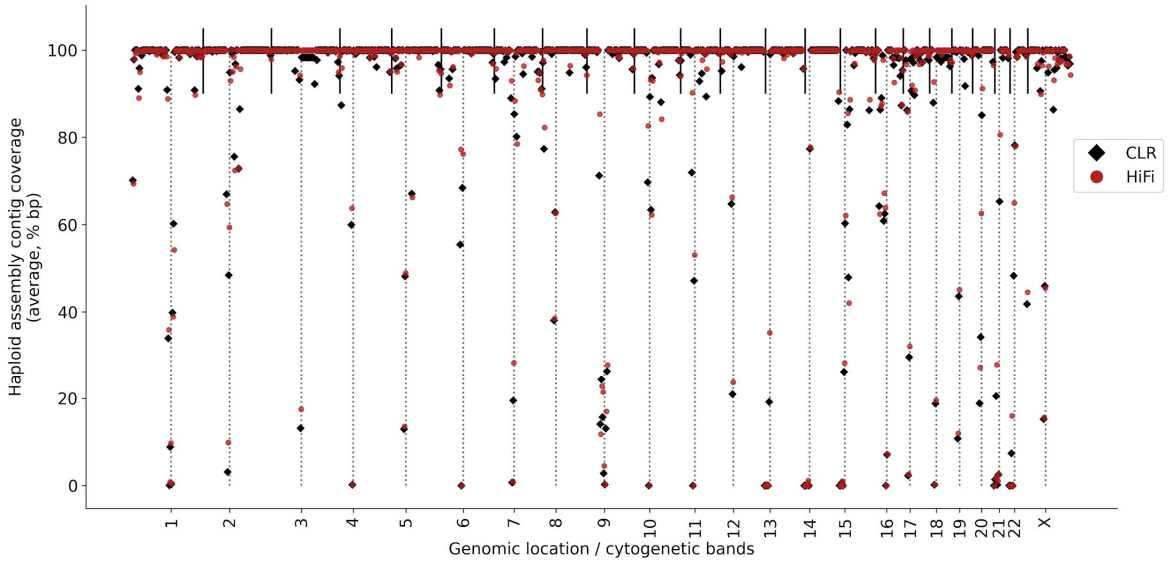


Fig. S6. Haploid assembly contig coverage relative to GRCh38

Cytogenetic bands are plotted with fixed width along the x-axis for chromosomes 1-22 and X. The contig coverage (y-axis, % covered bp in region) is depicted as an average over all CLR (black diamonds) and HiFi (red circles) haplotype assemblies. Centromere locations are indicated as gray dotted lines, and chromosome boundaries as black vertical lines at the top.

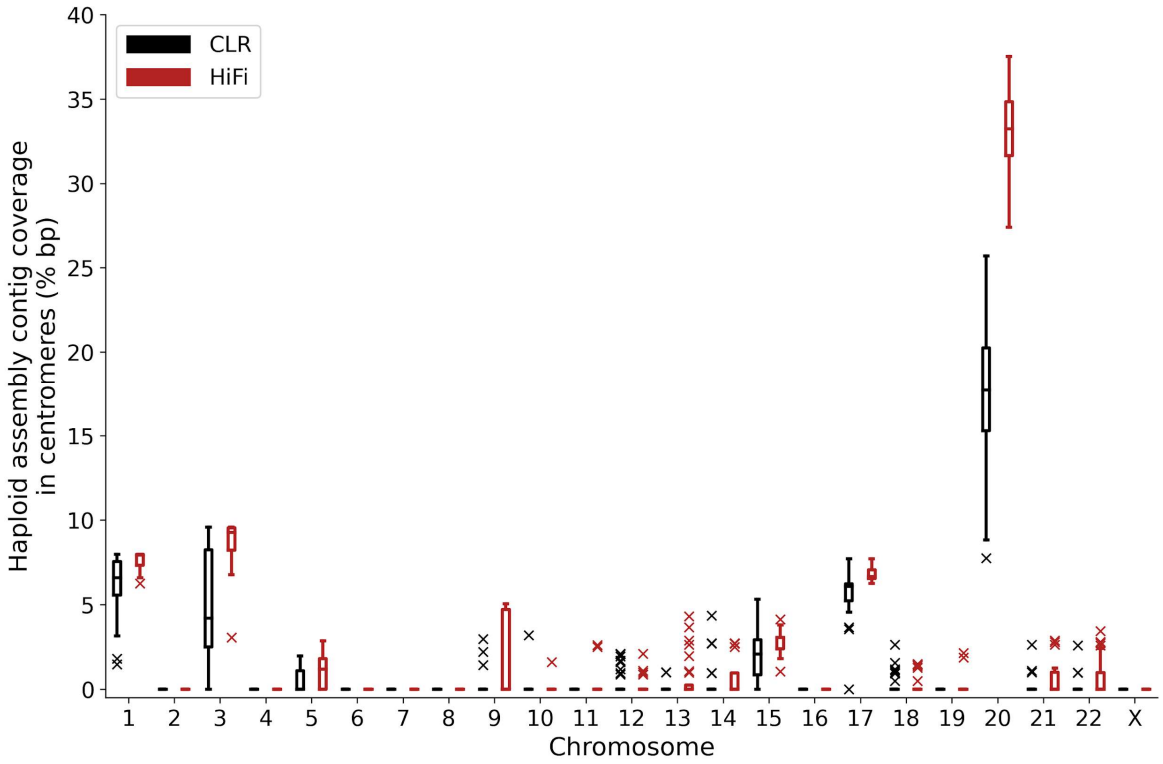


Fig. S7. Haploid assembly contig coverage in GRCh38 centromeres

Variation in haploid assembly contig coverage in centromeres for chromosomes 1-22 and X (x-axis) is depicted separately for CLR (black) and HiFi (red) assemblies. The symbol “x” marks outliers outside of 1.5X interquartile range.

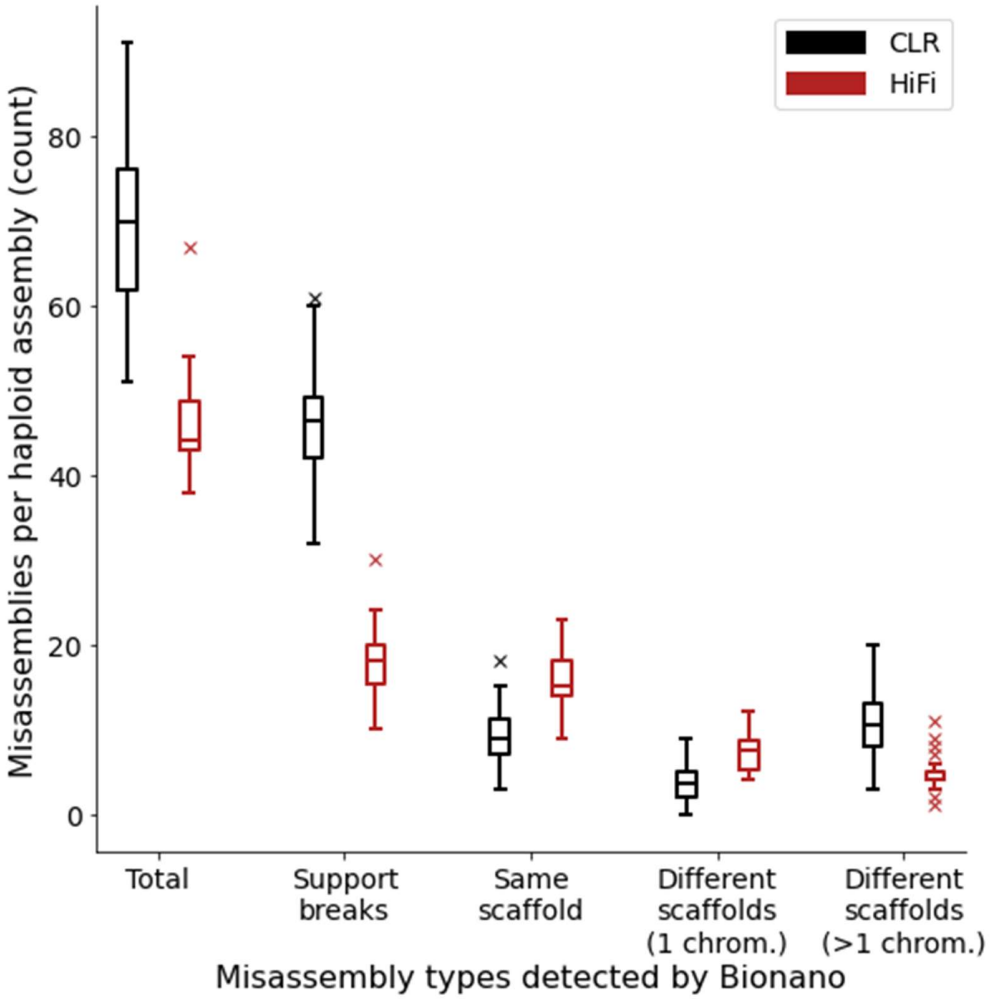


Fig. S8. Misassemblies detected via Bionano hybrid scaffolding

Contig-level phased assemblies were combined with Bionano optical maps to create scaffolded hybrid assemblies. Misassemblies detected per haploid assembly (x-axis, “Total”) were characterized based on the type of misassembly (x-axis, left to right). Part of contig has no Bionano support; contig is fragmented within scaffold; contig is fragmented between different scaffolds on the same chromosome; contig is fragmented between different scaffolds on different chromosomes (“chimerism”).

Fig. S9. Alignments of unscaffolded contig sequence

Unscaffolded contig sequence was split into 500 bp reads and aligned to GRCh38. Read alignment coverage was aggregated in bins of 500 kbp and are plotted as an average for all HiFi (red track) and all CLR (black track) assemblies.

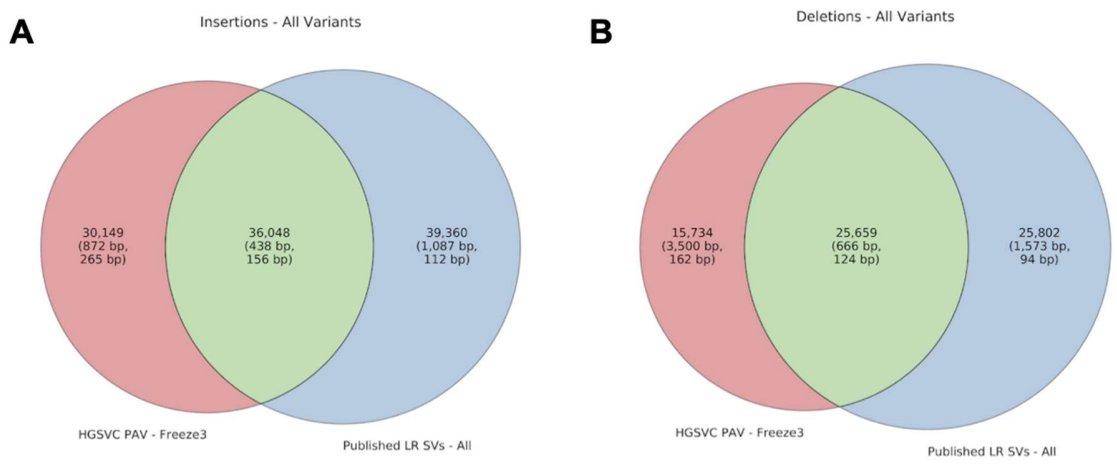
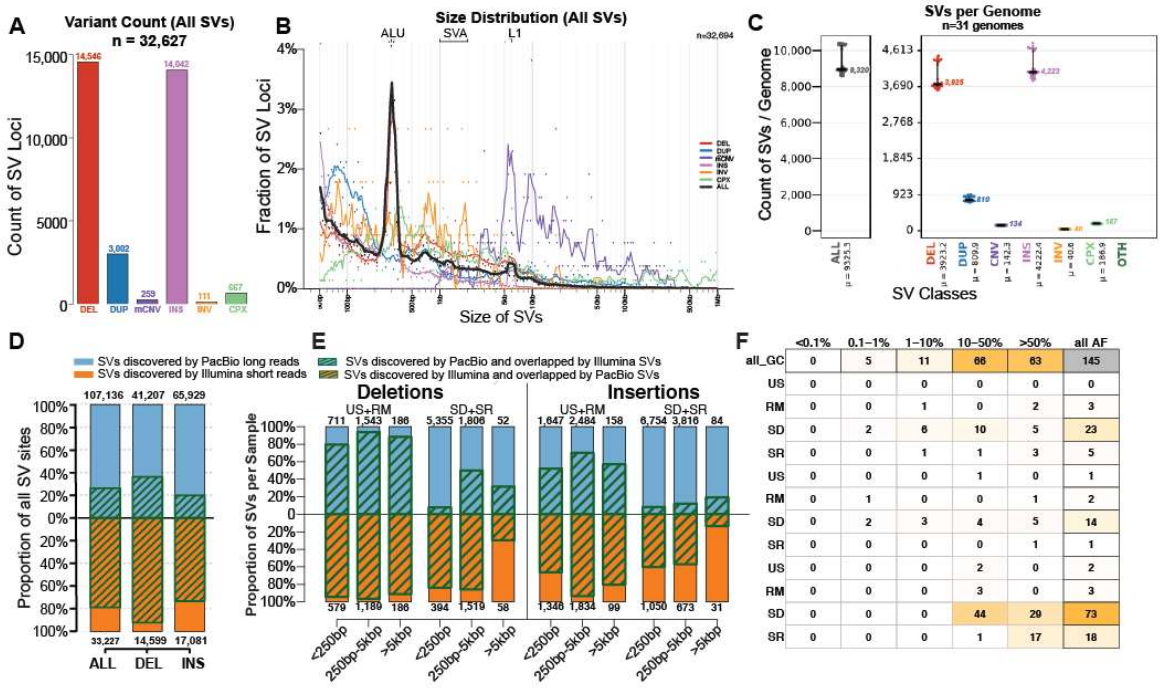


Fig. S10. New sequence-resolved SVs

Merged SVs in red are compared to SVs published in five other long-read studies in blue. (A) Insertions - All Variants, (B) Deletions - All Variants.



	<0.1%	0.1–1%	1–10%	10–50%	>50%	all AF
all_GC	0	5	11	66	63	145
US	0	0	0	0	0	0
RM	0	0	1	0	2	3
SD	0	2	6	10	5	23
SR	0	0	1	1	3	5
US	0	0	0	1	0	1
RM	0	1	0	0	1	2
SD	0	2	3	4	5	14
SR	0	0	0	0	1	1
US	0	0	0	2	0	2
RM	0	0	0	3	0	3
SD	0	0	0	44	29	73
SR	0	0	0	1	17	18

Fig. S11. Overview of the Illumina short-read integration callset on the 31 genomes with matched PacBio sequences and callsets

(A) Count of SV loci by variant type in the Illumina integration callset; (B) Size distribution of SV calls; (C) Count of SVs per sample by variant type; (D) Comparison of SV sites across 31 samples shared by PacBio and Illumina sequences; (E) Concordance of SVs between Illumina and PacBio by genomic location and SV sizes. US - unique sequences, RM - repeat masked regions, SD - segmental duplications, SR - simple repeats; (F) Distribution of large CNVs (>5 kbp) specifically discovered by Illumina short-read across genomic locations and ranges of allele frequencies. Abbreviations of genomic context are the same as in E.

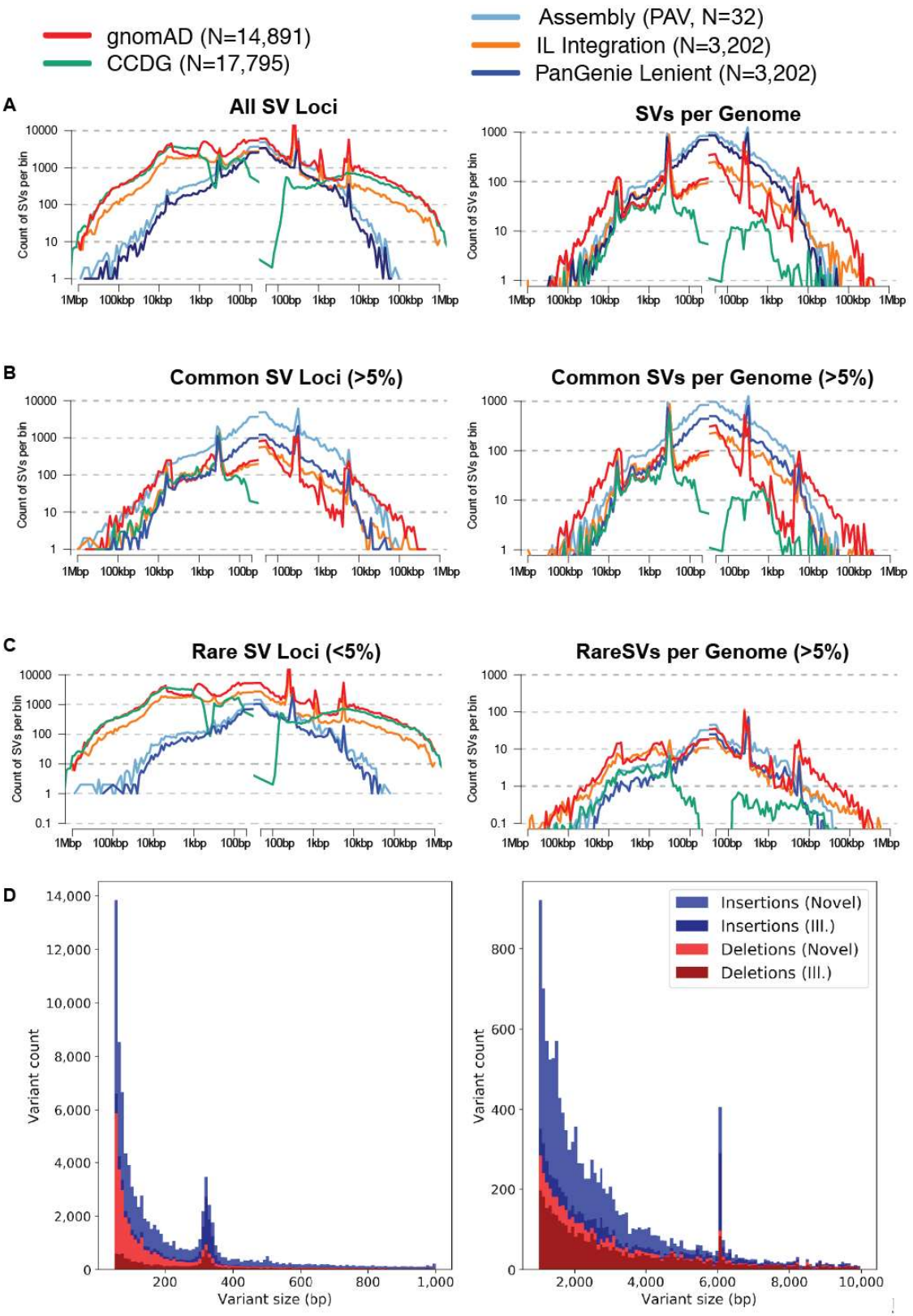


Fig. S12. Length distribution of SVs from HGSVC, gnomAD, and CCDG (caption next page)

Fig. S12. Length distribution of SVs from HGSVC, gnomAD and CCDG

(A) Length distribution of SVs from the HGSVC Illumina short-read integration callset (orange), PacBio long-read callset (light blue) and PanGenie genotypes (blue), gnomAD (red) and CCDG (green). Left panel shows the overall distribution of all SV loci, and the right panel shows the averaged distribution per genome. (B-C) Length distribution of common (>5% AF, B) and rare (<5% AF, C) SVs from the HGSVC Illumina short-read integration, PacBio long-read callset, PanGenie genotypes, gnomAD, and CCDG. Color and scheme are the same as in A. Length distribution of common SVs that are of 5% or higher allele frequencies from the HGSVC Illumina integration (orange), PacBio (light blue), PanGenie (blue), gnomAD, and CCDG. (C) Length distribution of rare SVs (<5% AF) from the HGSVC Illumina integration, PacBio. PanGenie, gnomAD, and CCDG. (D) Size distribution of insertions (blue) and deletions (red) stratified by Illumina support in a 1000GP deep-coverage callset for the same discovery sample (dark color) or not seen in the Illumina callset (light color). Most SV types and sizes are under-called by Illumina except for large deletions, SINEs (~300 bp), and LINEs (~6 kbp).

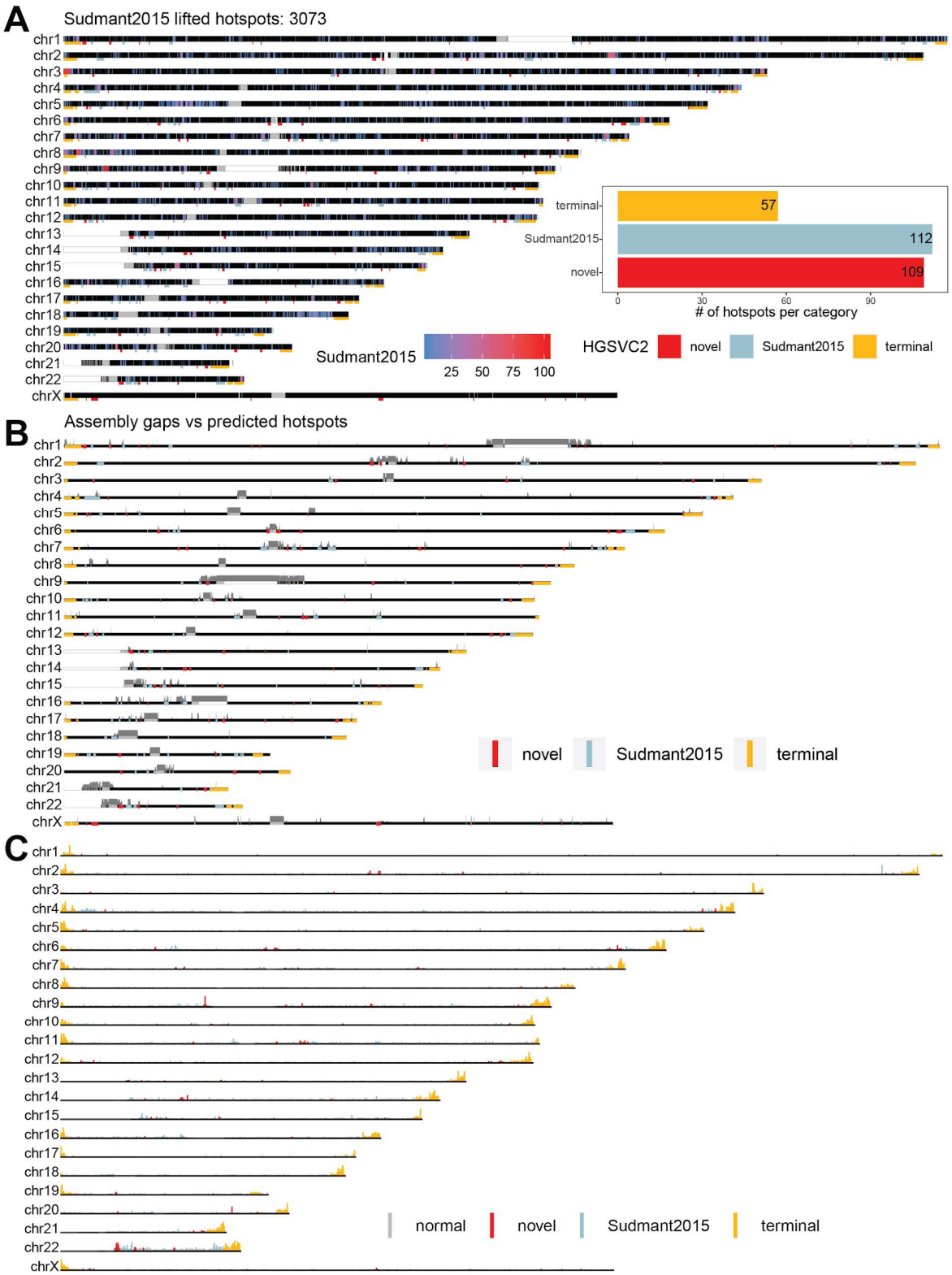

A
Sudmant2015 lifted hotspots: 3073
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
terminal
Sudmant2015
novel
57
112
109
0
30
60
90
# of hotspots per category
Sudmant2015
25 50 75 100
HGSVC2
novel
Sudmant2015
terminal
B
Assembly gaps vs predicted hotspots
novel
Sudmant2015
terminal
C
normal
novel
Sudmant2015
terminal

Fig. S13. Raw data supporting detected SV hotspots (caption next page)

Fig. S13. Raw data supporting detected SV hotspots

(A) Genome-wide distribution of SV hotspots detected by Sudmant et al. (*23*) (blue->red heatmap). SV hotspots are categorized into three groups: 'novel' - unique for this study, 'Sudmant2015' - overlapping with previous study, and 'terminal' - residing in the last 5 Mbp of each chromosome end. Inset: Shows a total count of hotpots in each previously defined category. (B) Gray bar along each ideogram shows assembly gaps, across all 64 assembled phased genomes, as regions where contigs map with <60 mapping quality. Underneath each ideogram, we plot a position of each hotspot as a rectangle colored by previously defined hotspots category. (C) An ideogram showing the binned counts (200 kbp bin) of SVs along each chromosome. Each bin is colored based on the overlap with the previously defined hotspot category. Bins not overlapping any hotspot are shown in gray ('normal').

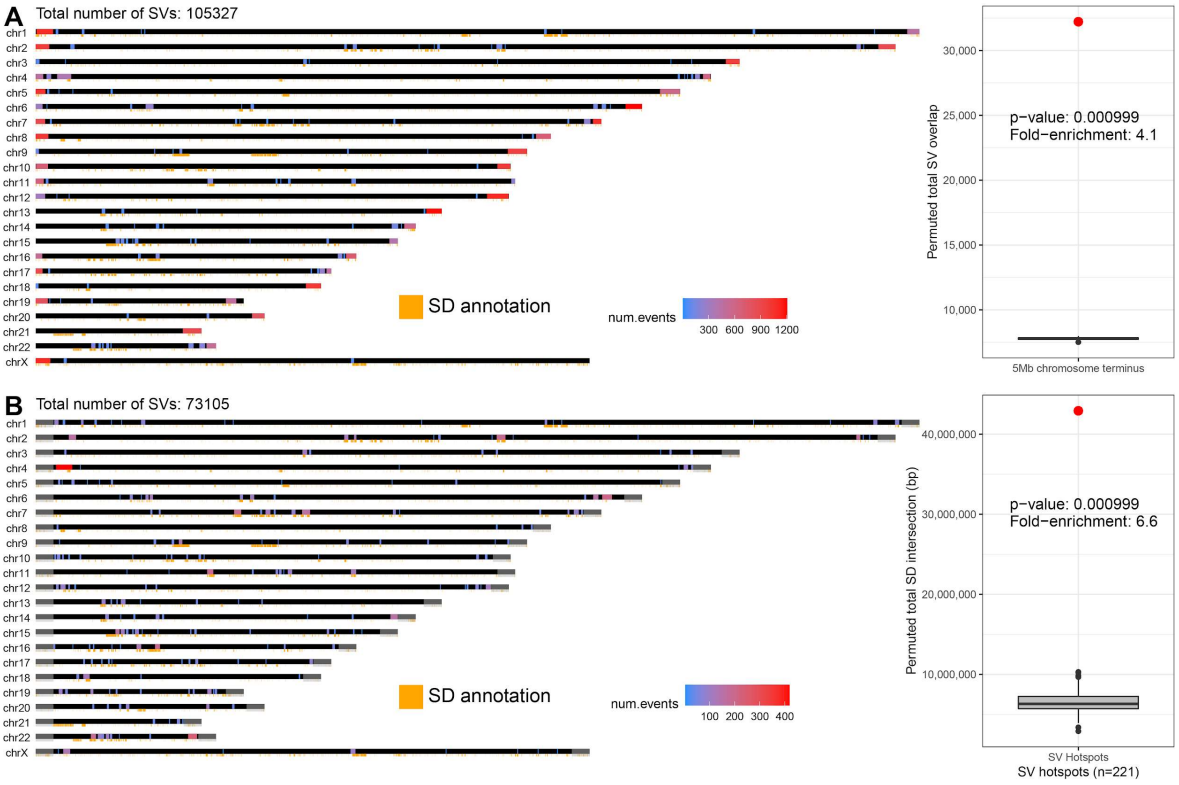


Fig. S14. SV hotspot detection and enrichment for chromosome ends and SDs

(A) An ideogram showing detected SV hotspots using all SVs (≥50 bp). The total number of SVs in each detected hotspot is shown by a scale going from blue to red. Positions of SDs are added at the bottom of each chromosome ideogram as an orange rectangle. Left: Enrichment analysis of SVs with respect to the last 5 Mbp of each chromosome end. The red dot represents the observed number of SVs at the 5 Mbp terminus while the box plot shows the distribution of SV counts at the 5 Mbp terminus, after 1000 random shuffling of SVs. (B) An ideogram showing detected SV hotspots using after filtering SV at the 5 Mbp terminus of each chromosome (highlighted by a gray rectangle at the end of each chromosome). The total number of SVs in each detected hotspot is shown by a scale going from blue to red. Positions of SDs are added at the bottom of each chromosome ideogram as an orange rectangle. Left: Enrichment analysis of SVs in respect to SD content. The red dot represents the observed number of SD bases intersecting with detected hotspots while the box plot shows the distribution of SD overlap after 1000 random shuffling of detected hotspots.

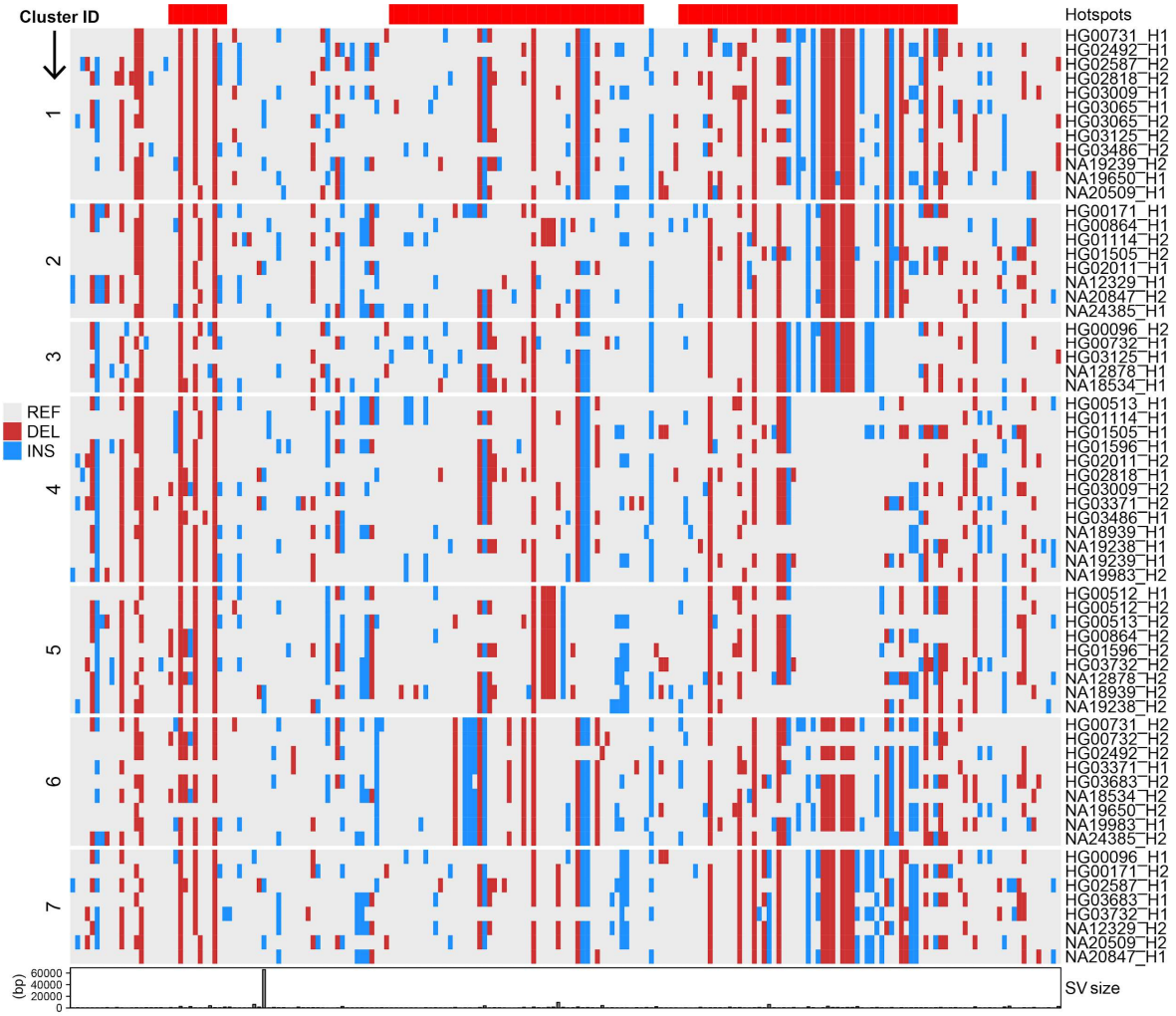


Fig. S15. Definition of distinct HLA haplotypes based on SV clustering

A heatmap of clustered SVs that defines seven distinct haplotypes where each column represents an SV and each row represents a unique haplotype (n=64) assembled over the HLA region. Haplotype clusters have been defined by k-means clustering after removal of SVs from regions where assemblies map with lower confidence to GRCh38. The final number of clusters was decided based on manual curation and separated by a white line and the cluster number (n=7). At the top of the heatmap, we point to the SVs that overlap with predicted SV hotspots. At the bottom of the heatmap, we show the size of each SV as a barplot.

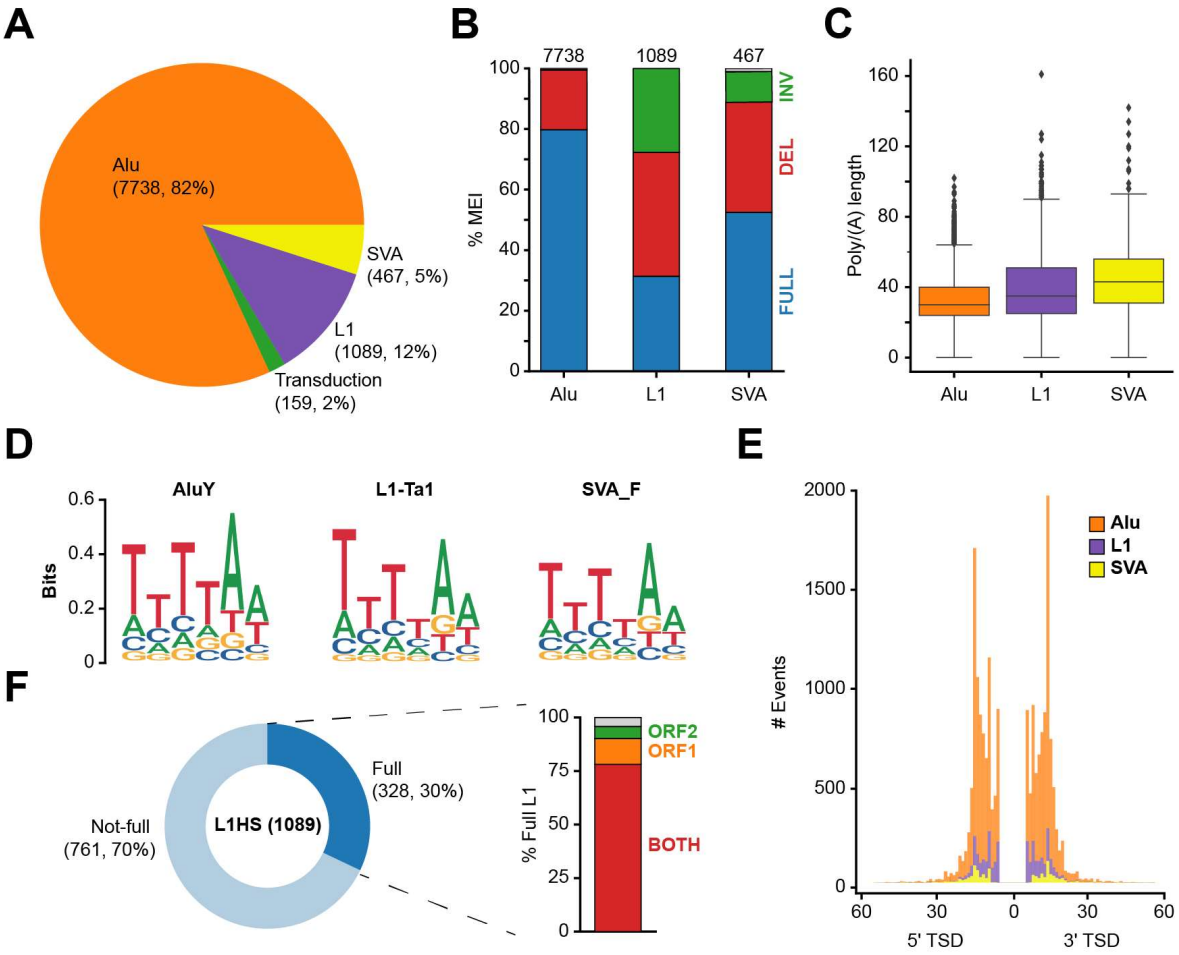


Fig. S16. Collection of sequence-resolved MEIs

(A) Number of MEIs per subfamily identified via PAV. (B) Percentage of events that are full length (blue), 5' deleted (red), 5' inverted (green), or with undetermined structure (gray) per subfamily. (C) Poly(A) tail length distributions. (D) Endonuclease cleavage site sequence logos for the youngest subfamilies (5' NNNN/NN 3'). (E) Target site duplication size distributions annotated using PALMER. (F) Left: donut chart showing the fraction of full-length L1s among all L1 insertions identified. Right: ORF status for full-length L1s displayed as red (intact ORFs), orange (truncated ORF1), green (truncated ORF2) and gray (both truncated) stacked bars.

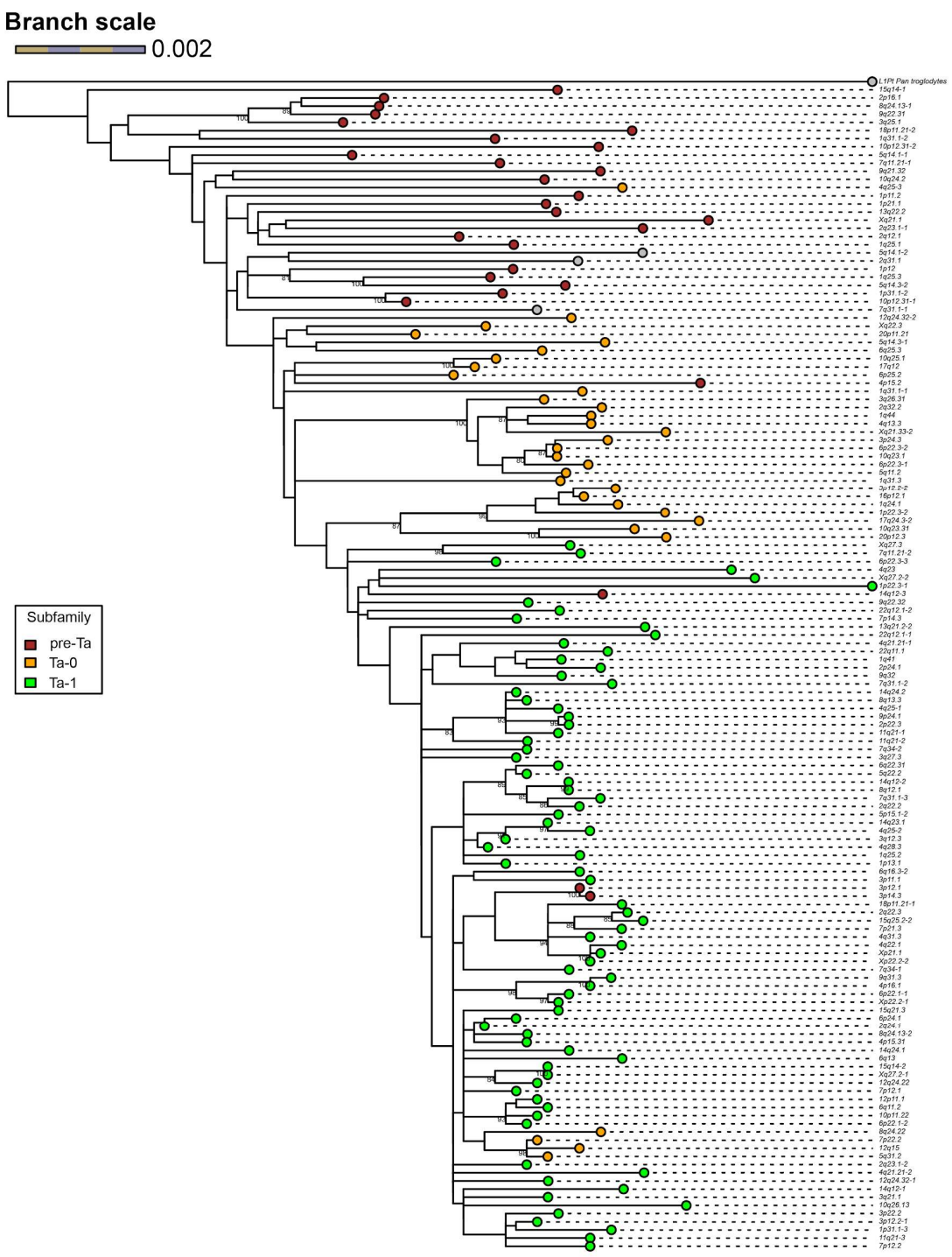

Branch scale
0.002
Subfamily
pre-Ta
Ta-0
Ta-1
L1Pt Pan troglodytes

Fig. S17. Complete phylogeny for all active sequence-resolved [...] (caption next page)

Fig. S17. Complete phylogeny for all active sequence-resolved full-length (FL)-L1s

Tree branch lengths are scaled according to the average number of substitutions per base position. Active L1s are named according to the cytoband where they reside. If multiple L1s are located over the same cytoband, a numerical index is appended to the cytoband identifier end. Subfamily assignments based on diagnostic nucleotides for each FL-L1 are indicated as colored circles on the tip of each branch. L1 *Pan troglodytes* (L1Pt) is included as an outgroup. Bootstrap support values ≥80% are indicated.

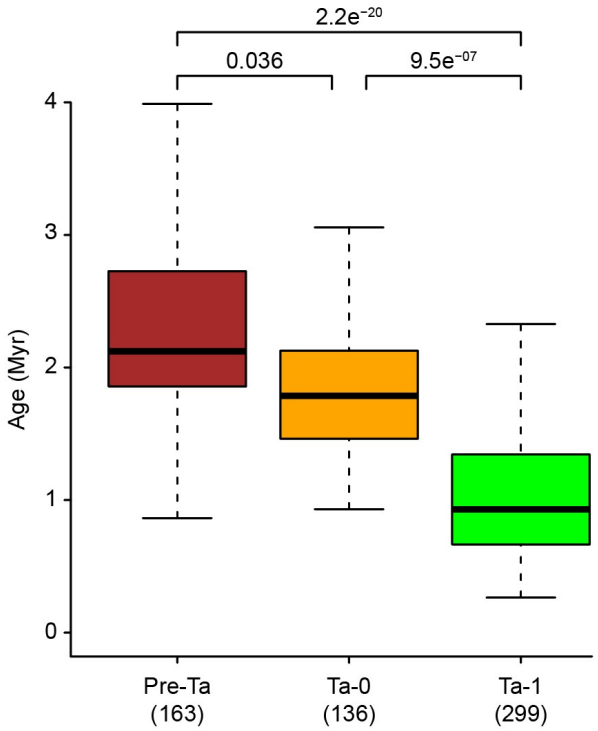


Fig. S18. Association between evolutionary age and subfamily

Age estimates in million years (Myr) for FL-L1 grouped according to subfamily. Subfamily assignments are based on diagnostic nucleotide positions. Subfamily distributions were compared pair-wise via a Student's t-test and the p-values are displayed for each comparison.

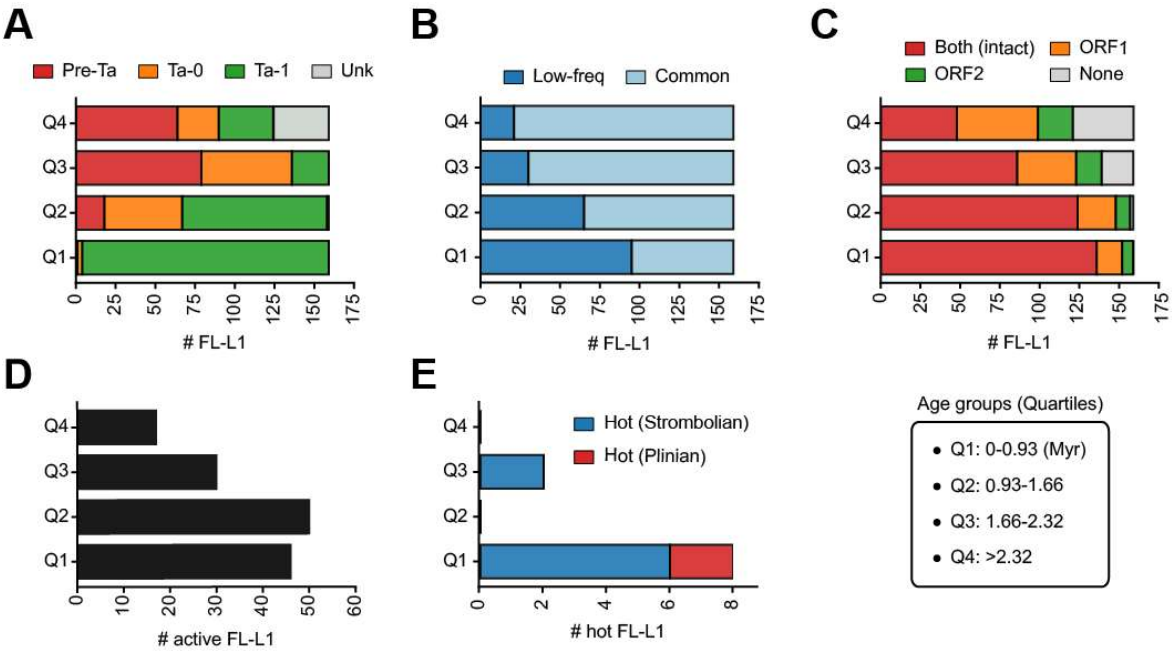


Fig. S19. Association between evolutionary age and FL-L1 features

(A) Number of FL-L1s belonging to each L1 subfamily at age quartiles (Q1-Q4). FL-L1s with no subfamily diagnostic nucleotides found are classified as unknown. (B) Number of low-frequency (<5% MAF) and common (≥5% MAF) FL-L1s per quartile. (C) Number of FL-L1s with both ORFs intact (Both), one ORF intact (ORF1 or ORF2), or both disrupted (None) per quartile. (D) Number of FL-L1s known to be active in the population, in cancer or in vitro per quartile. (E) Number of FL-L1s reported to be “hot” in cancer genomes per quartile. “Hot” elements classified according to the two activity patterns described in (*34*).

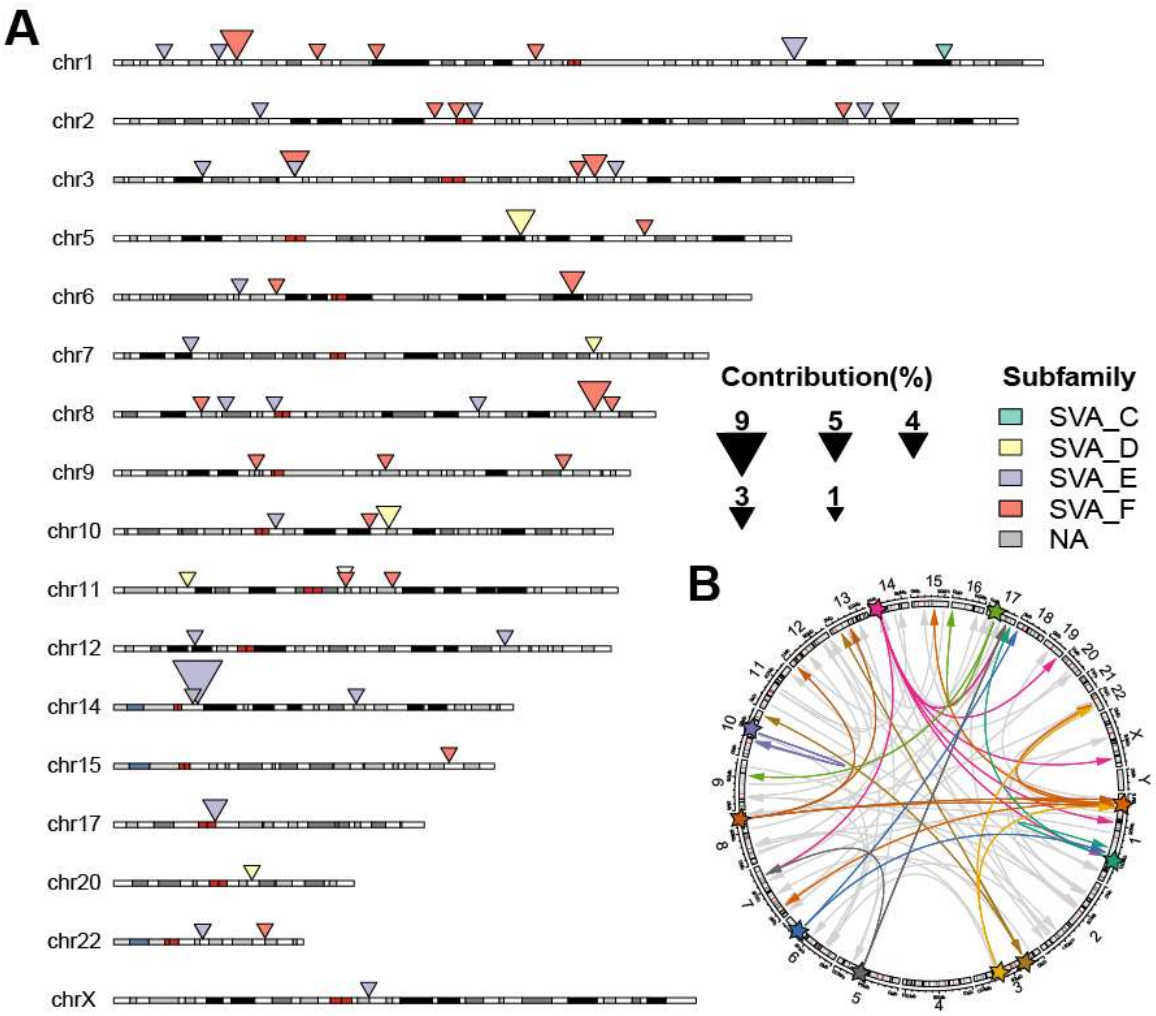


Fig. S20. Catalogue of source SVA active in HGSVC2 dataset

(A) Chromosomal map with germline SVA source SVA elements as triangles. Each triangle is color coded according to the source SVA subfamily. The contribution of each source locus (expressed as a percentage) to the total number of germline transductions identified is represented in a gradient of sizes, with top contributing elements exhibiting larger sizes. (B) Circos plot of SVA transductions and source SVA loci. Source elements mediating multiple transductions are highlighted with a star. Transductions derived from these copies are colored according to their source, while those derived from single-transduction source SVAs are in gray.

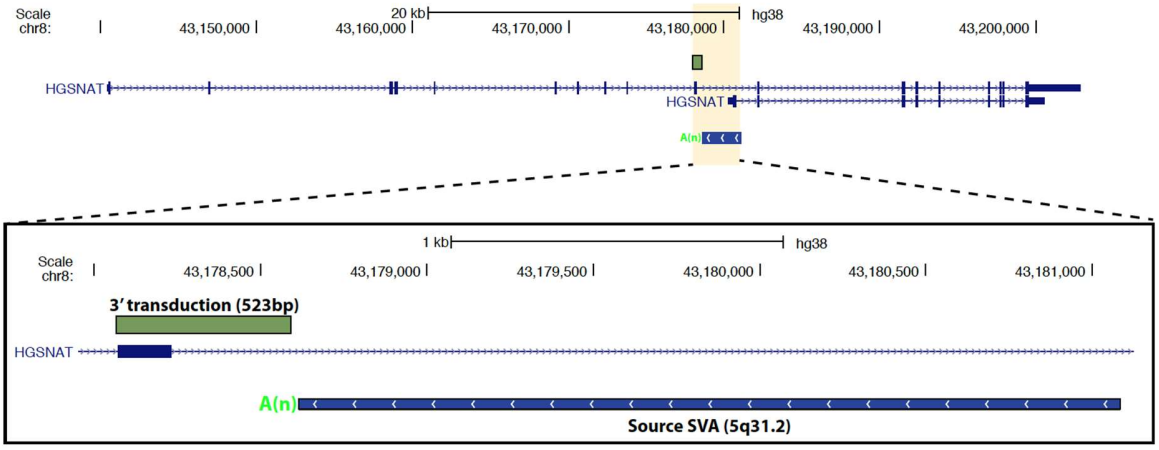


Fig. S21. Exon shuffling by intronic SVA source element

Partnered SVA-mediated 3′ transduction mobilizes complete exon from *HGSNAT* to another genomic location at chr5:138736711.

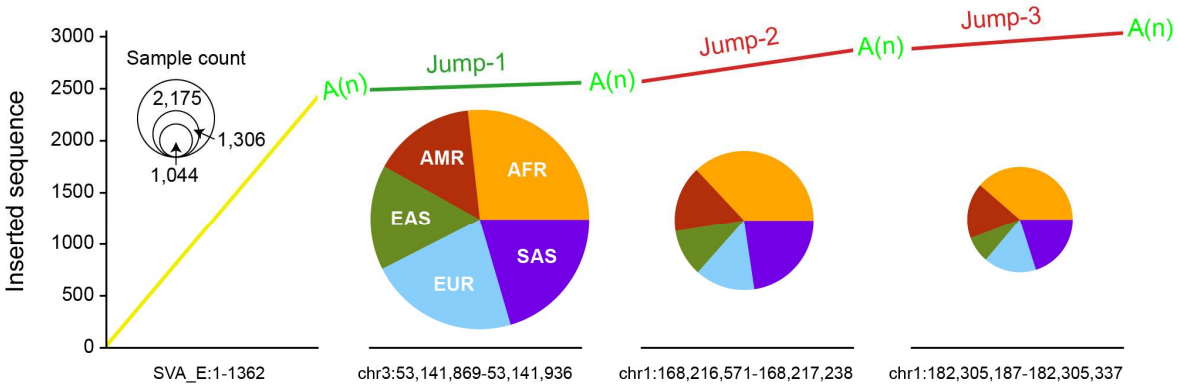


Fig. S22. Multi-transduction SVA insertion at 17q25.1 disentangled via assemblies

The configuration for the sequence-resolved nested transduction at 17q25.1 is depicted as a line plot. Alignment of the insert over the SVA_E consensus and three genomic loci at 3p, 1p (chr1:168216956), and 1p (chr1:182305187) are colored in yellow, green, red, and red, respectively. Poly(A) stretches between each transduction event are shown in dark green. Population distribution together with the average number of 1KG samples where k-mers derived from each transduced sequence are displayed as pie charts of different sizes.

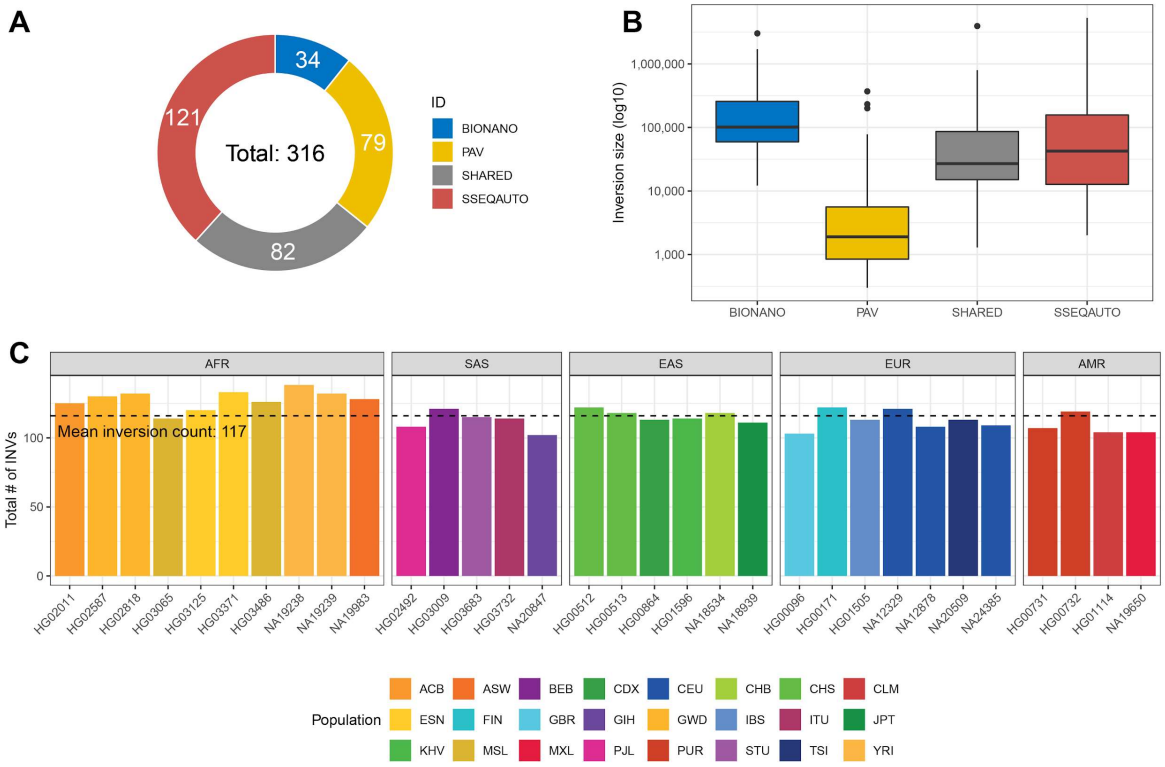


Fig. S23. Inversion callset summary (n=316)

(A) A donut plot showing the total number of inversions called exclusively by a certain technology ('PAV' - phased assembly inversion caller, 'BIONANO' - Bionano optical maps, and 'SSEQAUTO' - automated Strand-seq inversion calls) along with inversions called by at least two independent technologies ('SHARED') (Number of Strand-seq inversion calls per category.); (B) Size distribution of inversions calls per above-mentioned category; (C) A barplot showing the total number of inversions per sample. Dashed horizontal line shows a median inversion count across all samples (n=32).

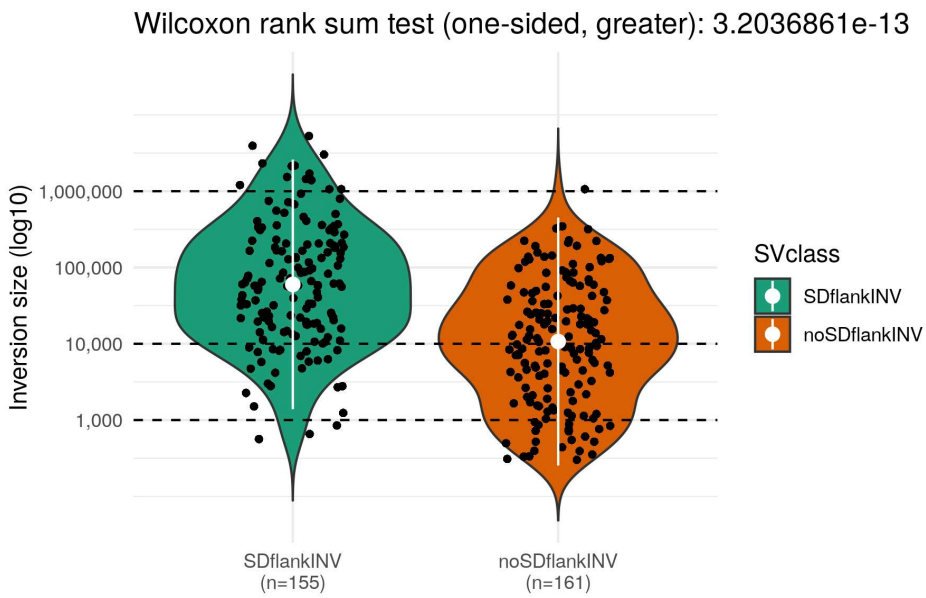


Fig. S24. Inversions flanked by SDs

Size distribution of inversions flanked by SDs (SDflankINV, n=155) and inversions not flanked by SDs (noSDflankINV, n=161). White dot shows the mean of each distribution along with IQR range.

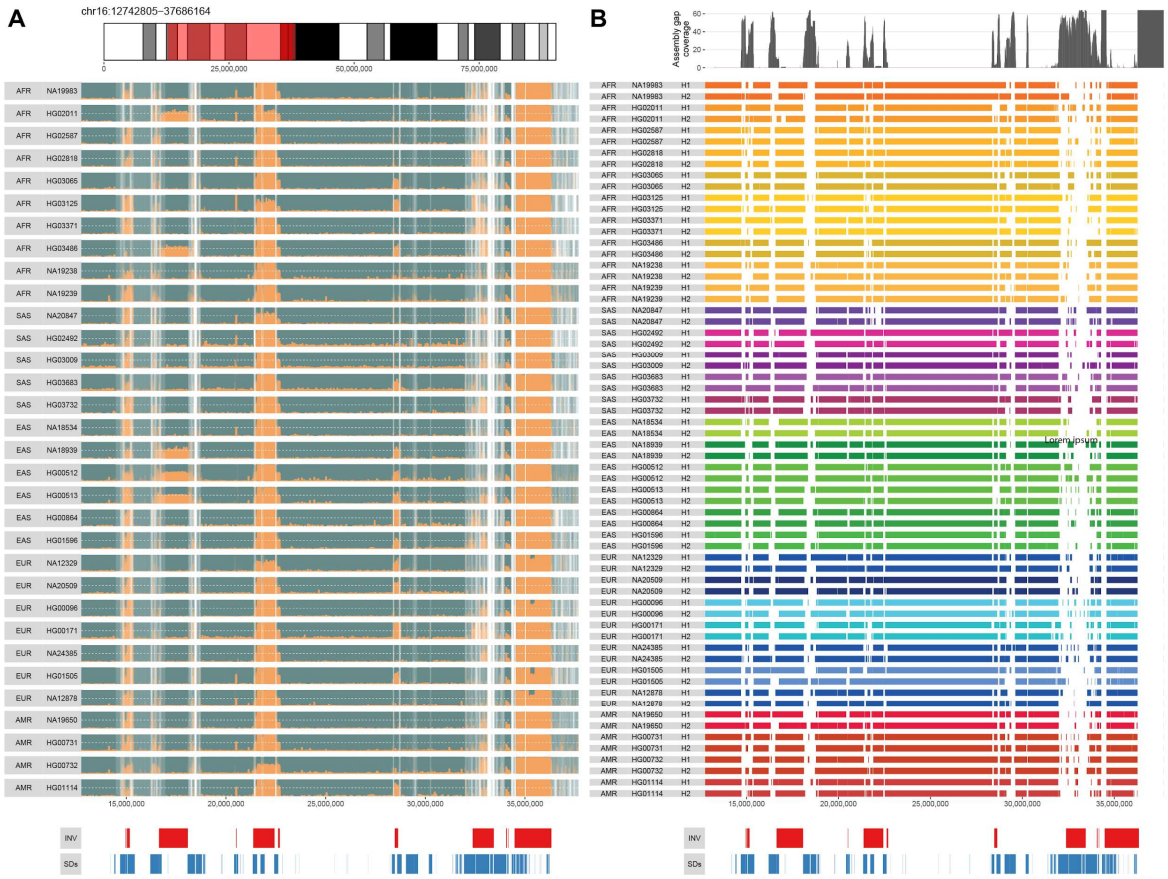


Fig. S25. Phased assembly alignments in 16p12 region

(A) Top panel: Shows a region on chromosome 16 presented in detail in other figure panels. Middle panel: Shows binned (bin size: 10 kbp, stepsize: 5 kbp) ratio of Crick (positive strand, '+', teal) and Watson (negative strand, '-', orange) reads for each sample-specific composite file in a given region; (B) Top panel: Gap coverage of phased assemblies aligned to GRCh38. Here, only regions where assembled contigs align with mapping quality ≤60 are summarized. Middle panel: Each row represents sample-specific haplotype-resolved assembly (H1 - haplotype1, H2 - haplotype2) aligned to GRCh38. Regions where assembled contigs do not align with mapping quality ≥60 are visible as white gaps between colored rectangles specific for each sample and haplotype; (A) & (B) Bottom panels: Red rectangles highlight inverted regions detected by Strand-seq while blue rectangles show the positions of SDs in GRCh38.

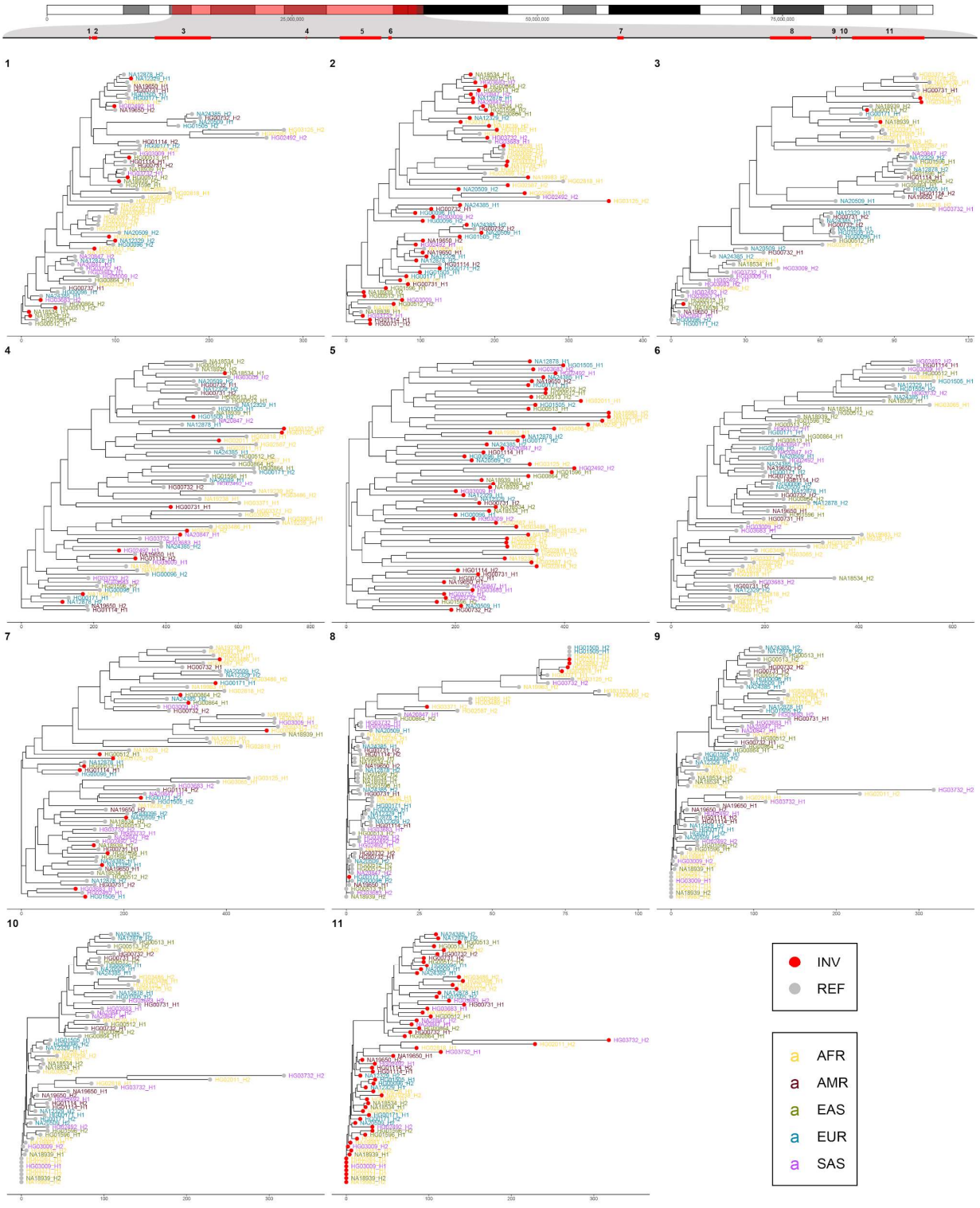

1 2 3 4 5 6 7 8 9 10 11
INV
REF
AFR
AMR
EAS
EUR
SAS

Fig. S26. Genetic background of inverted haplotypes at the 16p12 [...] (caption next page)

Fig. S26. Genetic background of inverted haplotypes at the 16p12 chromosome region

Top panel: A chromosome 16 ideogram with region of interest highlighted by red transparent rectangle. Below inverted regions are shown as red rectangles with numbers pointing to a neighbor-joining tree constructed based on phased SNVs from inversion flanking region (+/-500 kbp). SNVs from within the inverted region and those overlapping SDs (≥98% identity) were not considered for the tree construction. Haplotypes flanking inverted regions are highlighted by a red dot and those with a reference orientation (or unknown phase) are shown in gray. Each sample and haplotype label is colored based on the superpopulation of origin. Given the fact that this analysis does not take into account meiotic recombination, we suggest that inversions 1, 3, 4, 7 and 8 show recurrent traits as they seem to occur at different genetic backgrounds.

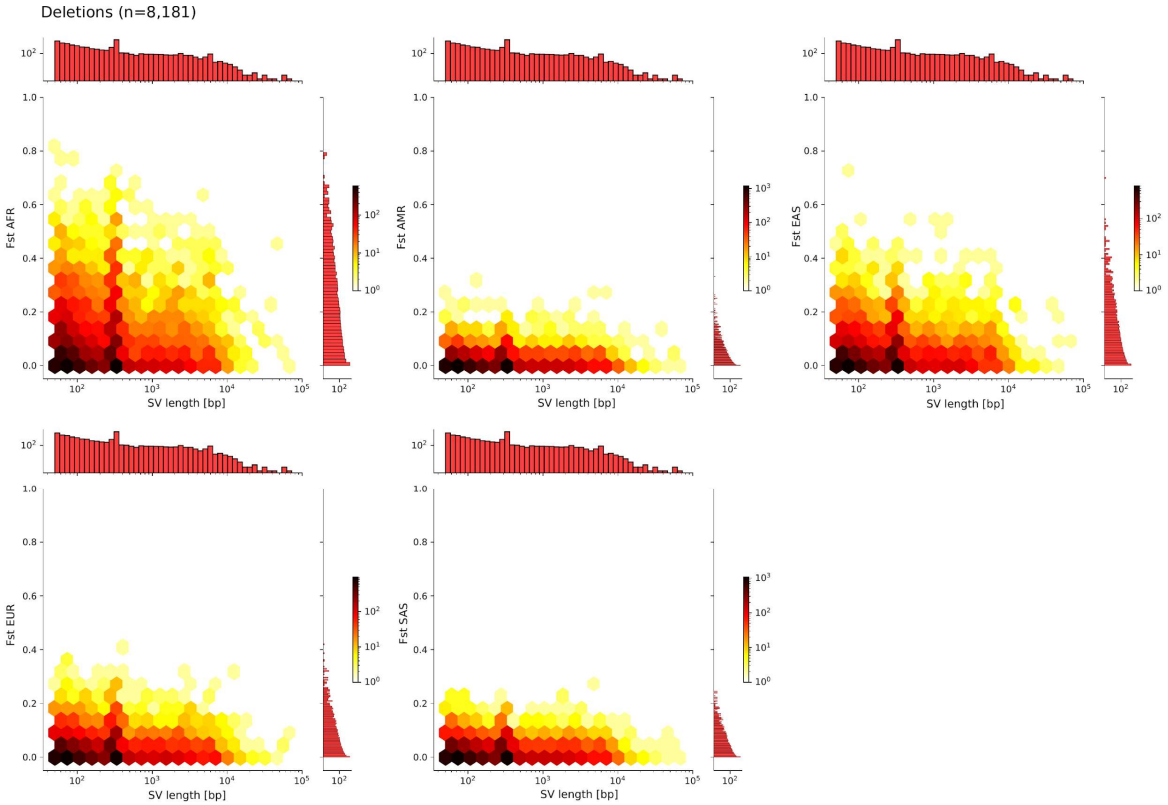


Fig. S27. Fst versus SV length for all superpopulations (deletions)

For each superpopulation, Fst values were computed by comparing to the union of the remaining populations. The plots are based on the filtered PanGenie calls containing 8,181 deletions.

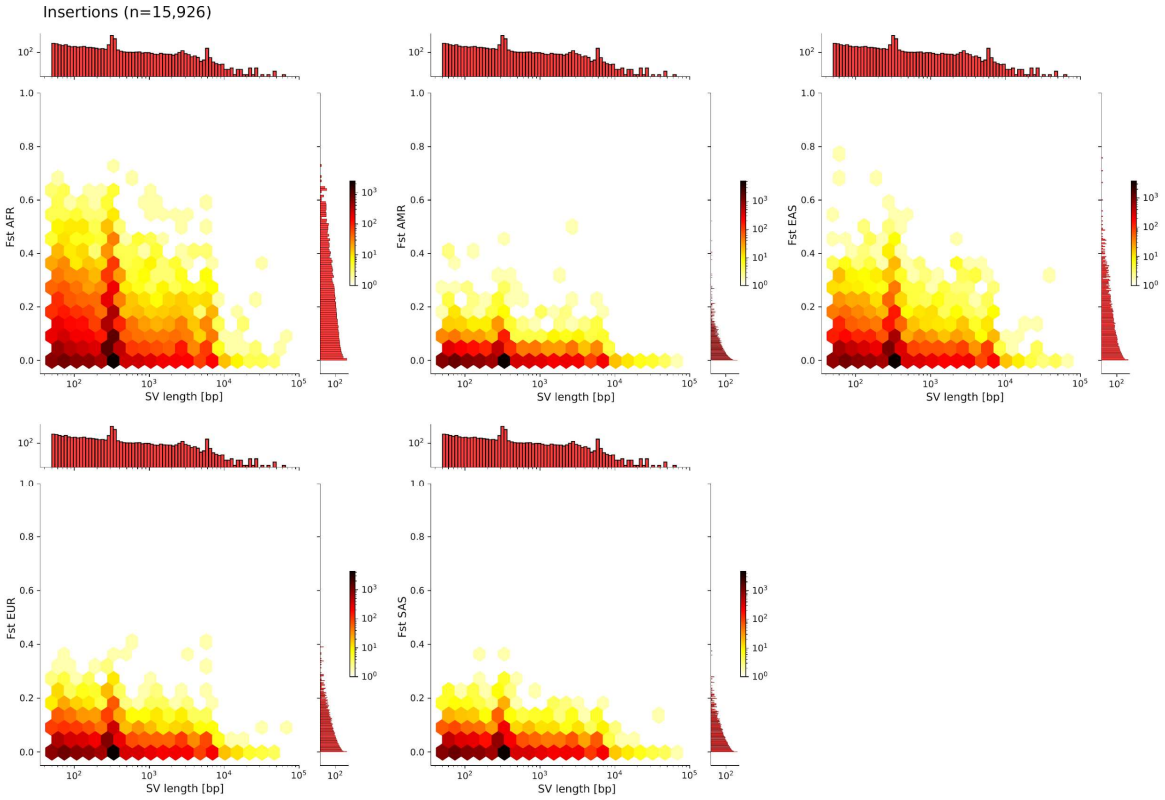


Fig. S28. Fst versus SV length for all superpopulations (insertions)

For each superpopulation, Fst values were computed by comparing to the union of the remaining populations. The plots are based on the filtered PanGenie calls containing 15,926 insertions.

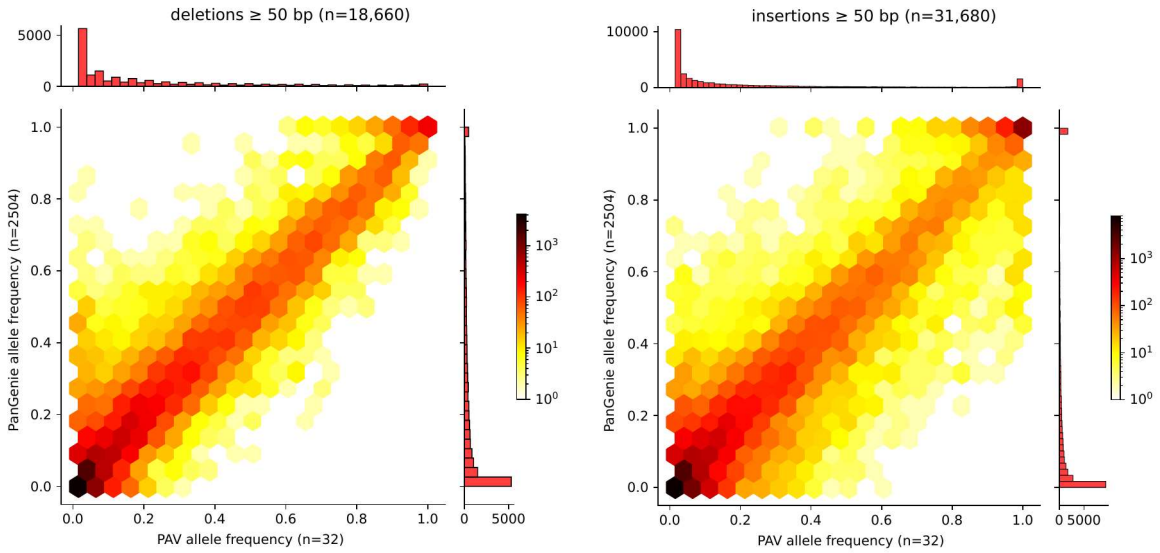


Fig. S29. Allele frequencies of SVs in the lenient callset

For PanGenie, allele frequencies were computed based on the genotypes of all 2,504 unrelated samples. The PAV allele frequencies were computed based on all 64 assemblies. Only SVs (≥50 bp) contained in our lenient callset (cutoff -0.5, n=50,340) were considered.

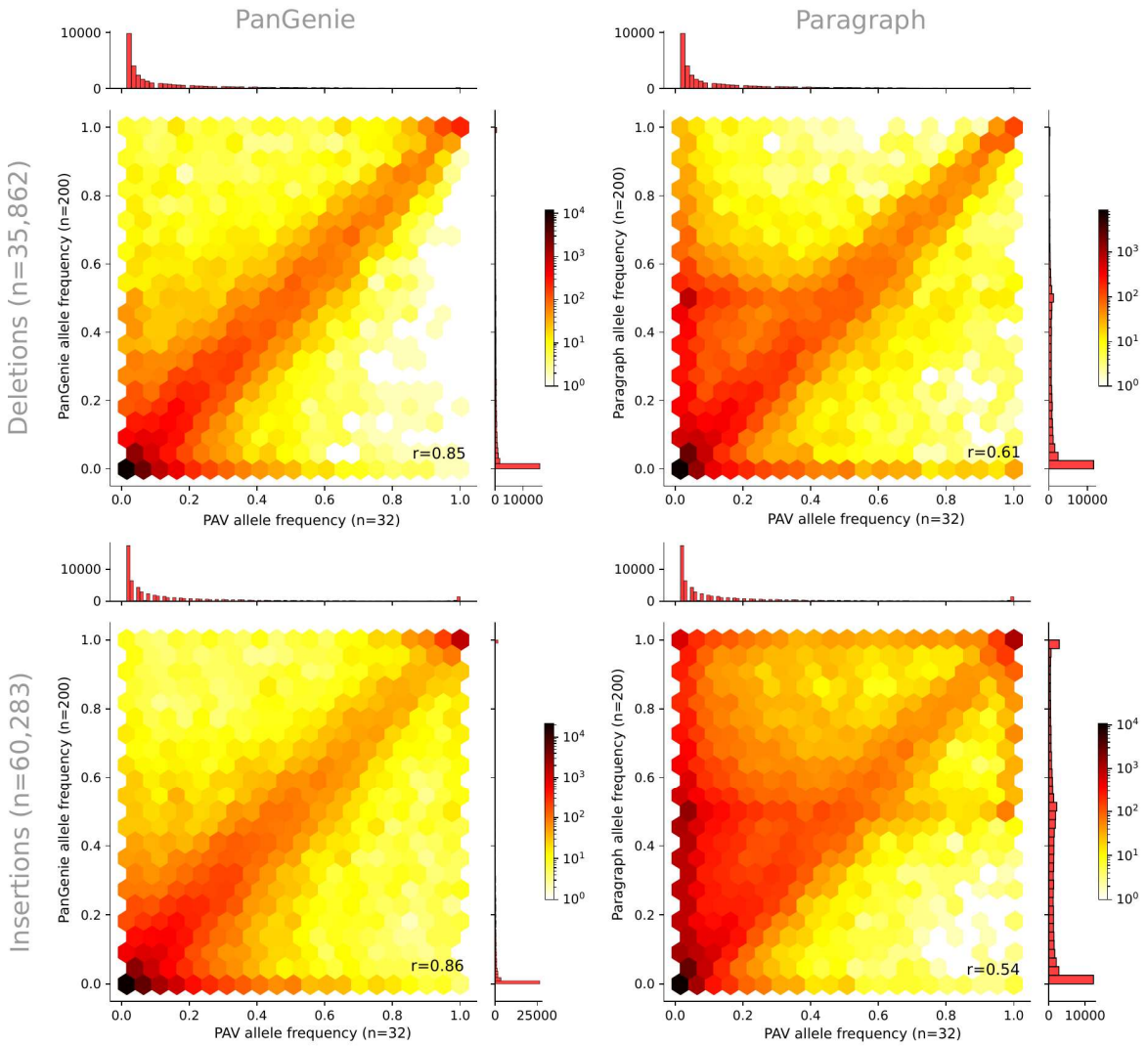


Fig. S30. Comparison of PanGenie and Paragraph allele frequencies on the pilot set

Paragraph and PanGenie were run on a subset of 100 trios (300 samples) in order to derive genotypes for all SVs (n=96,145). Allele frequencies were computed based on the genotypes of both methods for all 200 unrelated samples. PAV allele frequencies were computed based on all 64 assembly haplotypes.

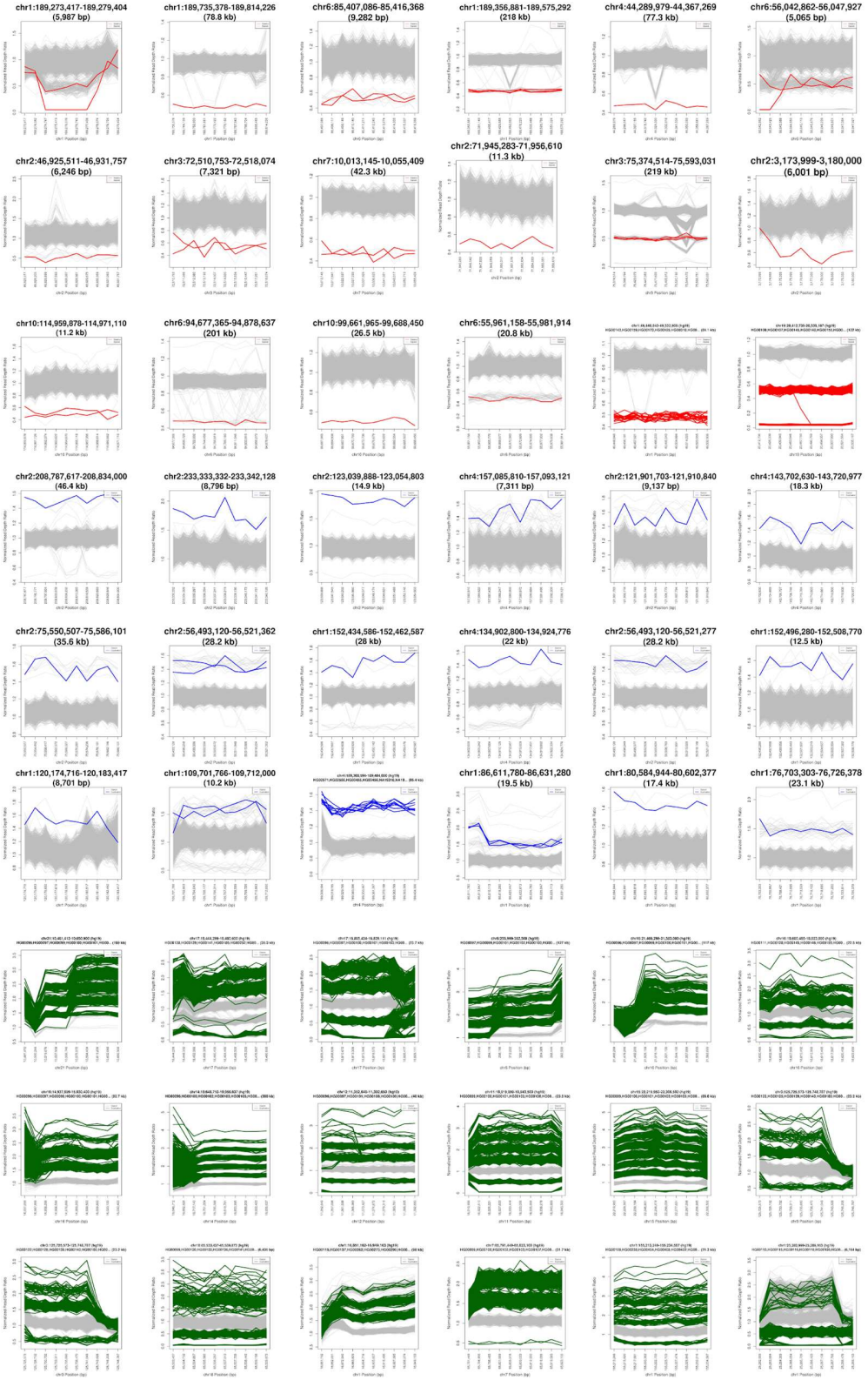


Fig. S31. Distribution of normalized sequencing depth of CNVs (caption next page)

Fig. S31. Distribution of normalized sequencing depth of CNVs

A randomly selected subset of CNVs over 5 kbp in size, discovered by short-read Illumina sequences but missed by long-read PacBio sequences are shown in this plot. Y-axis represents the normalized sequencing depth of each sample, with 1 representing copy number 2, 0.5 representing copy number 1, etc. Samples with deletions, duplications, and multi-allelic CNVs were shown in red, blue, and green colors, respectively.

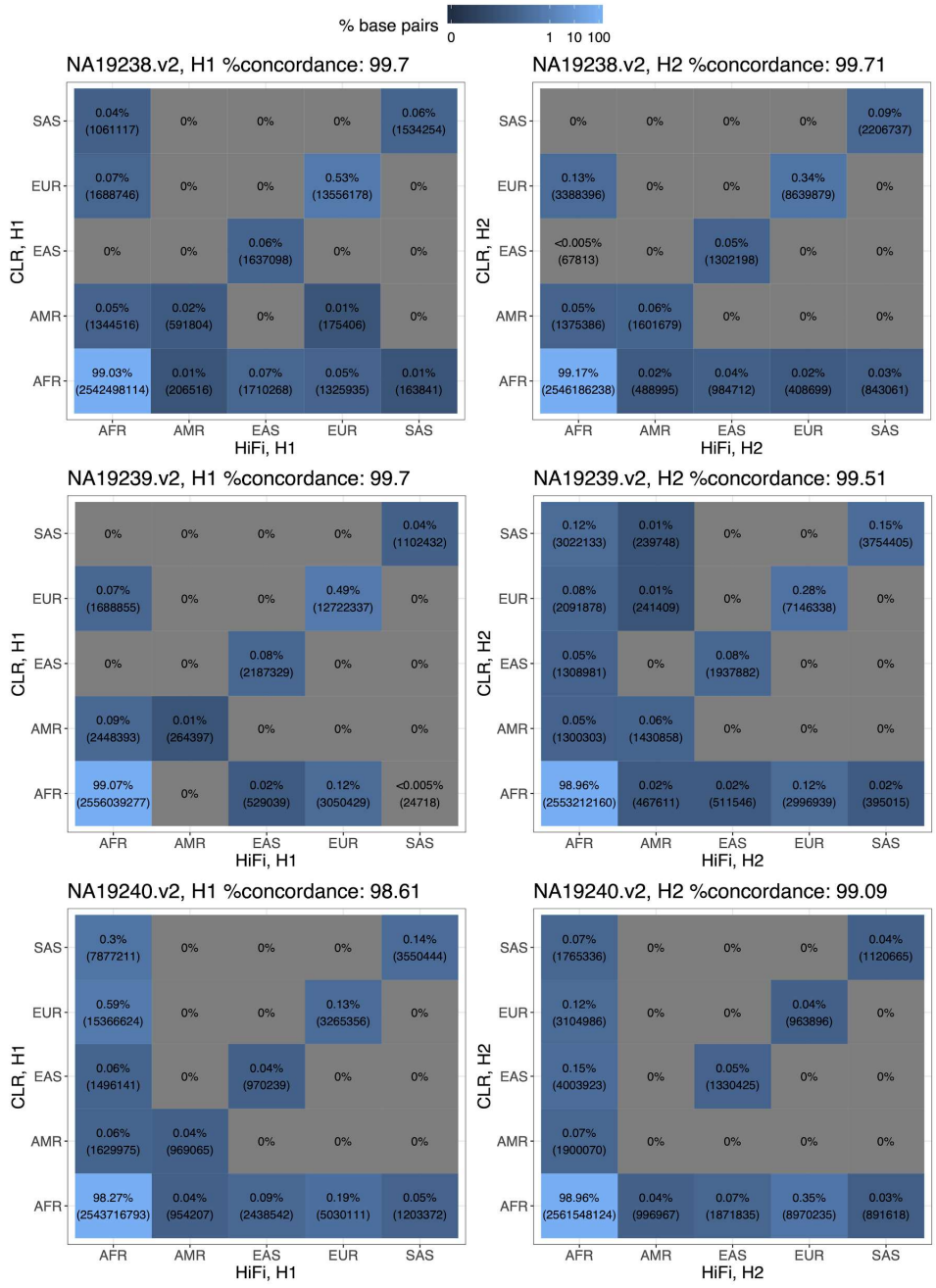


Fig. S32. Concordant ancestry calls between HiFi and CLR data for the YRI trio

In each row, the heatmaps indicate the fractions of concordant (diagonal entries) and discordant (off-diagonal entries) ancestry calls between the HiFi and CLR haplotypes from an individual genome. The numbers inside parentheses are the total number of base pairs in individual ancestry call combinations between the two datasets.

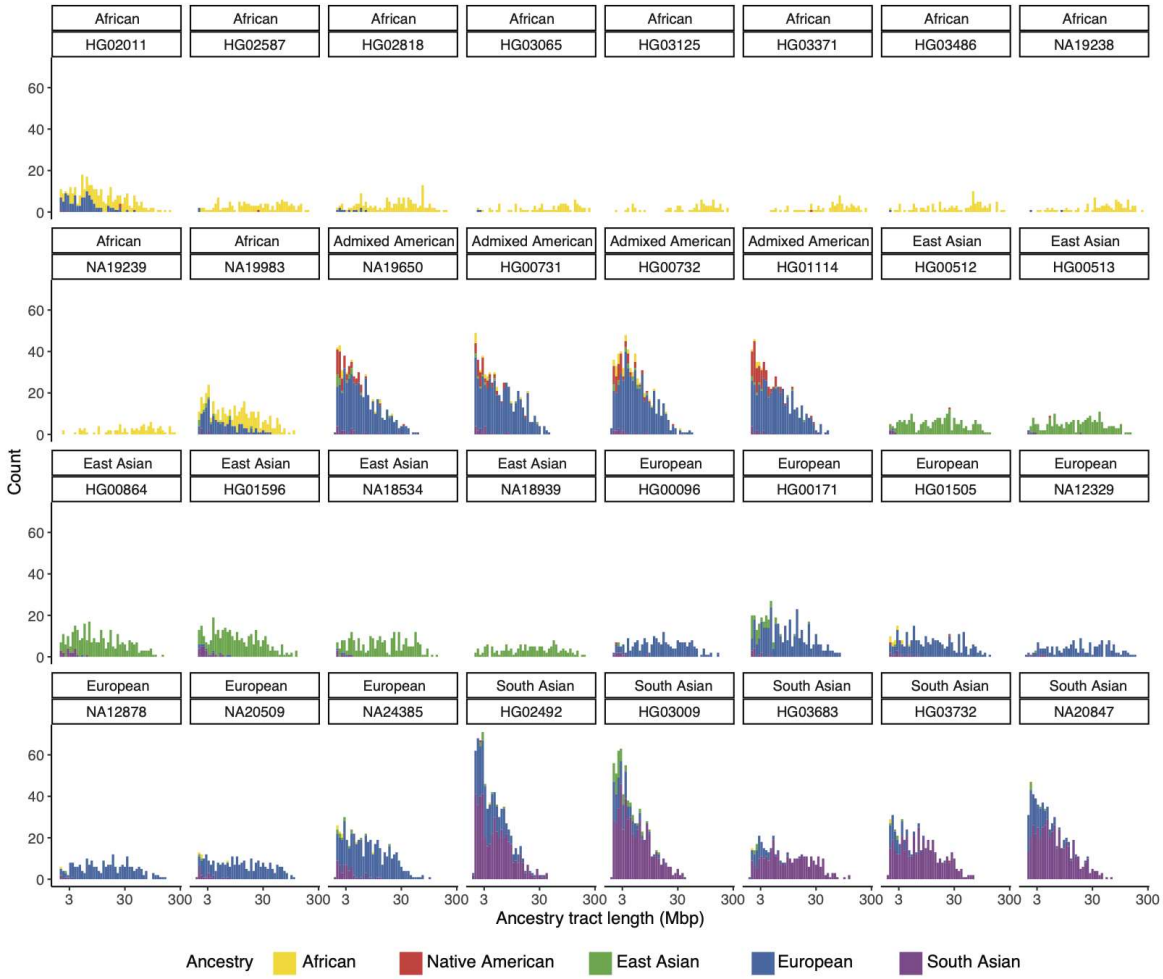


Fig. S33. Length distributions of inferred genome-wide ancestry tracts for all haplotype-phased assemblies

A per-individual version for Figure 6B, which shows the length distribution (log10 scale) of ancestry tracts for the 32 genomes.

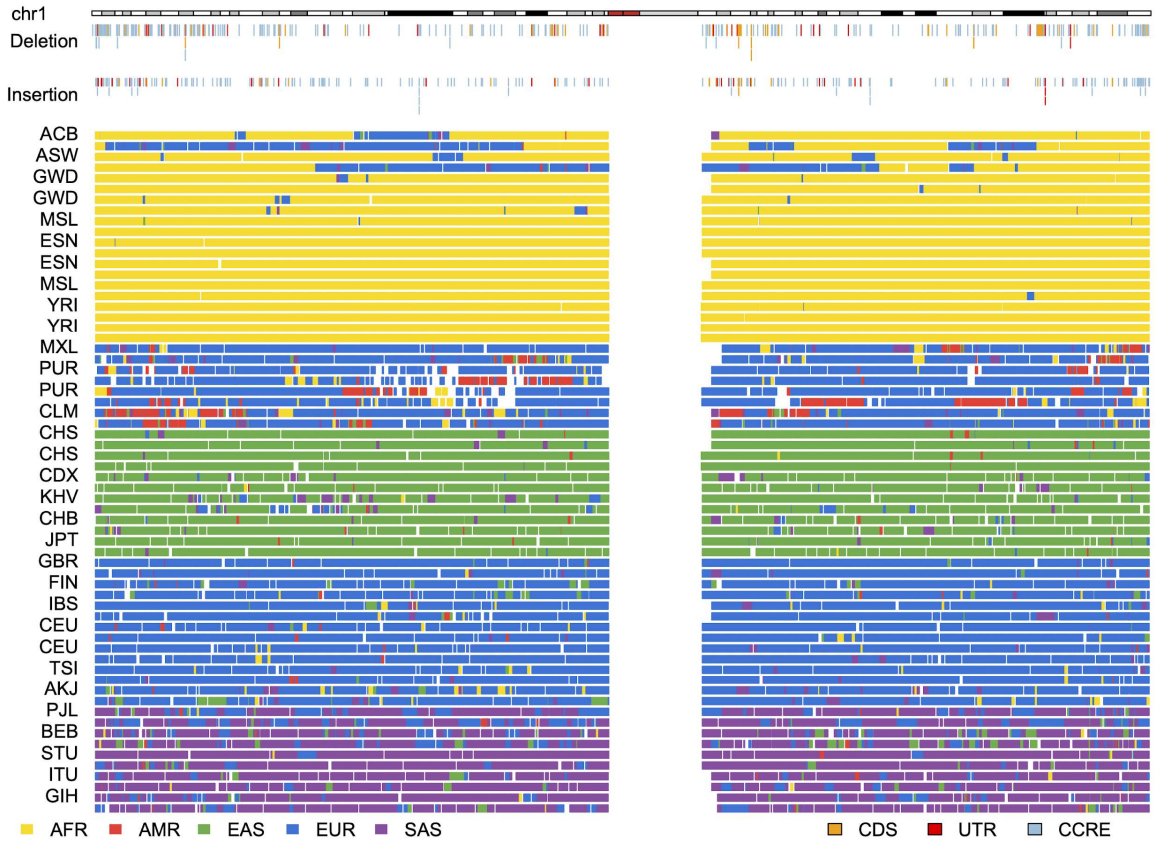


Fig. S34. Inferred local ancestry blocks across chromosome 1 for haplotype-phased assemblies

Ancestry calls are determined using the procedure described in the Methods. White spaces are discordant calls, known gaps, centromeres, and/or SDs. Deletion and insertion SVs that overlap with coding sequences (CDS), UTRs, and putative promoter sequences (CCRE) are annotated above.

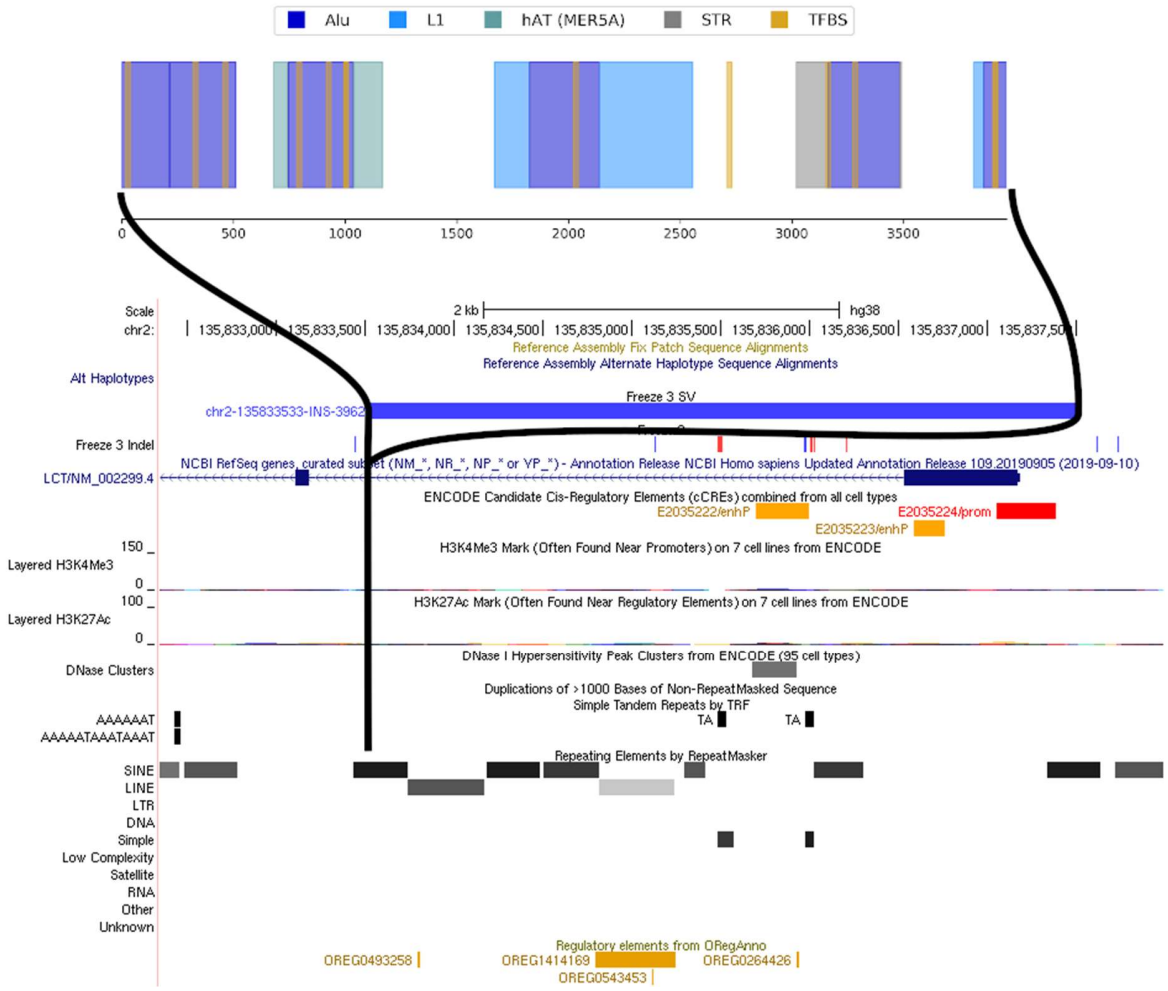


Fig. S35. Structure of a 4.0 kbp insertion in the first *LCT* exon

A 3,962 bp insertion in the first exon of *LCT* contains ancestral sequence with strong evidence for an AluY-mediated deletion captured as the reference allele. The insertion site is in a reference AluY, and 312 bp of AluY sequence (dark blue) is split between both ends of the inserted sequence. MEME predicts 11 transcription factor binding sites (yellow).

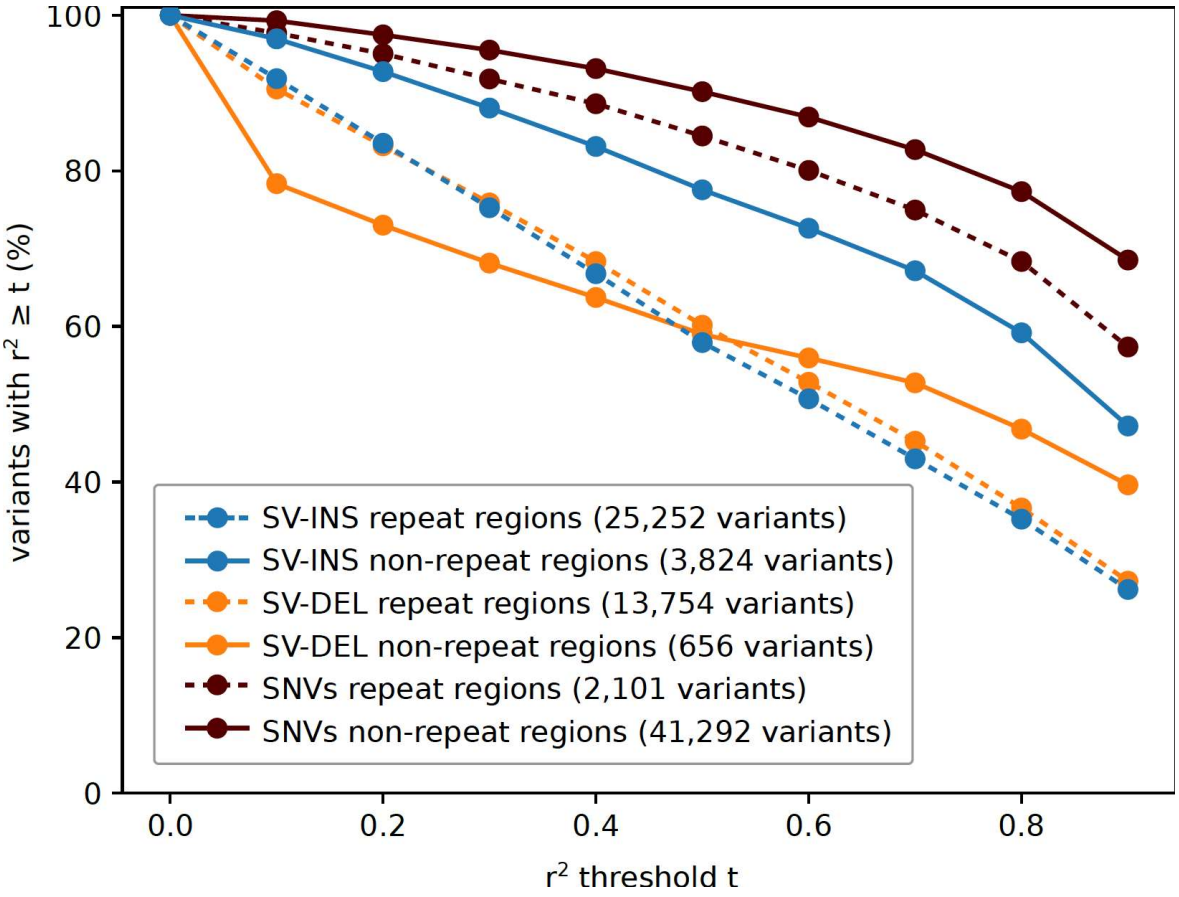


Fig. S36. PanGenie SV-SNV LD

For each SV not discovered by Illumina but genotyped with an allele frequency >0 by PanGenie, we performed a linkage disequilibrium (LD) analysis computing r^2 for all SNVs within a 1 Mbp window. For different cutoffs t on r^2, the plot shows the number of SVs for which r^2 was greater or equal to t for at least one SNV (blue/orange curves for insertion/deletion SVs, respectively). We distinguish repeat regions (dashed lines, simple repeats or SDs) and non-repetitive regions (solid line). In addition, we randomly selected a comparable number of SNVs from our callset and repeated the same LD analysis (brown curves).

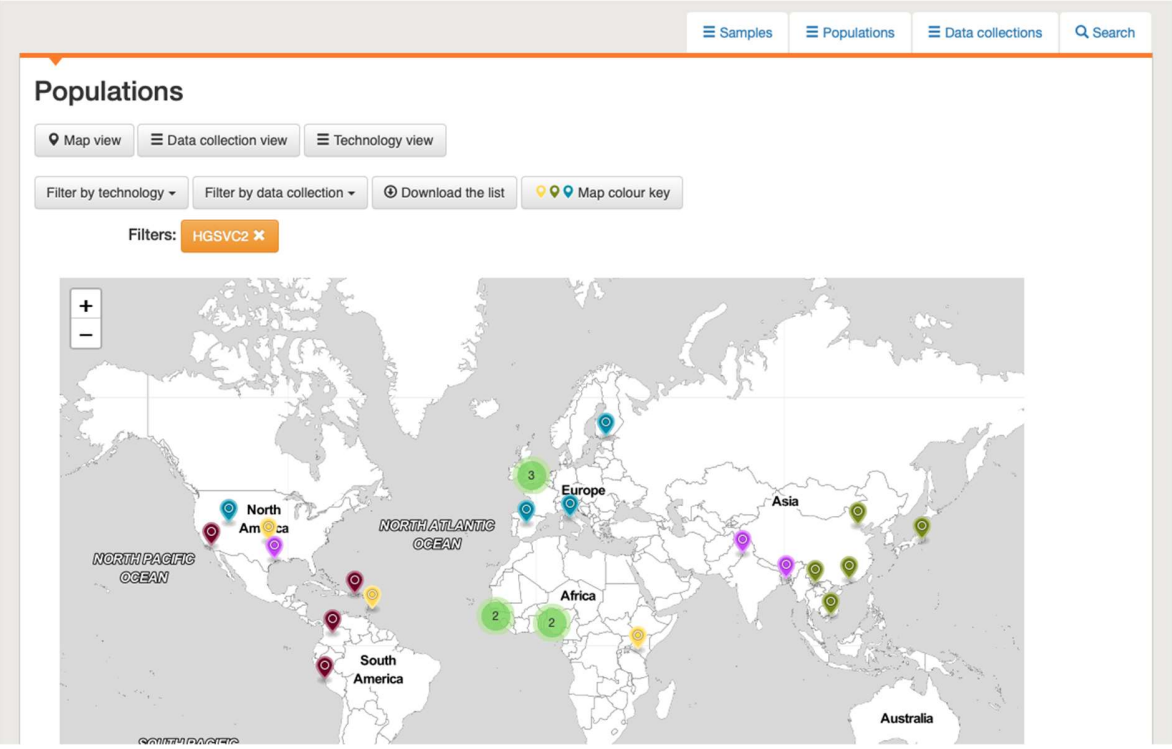


Fig. S37. Screenshot of HGSVC2 population distributions in the IGSR portal map view

Map view of the populations analyzed as part of this study (screenshot taken on 24th January 2021). This view can be accessed via the tab "Populations", button "Map view", and setting the data collection filter to "Human Genome Structural Variation Consortium 2" in the IGSR data portal, which includes the HGSVC2 official portal and integrates the HGSVC2 resources with other datasets generated on openly consented human samples.

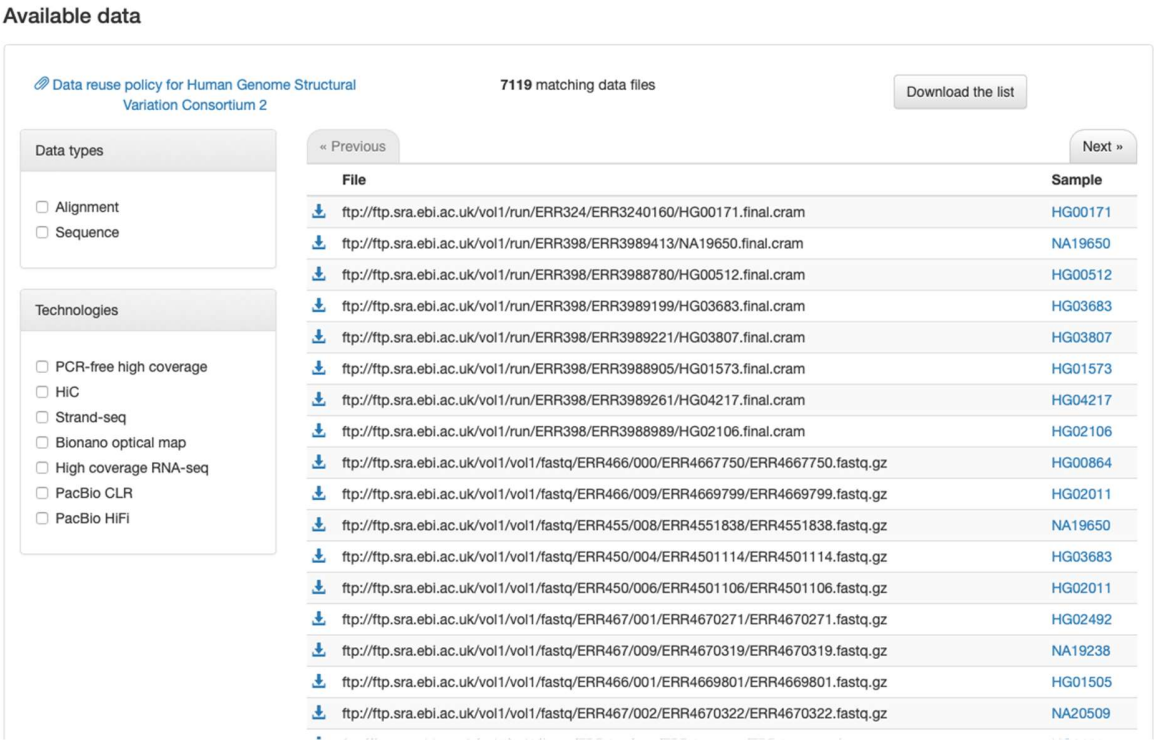


Fig. S38. Screenshot of HGSVC data resource in the IGSR portal

Landing page of the official HGSVC2 data portal listing available datasets (bottom of the page; screenshot taken on 24th January 2021). This list can be filtered by technology and data format (left menu), and includes a direct link to the data reuse policy for the HGSVC2 data (top left).

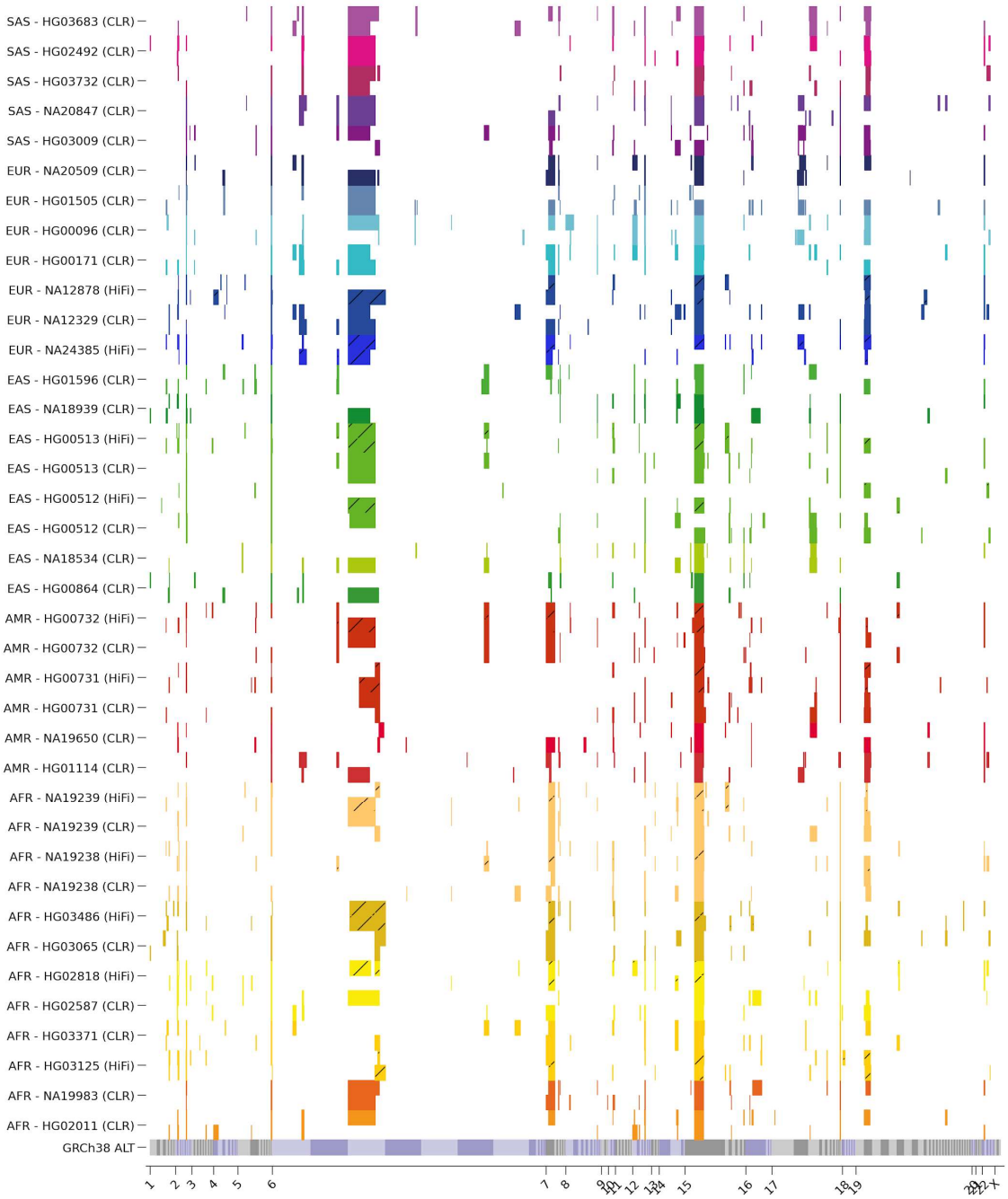


Fig. S39. Haploid assembly contig coverage on GRCh38 ALT contigs

GRCh38 ALT contigs were concatenated and plotted with a color scheme alternating between chromosomes (1-22, X) and between ALT contigs on the same chromosome (x-axis). Haploid assembly contig coverage is indicated as colored blocks (see Fig. 1 for color legend; HiFi assemblies plotted with hatched bars) for contig alignments at a MAPQ of 60 (y-axis).

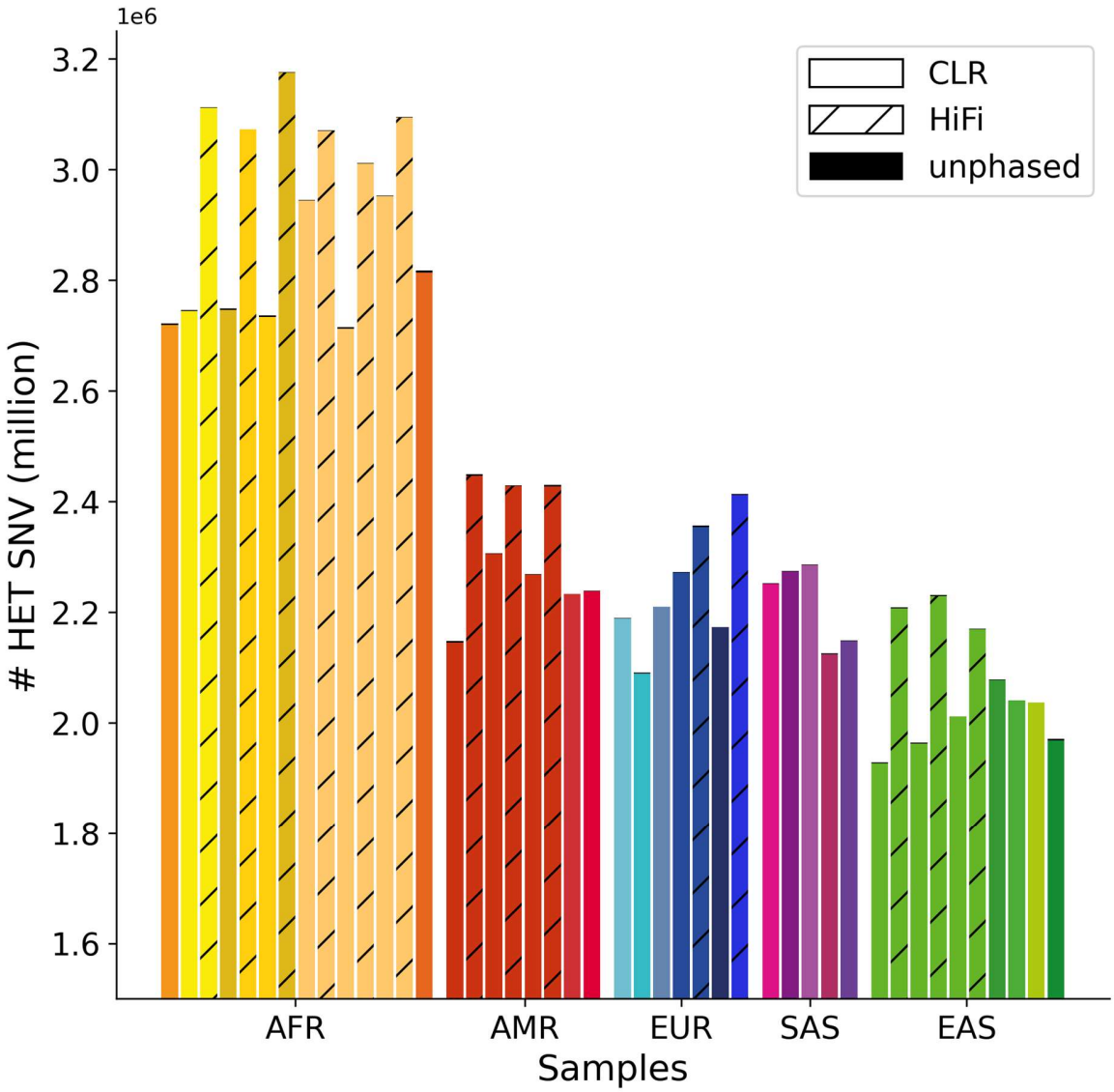


Fig. S40. Heterozygous SNVs per sample

Samples are grouped by superpopulation along the x-axis (for color code, see Fig. 1). Bar height (y-axis, in millions) depicts the number of heterozygous SNVs per sample, as called in the diploid assembly pipeline to obtain local phase information. The unphased fraction of variants is indicated as the top black part of each bar. HiFi samples are represented by hatched bars.

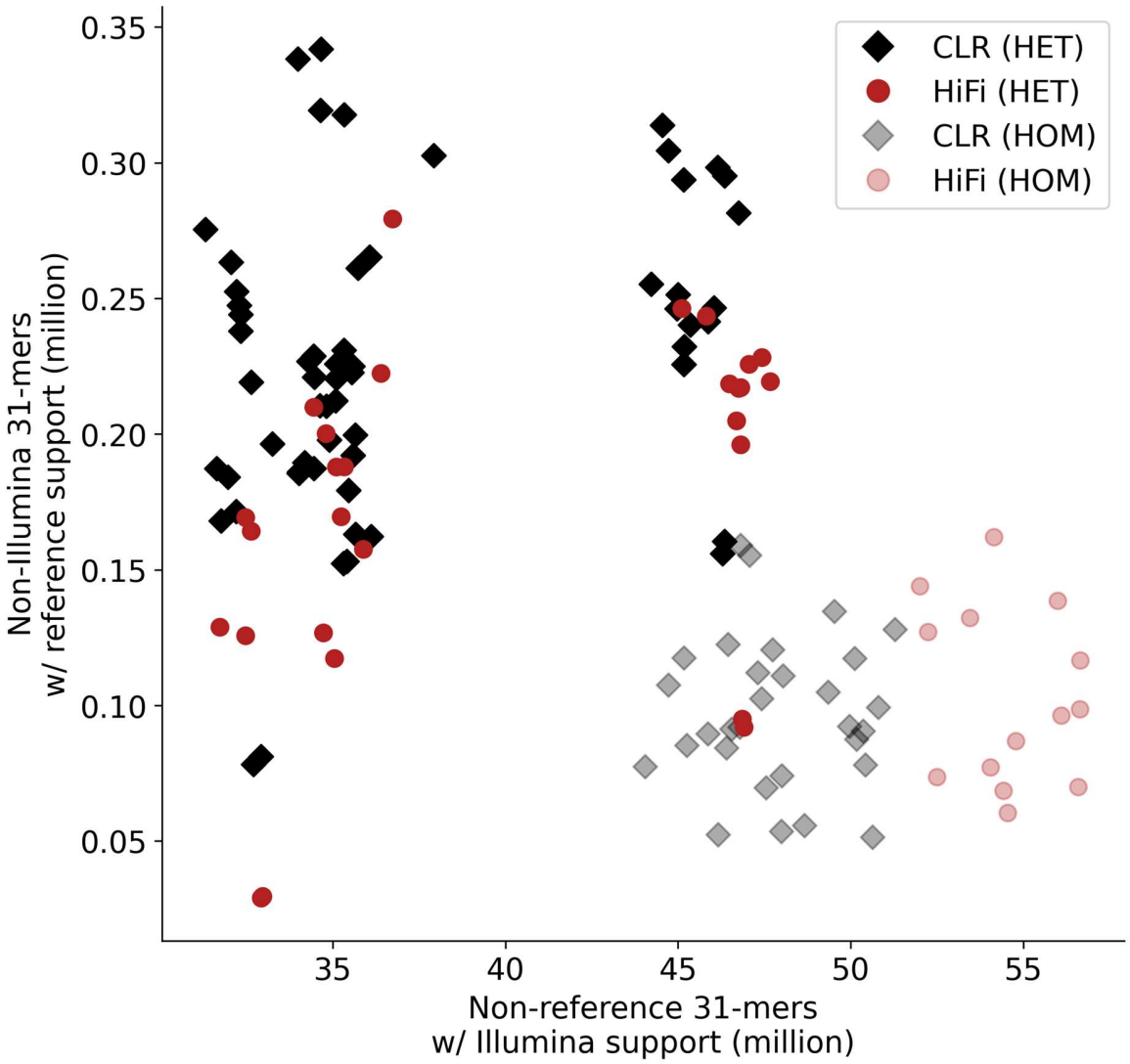


Fig. S41. Haploid assembly k-mer counts

Individual haplotype assemblies are depicted as black diamonds (CLR) or red circles (HiFi). 31-mer counts supported by Illumina data but not found in the GRCh38 reference are plotted on the x-axis (in million), and 31-mer counts supported by the GRCh38 reference and not found in Illumina data on the y-axis. The fraction of homozygous k-mers found in both assembled haplotypes is depicted as semi-transparent points.

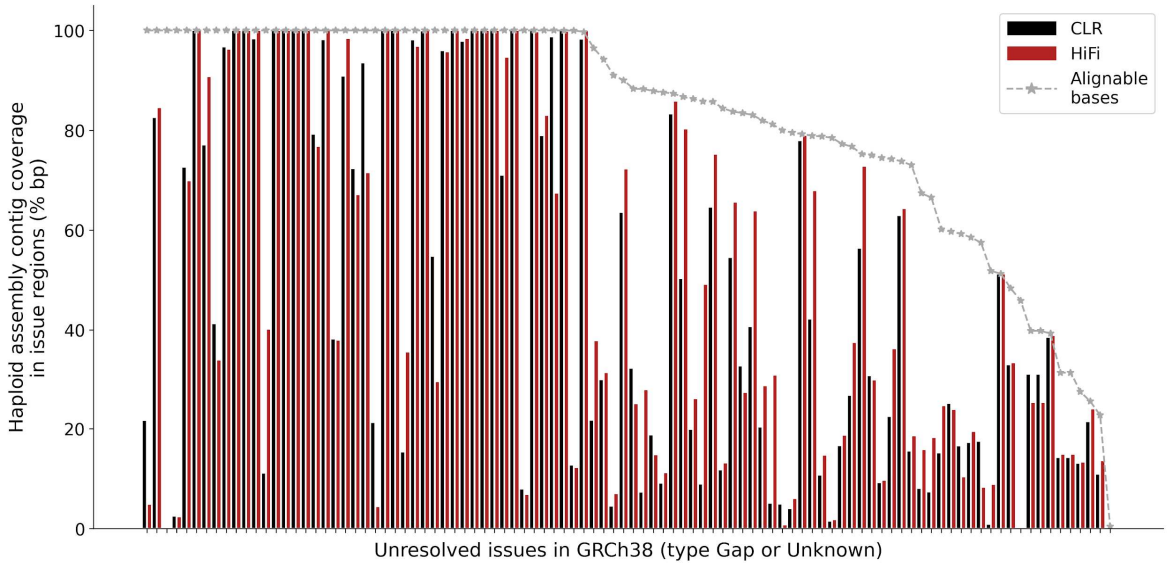


Fig. S42. Haploid assembly contig coverage in GRCh38 issue regions

Unresolved issues in GRCh38 of type “Gap” or “Unknown” are arranged in descending order by percent alignable bases (ACGT, but not N) on the x-axis. The maximal attainable contig coverage per issue region is depicted with the gray star line at the top. Bar height indicates average haploid contig coverage (y-axis, % covered bp in region) for all CLR (black) and HiFi (red) assemblies.

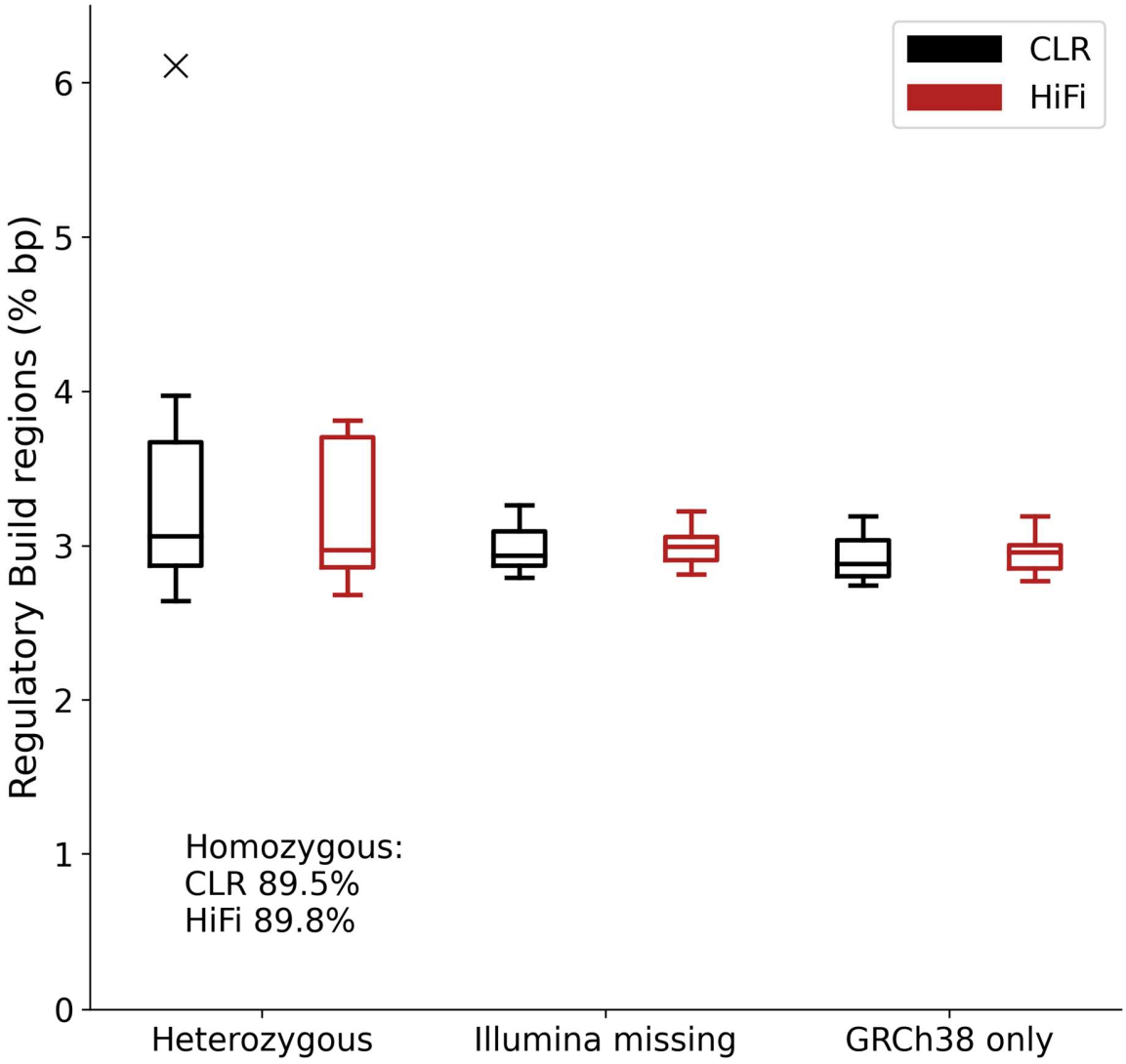


Fig. S43. Ensembl Regulatory Build regions in haploid assemblies

Regulatory regions are classified for each phased assembly as homozygous (CLR/HiFi average, text only for layout reasons), heterozygous (left), missing in Illumina short reads but present in at least one haplotype (middle), or only detected in the GRCh38 reference (right). Box plots illustrate the amount of detectable regulatory regions as percent-annotated base pairs over all Regulatory Build categories and separately for CLR (black) and HiFi (red) assemblies.

Fig. S44. Phased assembly contig coverage

GRCh38 regions covered with phased assembly contig alignments were defined based on Bionano hybrid scaffolded assemblies (dark green), or based on MAPQ60 threshold alignments (green). SDs with >95% identity are shown in shades of blue (steps >98% and >99% identity), and N gaps in the GRCh38 reference are indicated in gray.

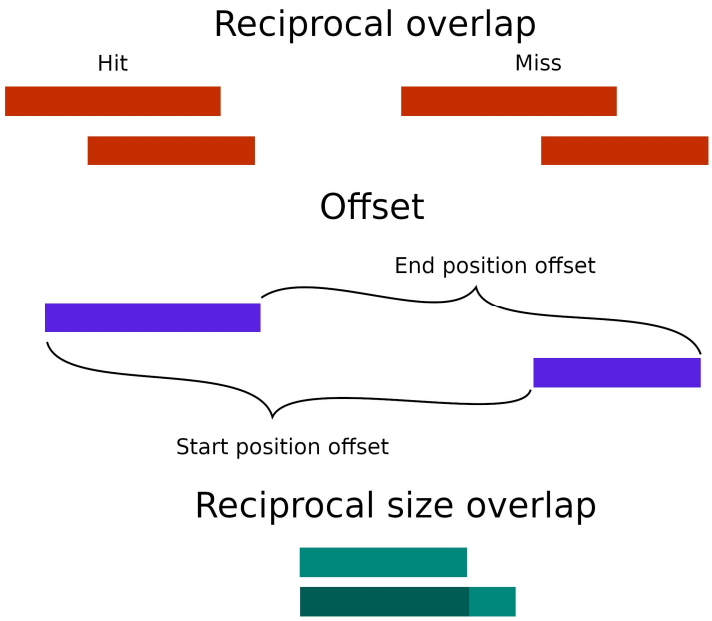


Fig. S45. Variant merging strategy

50% reciprocal overlap (red) often misses similar-sized events if they are shifted by a small distance, especially for smaller events. By considering a distance-size overlap, smaller variants within a given proximity (blue) and size overlap (green) can be intersected.

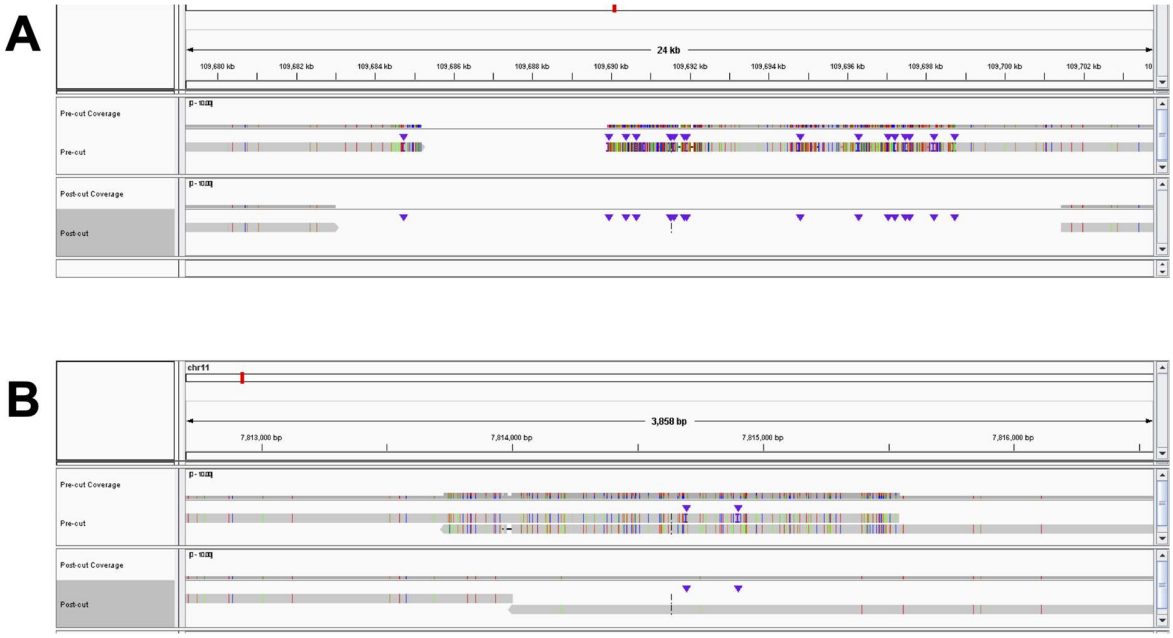


Fig. S46. PAV alignment trimming

(A) A contig alignment was fragmented into two records around an SV deletion (18.4 kbp). This deletion was flanked by an SD repeat, and the single repeat copy in the contig was aligned to each reference copy (top track). Alignment trimming to remove multiply-mapped contig bases trimmed back the contig map removing as many non-sequence-match CIGAR operations (op codes I, D, and X) as possible (bottom track). (B) A contig alignment was fragmented around an SV insertion (10.1 kbp tandem duplication). The contig has an extra copy of the duplication, and both were aligned to the same reference copy (top track). PAV trimming removed multiply-mapped reference bases leaving unaligned bases (the insertion) placed at the most likely breakpoint (bottom track).

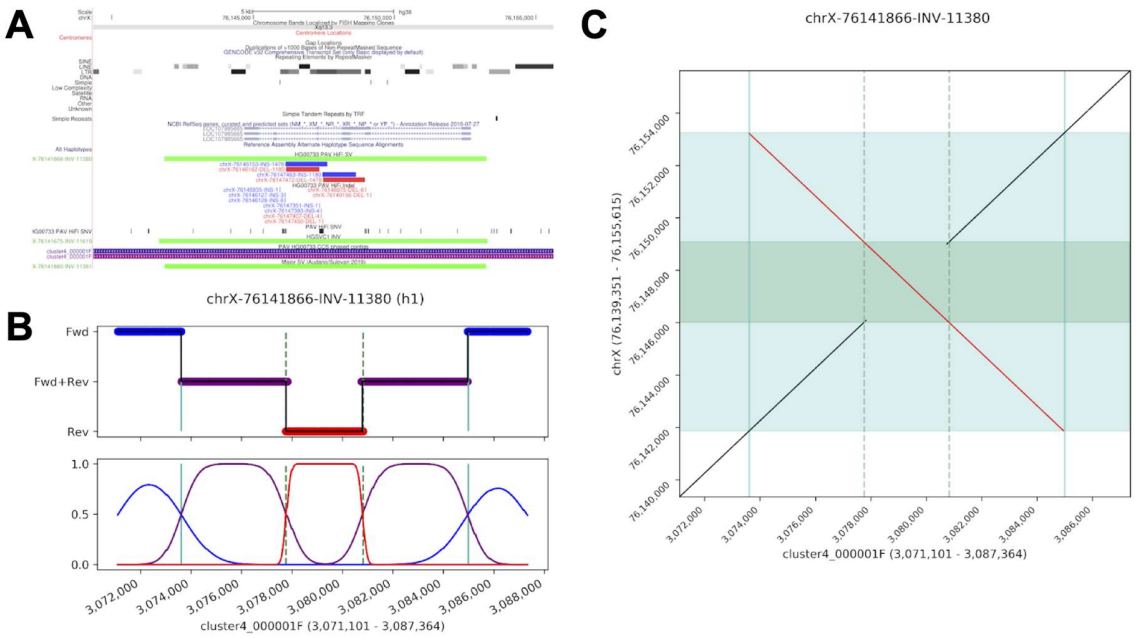


Fig. S47. PAV flags inversion sites

(A) PAV flagged a potential inversion site by identifying matched SV insertions and deletions of a similar size. In this region, the contig alignments (pink and purple bars) were not broken by the inversion. Clusters of indels and SNVs can also be seen near the center of the inversion. These events were used to seed an inversion search. (B) The k-mer density plot showing reference-oriented k-mers (blue), k-mers in reference and reverse orientation (purple), and k-mers strictly in reverse orientation (red). Top panel is the k-mers, and bottom panel is the scaled density plot for each orientation class. This inversion is flanked by a large repeat, which can be seen as large stretches of k-mers in both orientations, and the boundaries become the outer breakpoint (fwd to fwd-rev) and inner breakpoint (fwd-rev to rev) where the true inversion breakpoint is somewhere between these. (C) The dotplot of the resolved inversion. The inner strictly inverted region between inner breakpoints is shown in dark green (reference) and solid lines (contig), and the inverted duplications are shown as light green regions (reference) and the area between the solid and dashed lines (contig). The inversion call itself is reported using the outer breakpoints, and the inner breakpoints are carried as an annotation.

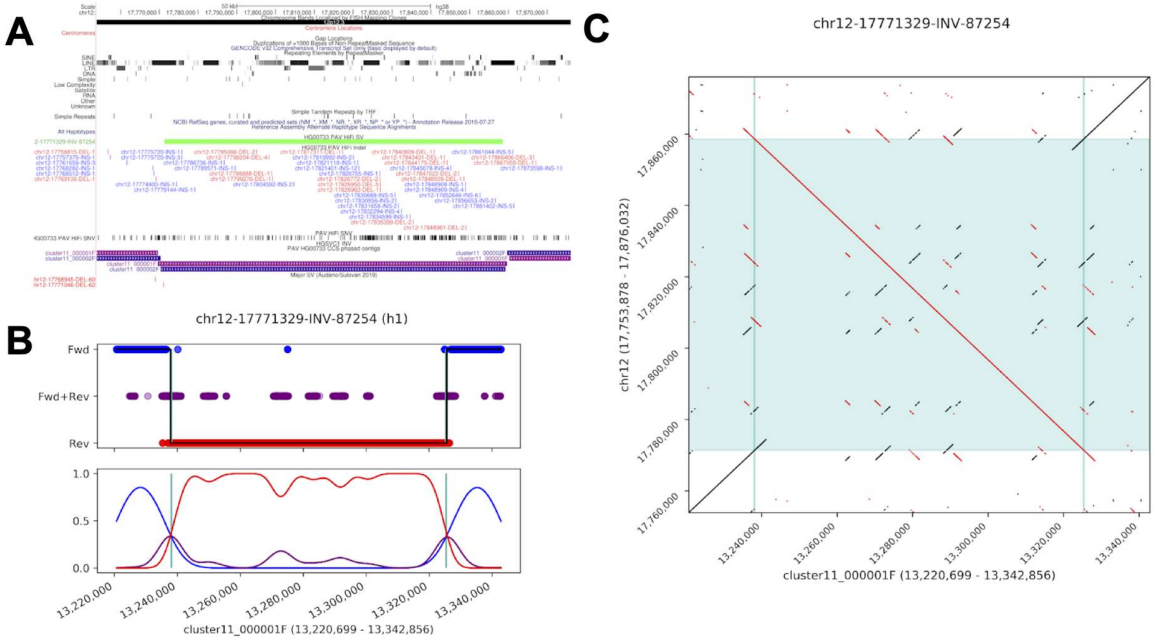


Fig. S48. PAV flags inversion sites

(A) Contig alignments (h1 pink, h2 purple) truncated at inversion breakpoints exhibiting prototypic alignment structure over an inversion with orientation reversed inside the inversion. (B) The k-mer density plot showing reference-oriented k-mers (blue), k-mers in reference and reverse orientation (purple), and k-mers strictly in reverse orientation (red). Top panel is the k-mers, and bottom panel is the scaled density plot for each orientation class. This inversion has no clear inner breakpoint. (C) The dotplot with the contig (horizontal) and reference (vertical) with inversion flanks in reference orientation.

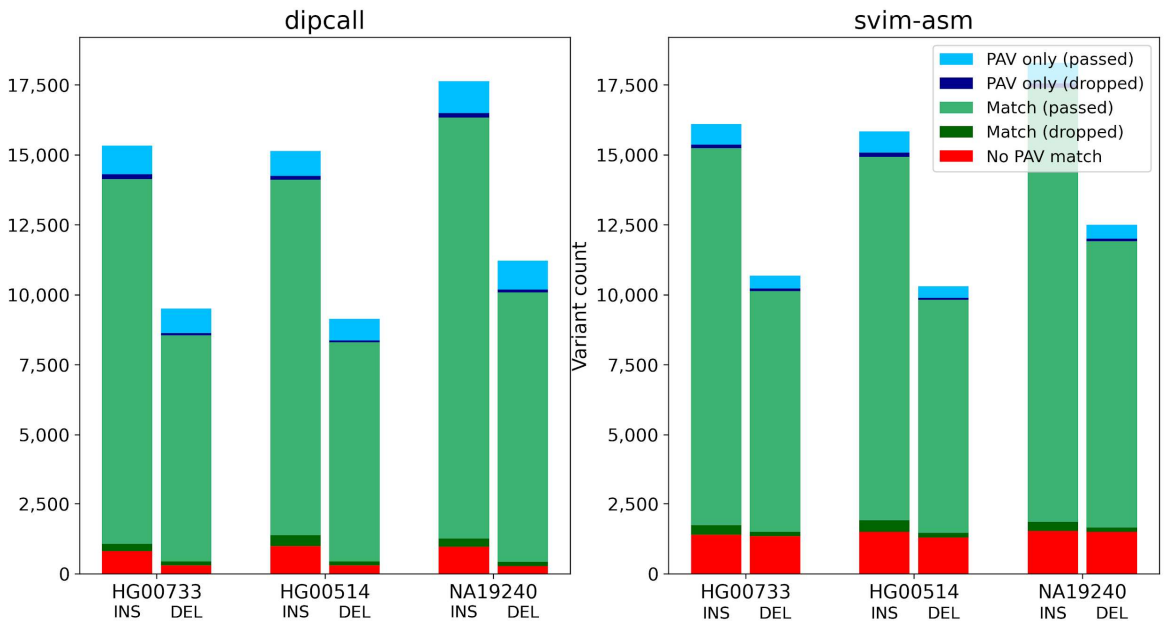


Fig. S49. PAV versus dipcall and svim-asm

PAV calls were matched with dipcall (left) and svim-asm (right) by searching for support within 1 kbp. dipcall has fewer calls that are not supported by PAV (red). A majority of SVs from both callers agree with PAV (light green) with some matching PAV calls dropped in QC (dark green), which may be over-filtering or systematic error among the methods. A majority of PAV-only calls passed QC (light blue) with a small number failing QC (dark blue). Compared to dipcall, there are fewer PAV-only calls generated by svim-asm.

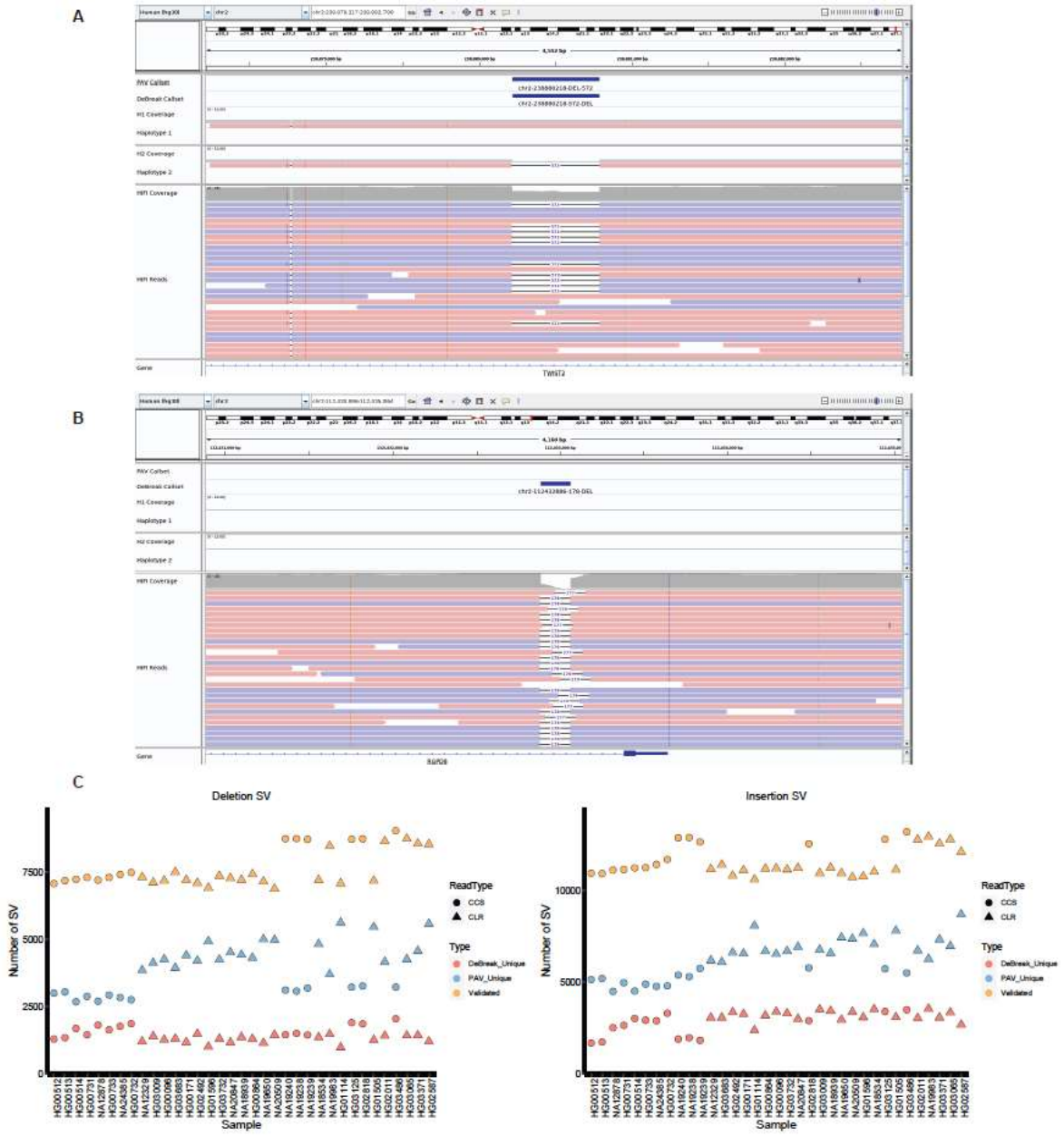


Fig. S50. DeBreak variant discovery

(A) An example of a deletion detected by both DeBreak and PAV in HG00733 HiFi sample; (B) An example of a deletion detected by DeBreak only. No assembly contigs were aligned to this region; (C) Number of SVs detected by DeBreak only, PAV only, and by both callers for deletion (left) and insertion (right) SVs in all samples.

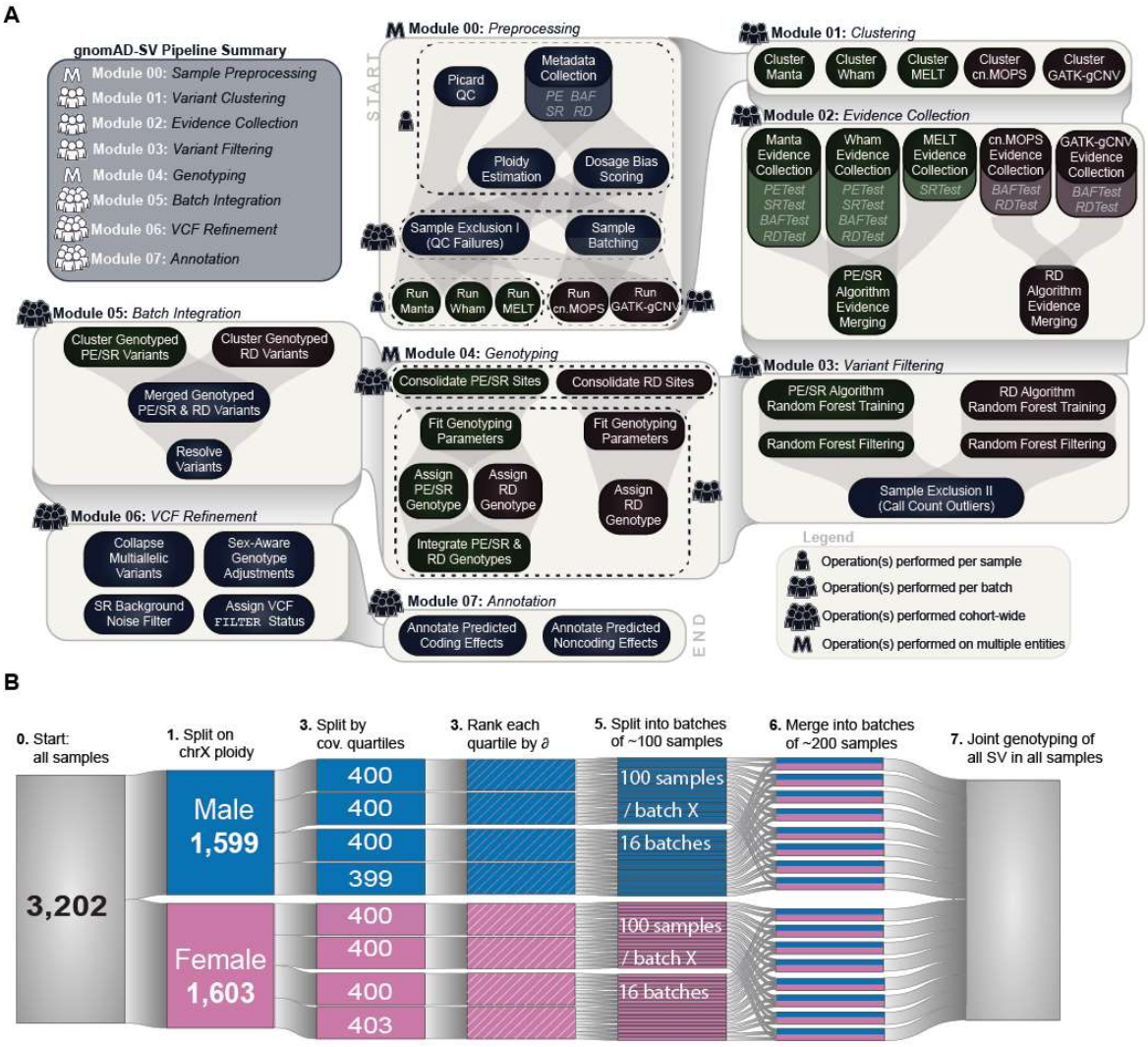


Fig. S51. Overview of GATK-SV pipeline

(A) An overview of the GATK-SV pipeline is summarized here, and details of the method were described in (*5*) as outlined in detail in Methods. The GATK-SV discovery pipeline contains seven sequential modules (light beige boxes). The sequence of modules is listed in the top left panel and is also indicated by connections between light beige boxes. Each module contains multiple sub-modules (smaller, dark boxes) that operate on the per-sample (N=1), per-batch (N~200), or cohort-wide (N=3,202) level. This pipeline has been made available as a series of publicly accessible methods on FireCloud/Terra to permit cloud-based analyses of SVs across WGS studies; (B) Batching strategy applied in GATK-SV. The 3,202 samples were grouped into 200 sample batches for efficient processing of the GATK-SV pipeline. Details of the strategy are described in the supplementary methods.

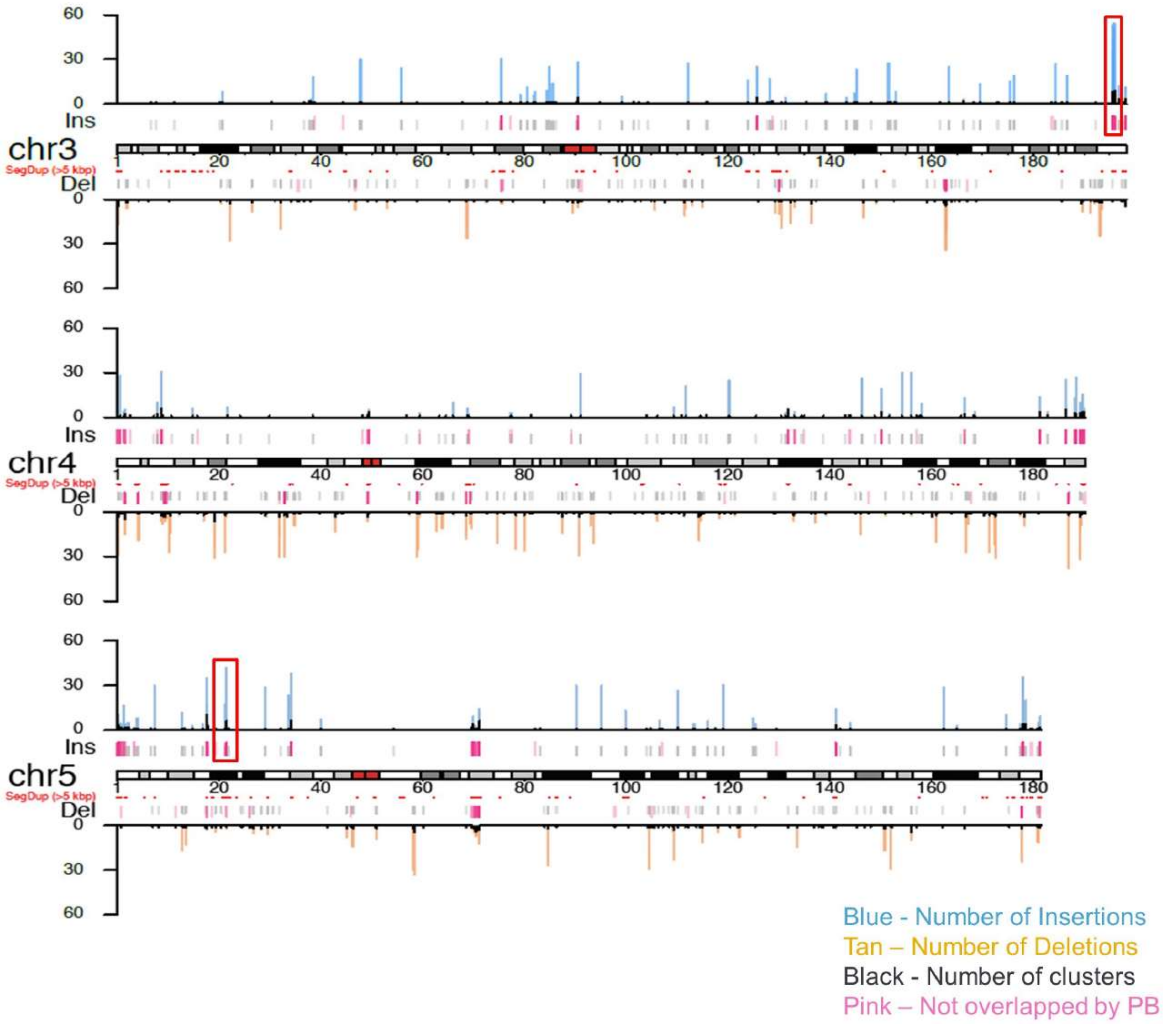


Fig. S52. Ideogram showing Bionano calls (≥5 kbp) and clusters

Large complex SV sites are regions with a high number of SV calls and with at least five clusters. Boxed in red on Chr3 and Chr5 are examples of these complex polymorphic regions . Please refer to the separate PDF version of this figure for the full-size image showing the complete ideogram.

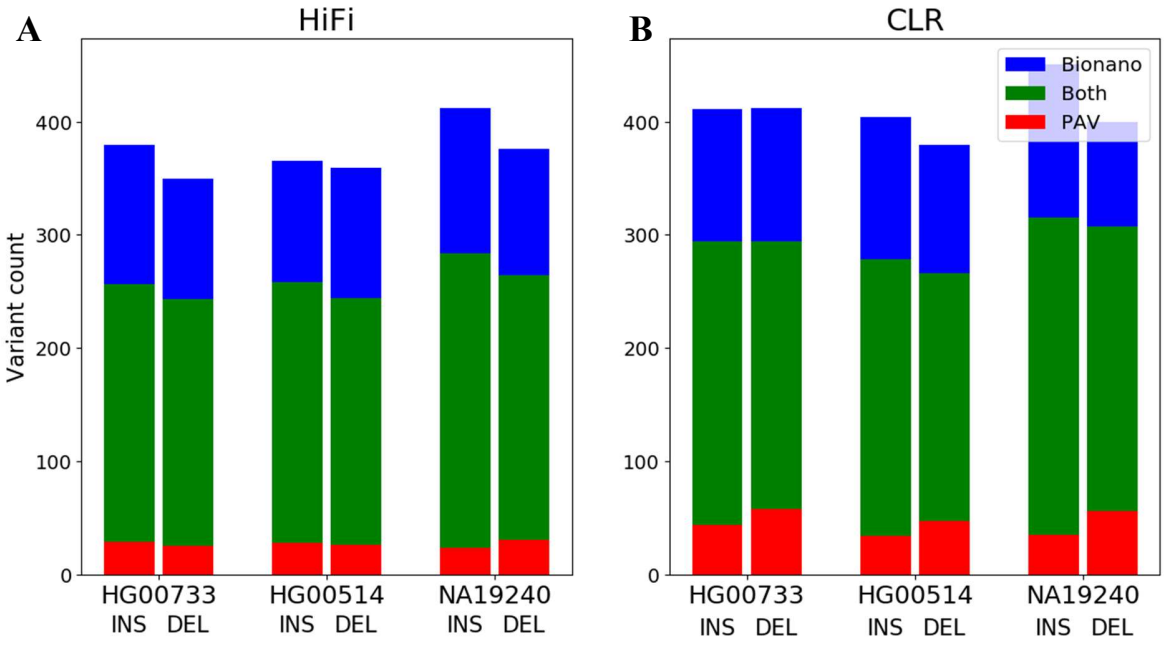


Fig. S53. Bionano and PAV intersection for SVs greater than 5 kbp

Each bar is a Venn diagram showing PAV only (red), PAV and Bionano (green), and Bionano only (blue) for SVs 5 kbp and greater in child samples for HiFi (A) and CLR (B).

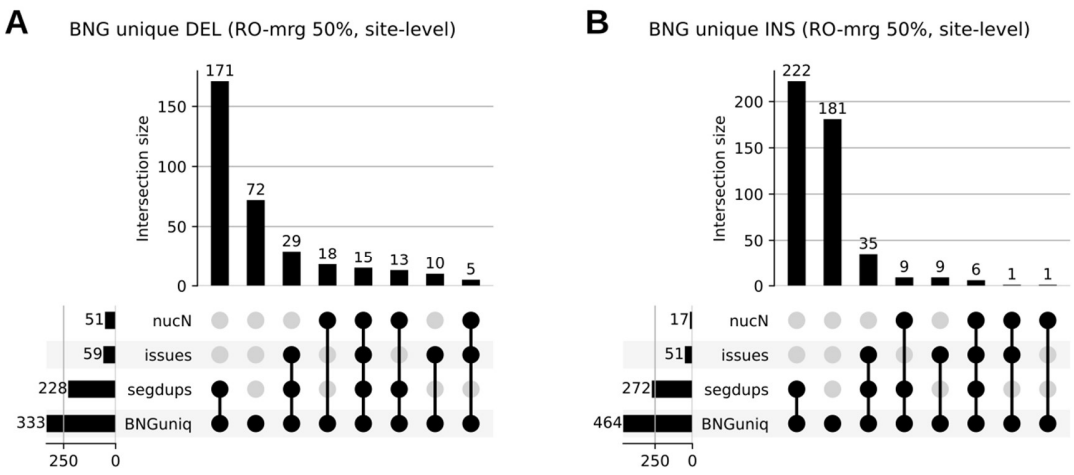


Fig. S54. Overlap between Bionano-unique SV clusters and GRCh38 regions

Upset plot of Bionano-unique (BNG unique) SV clusters. SVs (left: DEL; right: INS) were merged on the site-level (>50% reciprocal overlap; table S28 lists all Bionano-unique clusters unmerged, i.e., as individual sites) and intersected with GRCh38 annotations for SDs, GRC-defined issue regions (see Supplemental Material, Section “Reference-based analysis of phased assemblies”) and unresolved sequence (“N” gaps). Intersection sets were constructed by assigning all Bionano clusters with more than 20% overlap to the respective annotation category.

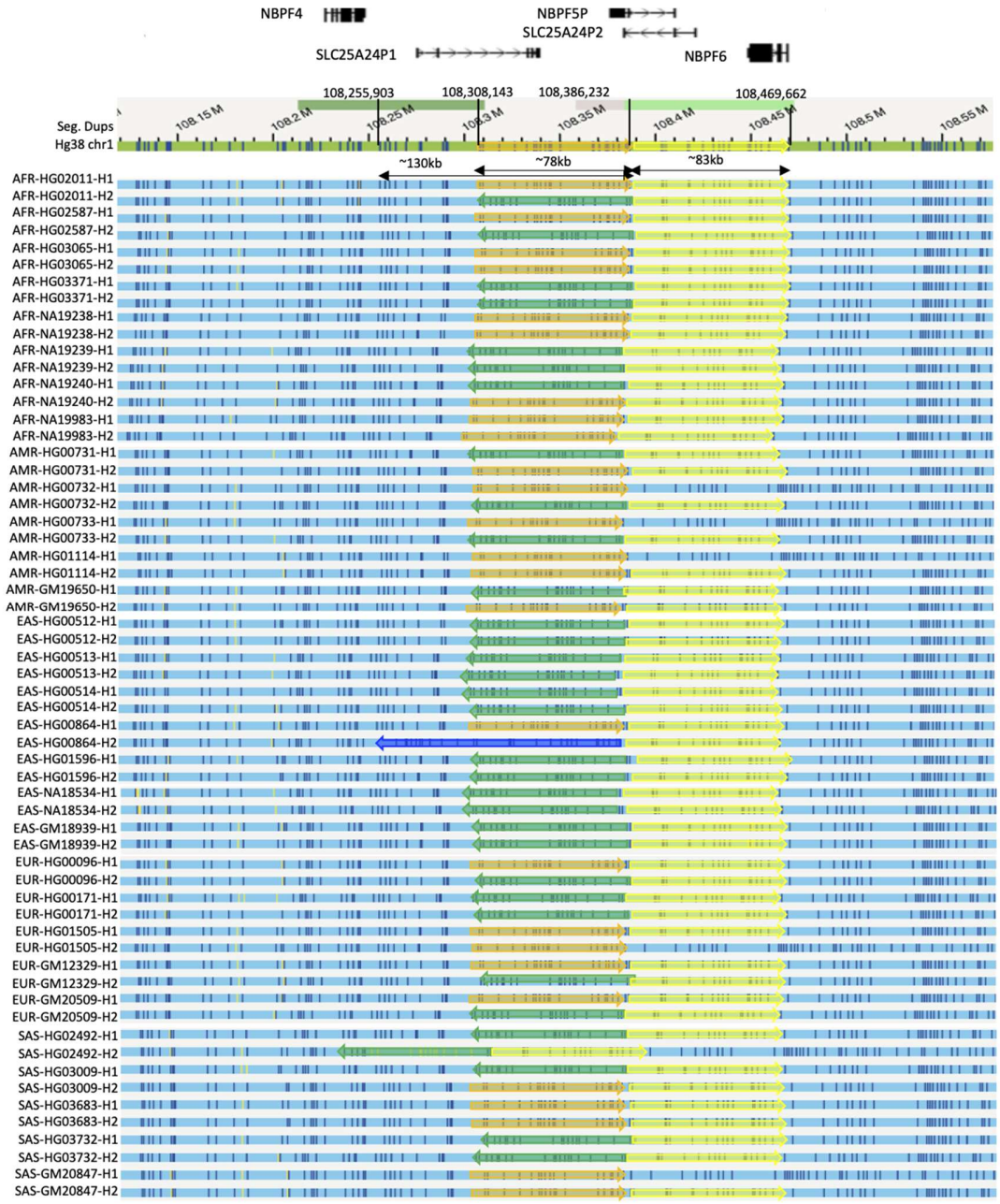


Fig. S55. Population distribution of SVs identified on 1p13.3 (108.2-108.5 Mbp)

Gene annotation (GENCODE v34) at the upper panel, and SDs between 108.2–108.3 Mbp and 108.35–108.5 Mbp. Green horizontal line represents GRCh38 reference assembly and blue horizontal lines represent sample contigs. Orange and yellow colored arrows represent the loci of inversions and deletions; green and blue arrows represent two different types of inversions.

Fig. S56. Population distribution of SVs identified on 5p14.3 (21.1–21.7 Mbp)

Gene annotation (GENCODE v34) at the upper panel, and SDs in light and dark green. Green horizontal line represents GRCh38 reference assembly and blue horizontal lines represent sample contigs. Red colored arrows represent copy number gains.

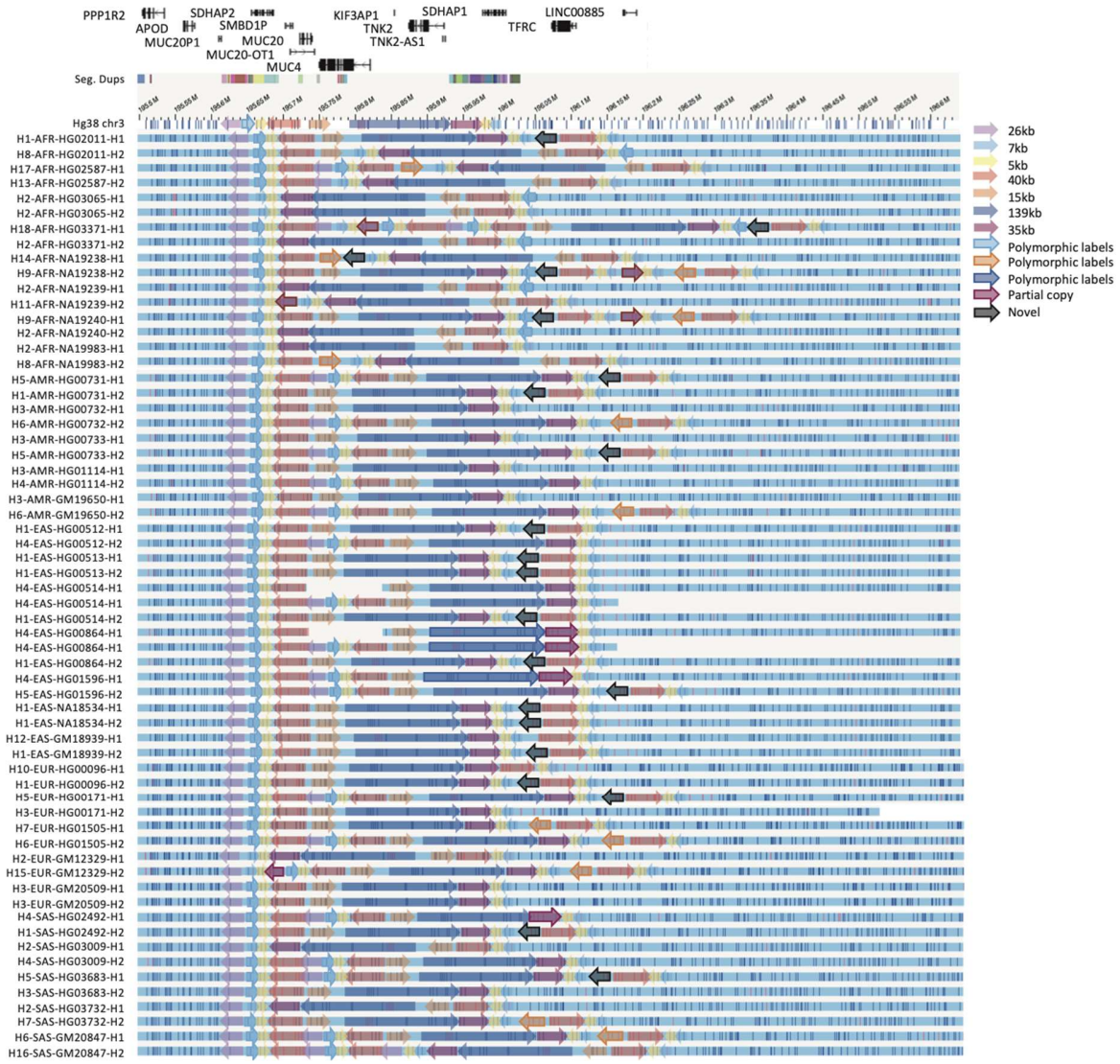


Fig. S57. Full configuration of 3q29 (195.4–196.1 Mbp)

Gene annotation (GENCODE v34) at the upper panel, green horizontal line represents GRCh38 reference assembly, and blue horizontal lines represent sample contigs. SDs are represented as colored blocks above GRCh38. Colored arrows represent different types of configurations in the 3q29 region in each sample.

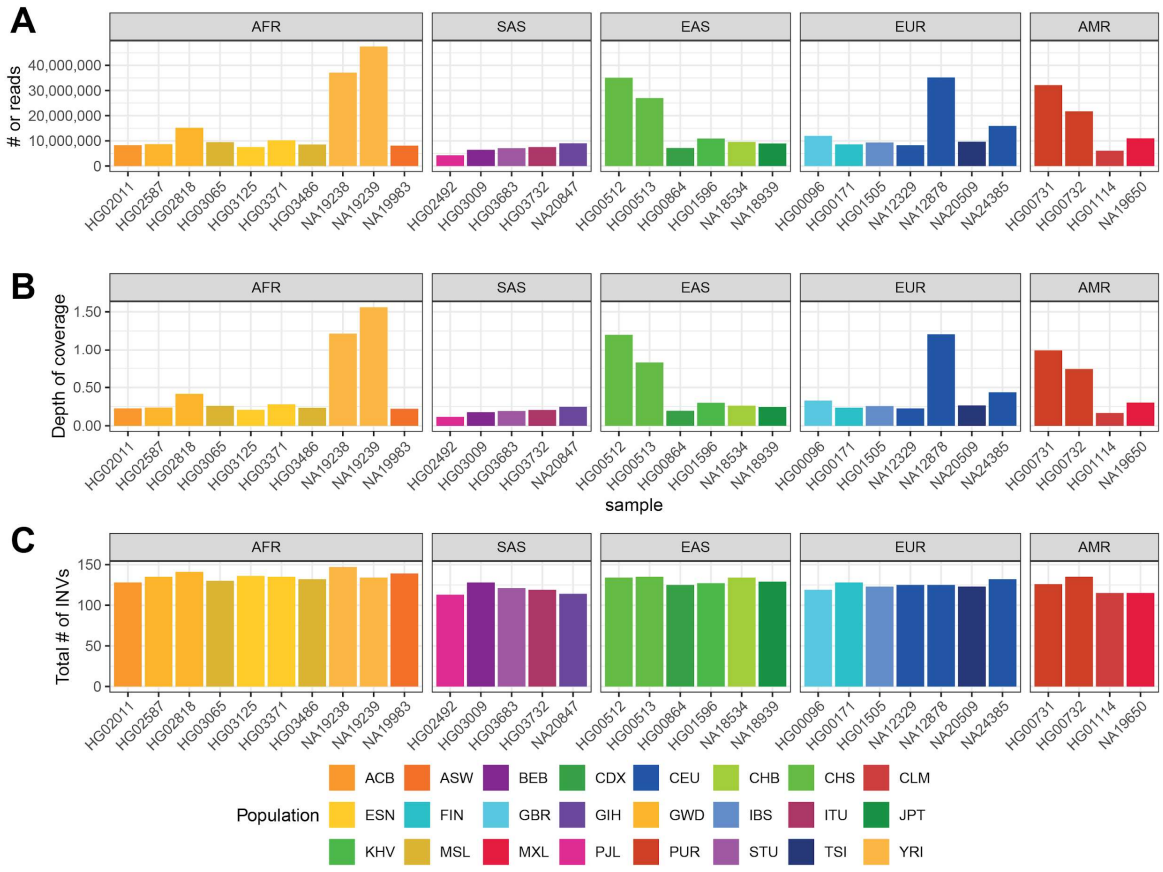


Fig. S58. Composite files summary (n=32)

Each Strand-seq composite file was created by concatenating reads across all informative Strand-seq libraries and homologs as previously described (*1*, *9*). (A) Shows the number of reads in each sample and population-specific (AFR - African, SAS - South Asian, EAS - East Asian, EUR - European, AMR - Admixed American; Tables S1, S36) composite file; (B) A barplot showing a depth of coverage of each composite file. Strand-seq data (NA19238, NA19239, HG00512, HG00513, NA12878, HG00731, HG00732) from previous studies sequenced at higher depth; (C) A barplot showing the total number of inversions detected by Strand-seq only per sample (n=32) showing that Strand-seq inversion calling is not skewed towards samples with a higher genome coverage.

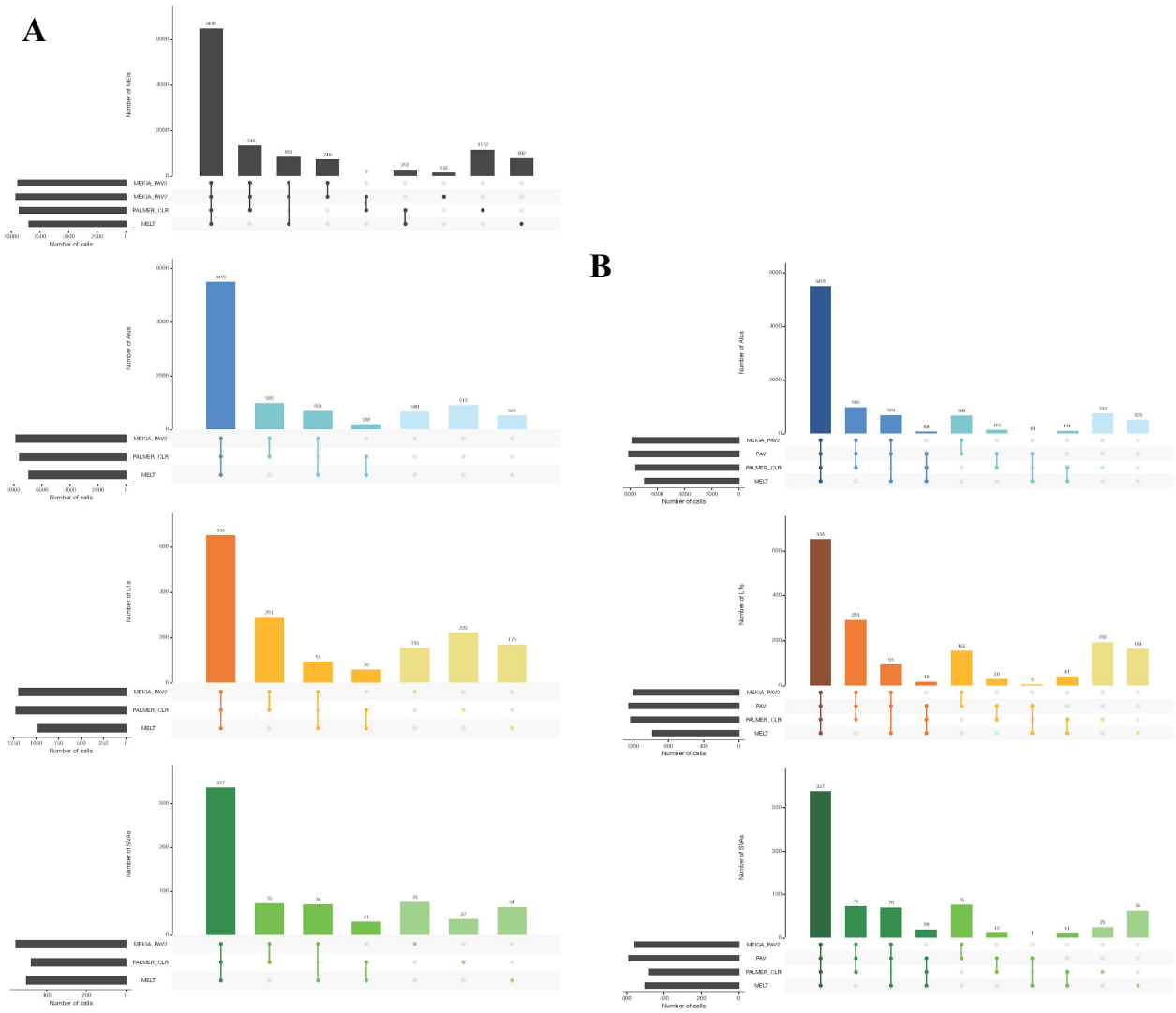


Fig. S59. UpSet plots for the integrated callset and PAV annotation in all HGSVC2 samples

A) The integration strategy was conducted on three categories of MEIs: Alu (blue), L1Hs (orange), and SVA (green), separately. The overall MEI intersection was shown by the top panel with black bars. Three calling methods are included: MELT (independent caller for Illumina), PALMER (PALMER_CLR, independent caller for PacBio mapping-based), and MEIGA_PAV (MEIGA_PAV2: the pre-filtered PAV assembly-based callset with MEIGA annotation; MEIGA_PAV3: the final PAV assembly-based callset with MEIGA annotation). B) Four callsets are included: MELT (independent caller for Illumina), PALMER (PALMER_CLR, independent caller for PacBio mapping-based), MEIGA_PAV (MEIGA_PAV2: the pre-filtered PAV assembly-based callset with MEIGA annotation), and PAV (the final PAV assembly-based callset).

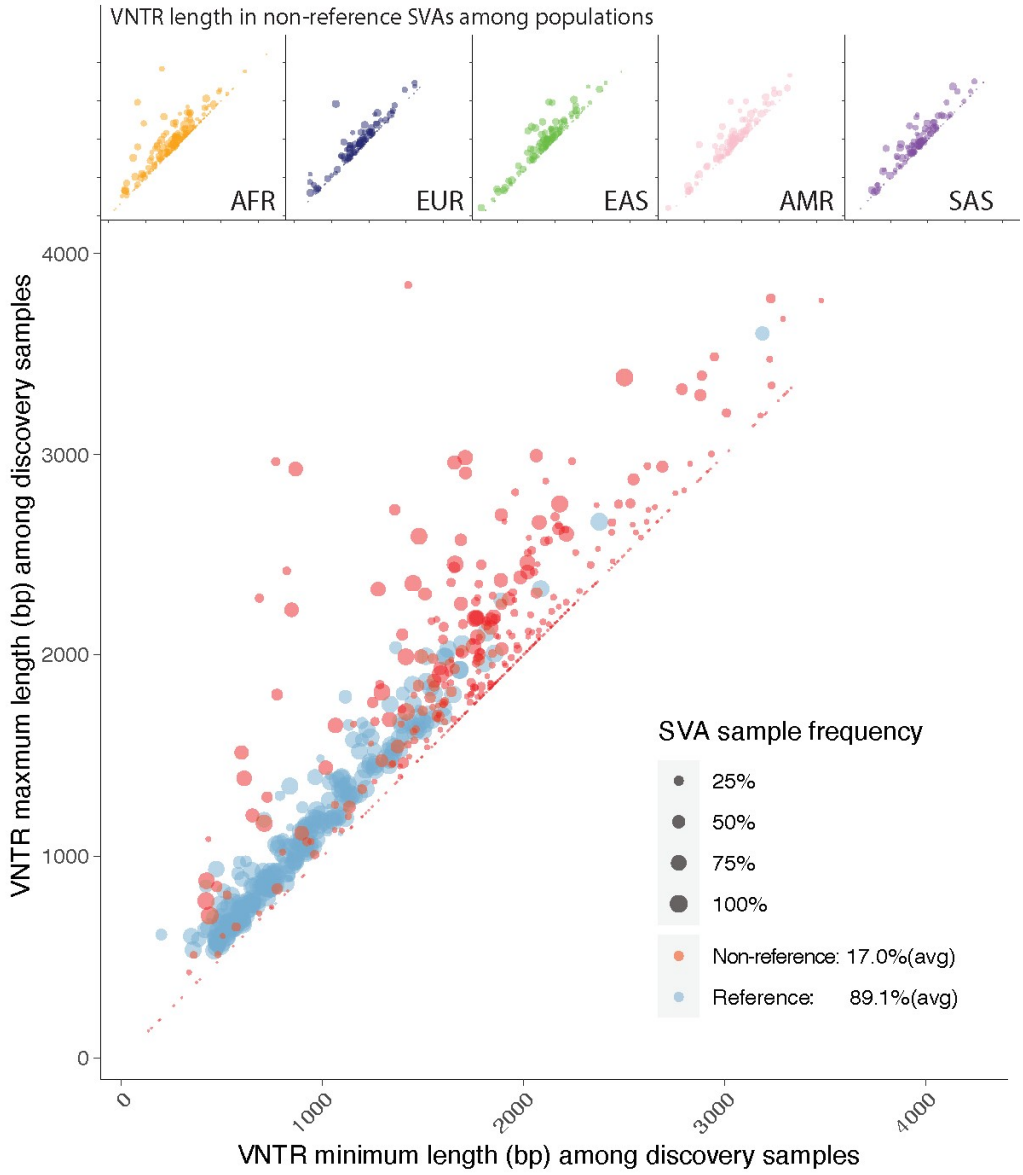


Fig. S60. Distributions of VNTR length in reference and polymorphic SVAs in each individual discovery sample

The main panel shows the minimum length (x-axis) versus maximum (y-axis) of VNTR regions in each SVA event (reference one as blue dot and polymorphic one as red dot). The size of dot represents the sample frequency of SVAs among discovery samples in HGSVC. The upper panel shows the separate distributions of VNTR regions in SVAs for five superpopulations: AFR (African, yellow), EUR (European, dark blue), EAS (East Asian, green), AMR (Admixed American, pink), and SAS (Southeast Asian, purple).

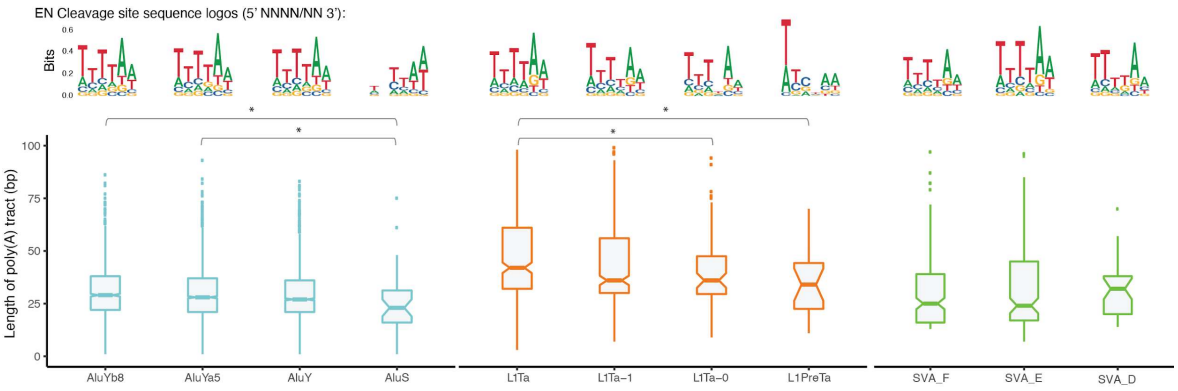


Fig. S61. Box plots of poly(A) tract and sequence logos of endonuclease cleavage sites for MEI subfamilies

The lower panel shows the length distributions for the polymorphic MEI families and their main subfamilies in the human populations: Alu (AluYb8, AluYa5, AluY, and AluS, blue), L1Hs (L1Ta, L1Ta-1, L1Ta-0, and L1PreTa, orange), and SVA (SVA_F, SVA_E, and SVA_D, green). The significant difference for poly(A) tract length between the subfamilies is indicated by an asterisk (p-value < 0.05, Student's t-test, two-sided). The upper panel shows the EN cleavage site sequence logos for the corresponding subfamilies in the x-axis.

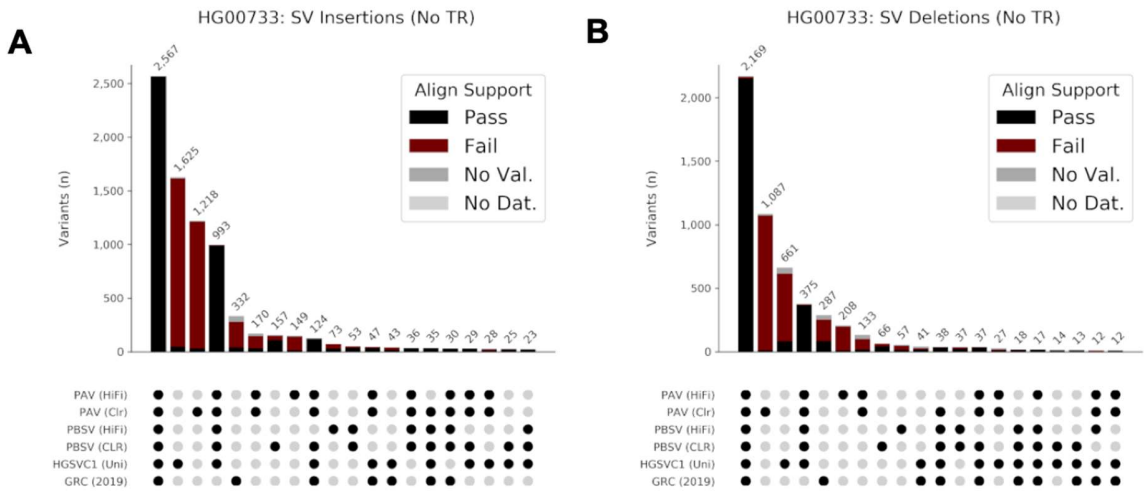


Fig. S62. PAV concordance among callers for HG00733 outside tandem repeats

Calls outside annotated tandem repeats from multiple sources for HG00733 were merged with the three-step approach. Shown are PAV with minimap2 alignments (PAV), PBSV, Chaisson 2019 unified callset (*1*) and Audano 2019 (*4*) for (A) SV insertions and (B) SV deletions. Bars are colored by subseq support.

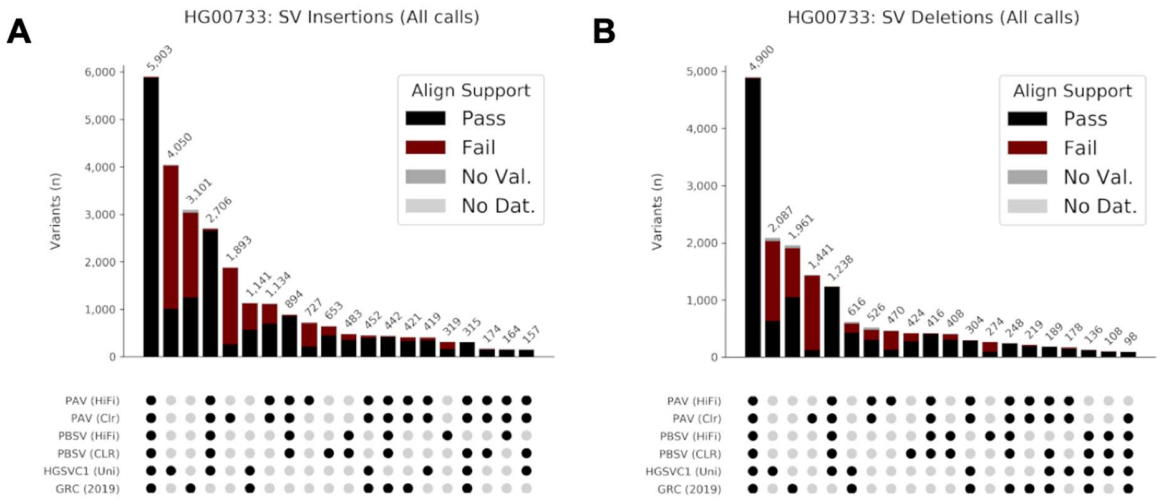


Fig. S63. PAV concordance among callers for HG00733

Calls from multiple sources for HG00733 were merged with the three-step approach. Shown are PAV with minimap2 alignments (PAV), PBSV, Chaisson 2019 unified callset (*1*) and Audano 2019 (*4*) for (A) SV insertions and (B) SV deletions. Bars are colored by subseq support. Variants supported by more than one caller also tend to have more subseq support. PAV calls from CLR contain more apparent false positives.

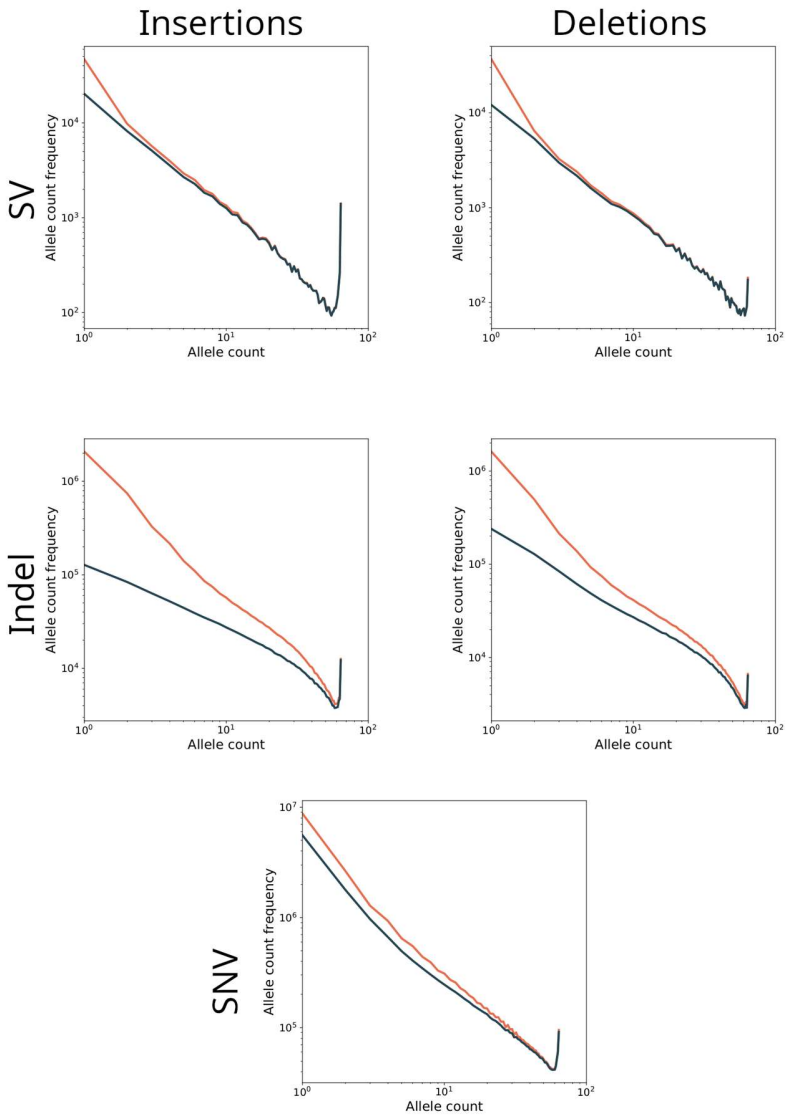


Fig. S64. Excess singleton rate is reduced by callset filtering

When allele counts (horizontal axis) are plotted against the frequency for each count (vertical axis), the distribution should be approximately linear. Before filtering (orange), an increase in singletons was observed, which was corrected by applying “+1” (SVs) and “+2” (indels and SNVs) filters to the callset (blue). Filtered SVs are enriched for low-frequency events. Higher indel error rate from long-read sequences produces more calls that persist from multiple callsources, but without more orthogonal support, these calls are over-filtered. Over-filtering is also likely with SNVs, but to a lesser extent than indels.

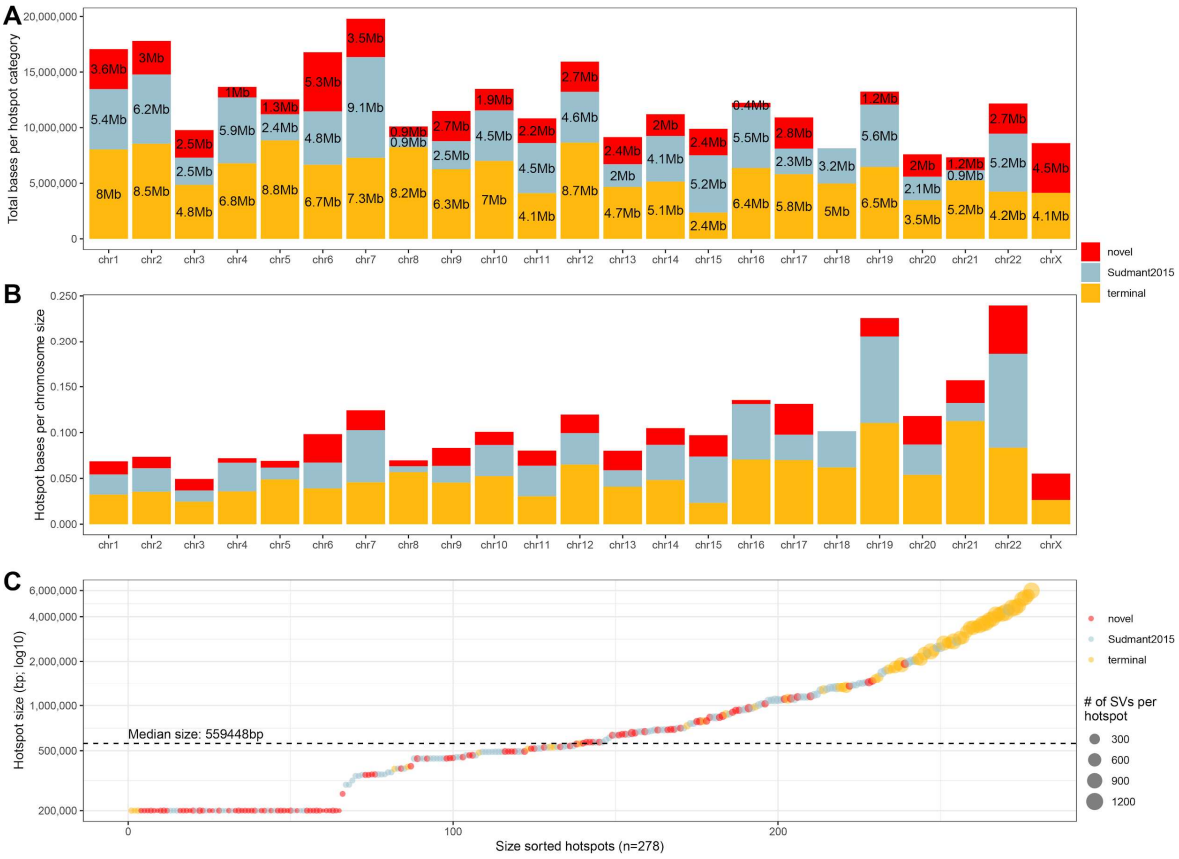


Fig. S65. Summary of detected SV hotspots per chromosome

(A) A total size of detected hotspots per category ('novel', 'Sudmant2015' and 'terminal') and per chromosome (x-axis). (B) Number of hotspot bases per category and per chromosome normalized by the chromosome size. (C) Size distribution of all detected SV hotspots (n=278) ordered from smallest to largest. Size of each dot reflects the number of SVs in each hotspot. Median hotspot size is highlighted by a horizontal dashed line.

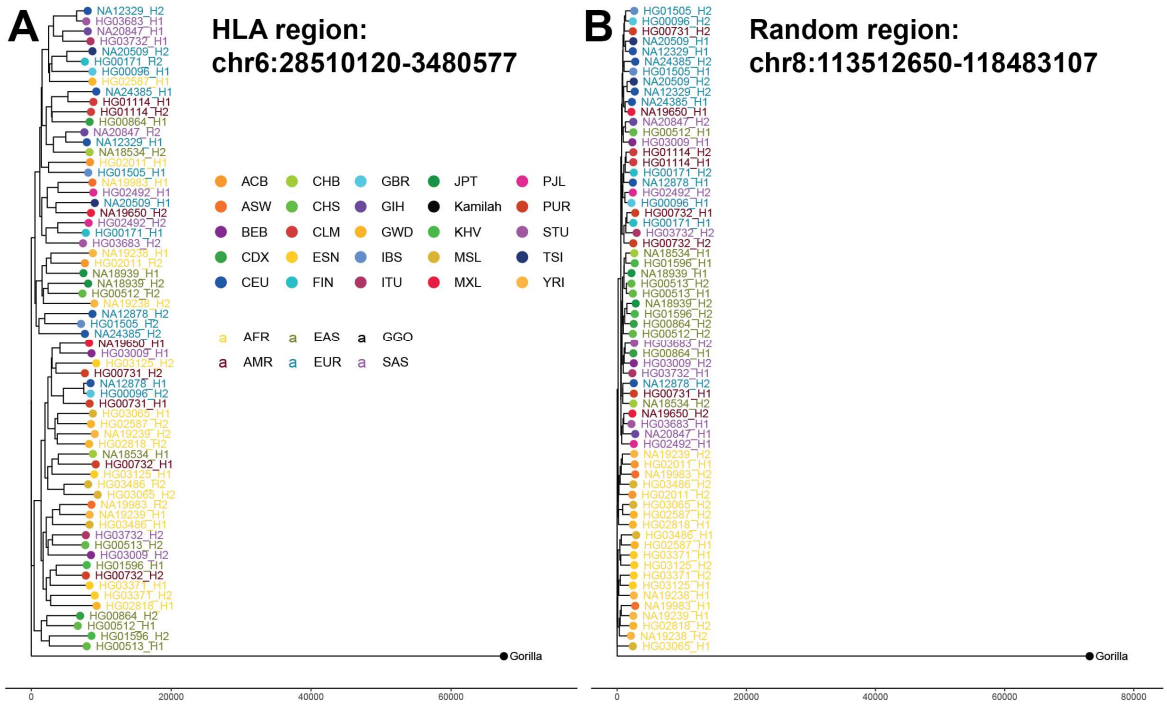


Fig. S66. Evolutionary distances inside and outside of the HLA region

A neighbor-joining evolutionary tree based on distances calculated between all phased assemblies over the (A) HLA region (chr6:28,510,120-33,480,577) and (B) randomly chosen region (chr8:113,512,650-118,483,107). Each dot is colored based on the 1KG population identifier and color of the label defines superpopulation. We used a gorilla (GGO) squashed assembly (Kamilah individual) as an outgroup.

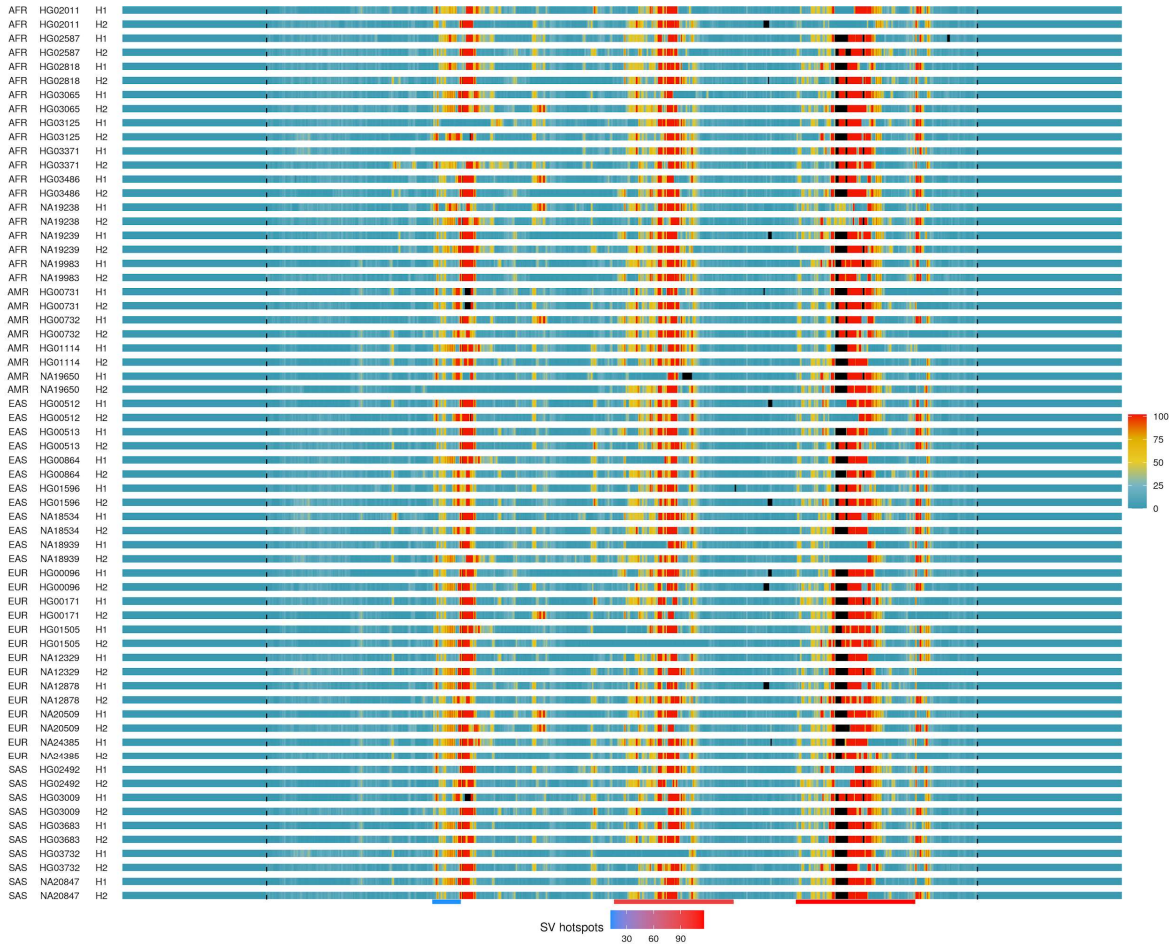


Fig. S67. Count of alternative alleles per haplotype for the HLA region

Each row represents binned counts (bin size – 10 kbp, stepsize – 1 kbp) of alternative alleles in each assembled haplotype with respect to GRCh38 as a reference. The number of alternative alleles in each bin are reflected in a color scheme going from blue to red (see legend). The HLA region (chr6:28,510,120-33,480,577) is highlighted by dashed lines. Regions of less reliable contig mapping with respect to GRCh38 are highlighted by black color. At the bottom of this heatmap, we show previously detected SV hotspots that overlap with the HLA region. The number of SVs in each hotspot is reflected by the color scheme going from blue to red.

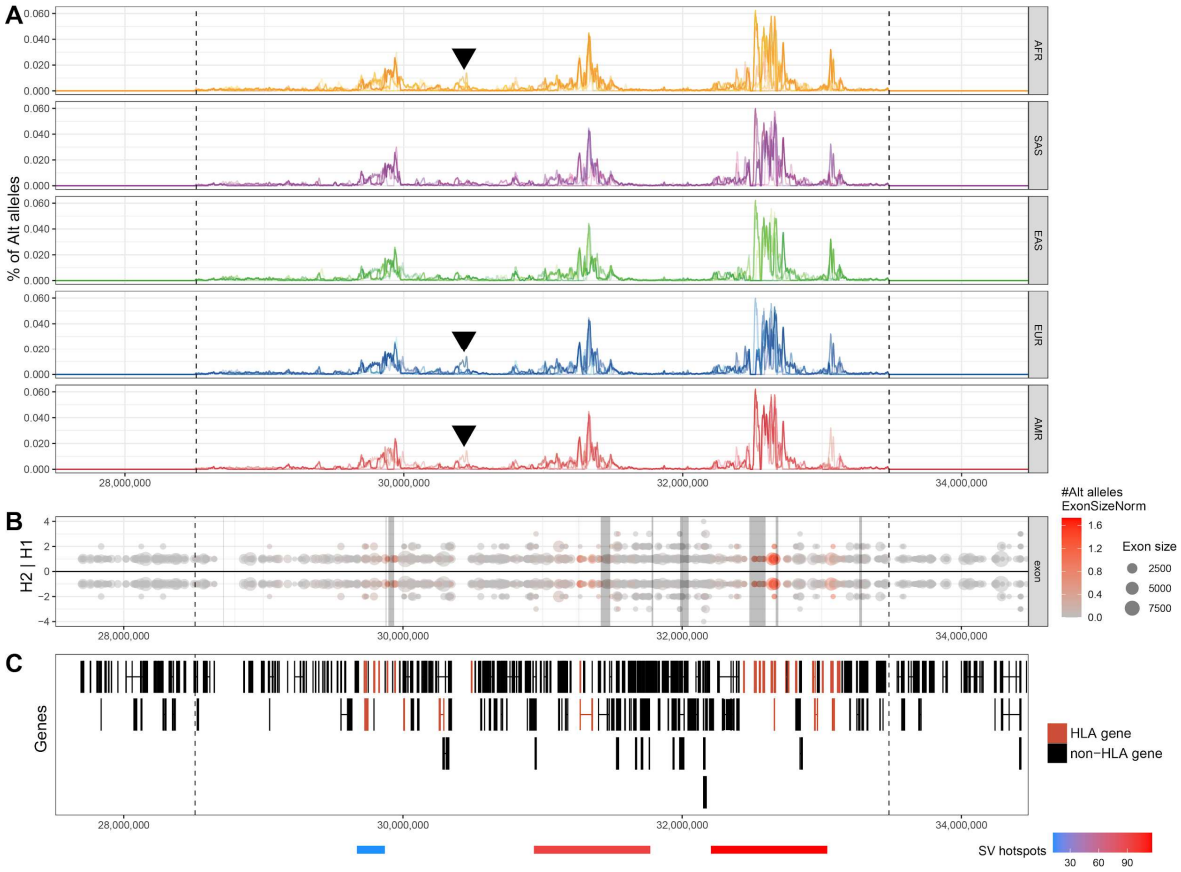


Fig. S68. Summary of genetic variability over the HLA region

(A) Binned (bin size – 10 kbp, stepsize – 1 kbp) percentage of alternative alleles per haplotype. Haplotypes from each superpopulation (AFR - African, SAS - Southeast Asian, EAS - East Asian, EUR - European, AMR - Admixed American) are grouped together and plotted as a single line. Transparency (alpha) level serves to highlight differences within and across superpopulations. Black arrowheads point to the extent of haplotype diversity unique to AFR, EUR and AMR populations. (B) Number of alternative alleles per exon overlapping the HLA region. The size of each exon is reflected by the size of each dot. The number of alternative alleles per exon is highlighted by a color scheme going from gray to red. (C) A gene model obtained from R package 'TxDb.Hsapiens.UCSC.hg38.knownGene'. The position of each exon is plotted as a black or red rectangle for non-HLA and HLA genes, respectively. At the bottom, we show previously detected SV hotspots that overlap with the HLA region. The number of SVs in each hotspot is reflected by a color scheme going from blue to red.

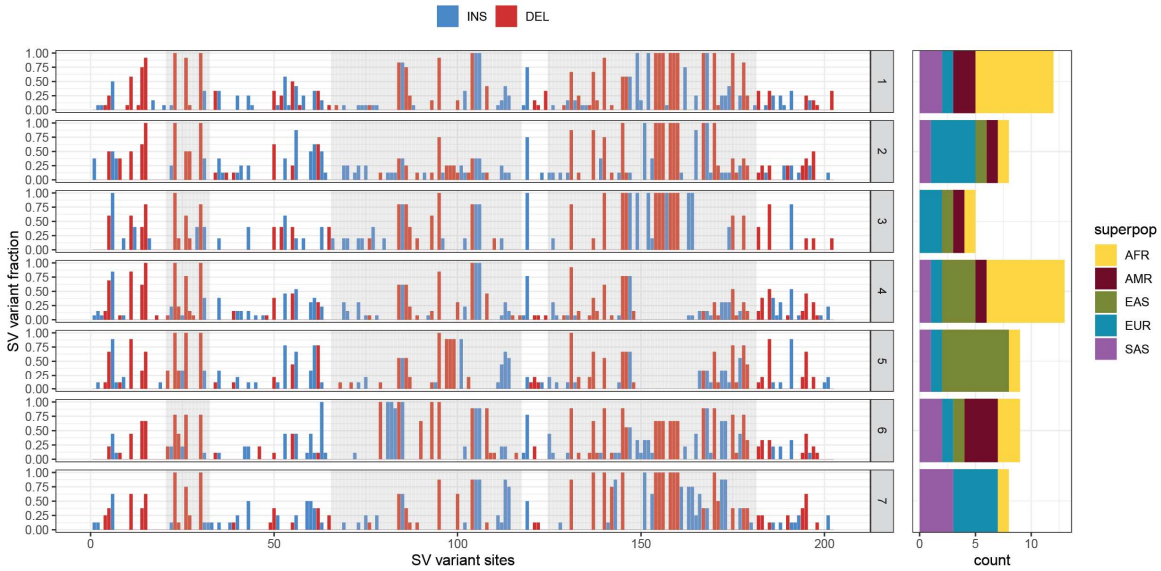


Fig. S69. Summary of HLA haplotypes with a similar SV distribution

Left plot shows the fraction of SVs per site for each cluster defined by k-means clustering in Fig. S15. Height of each bar represents frequency of a given variant and the given position across all SV haplotypes assigned to each cluster (n=7). Regions of previously defined SV hotspots are highlighted by vertical gray bars over each SV that overlaps with a hotspot region. Barplot on the right counts the superpopulation of origin for each haplotype in each cluster.

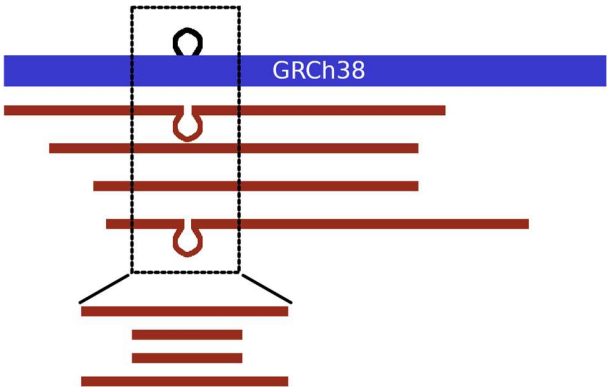


Fig. S70. Subseq illustration

Reads (red) aligned to the reference (blue) with an SV insertion (loop). A region surrounding the breakpoints is chosen (dashed box) with larger regions for larger SVs. All reads traversing that region (i.e., touching both ends) are extracted, and the length of the read in that region is determined by parsing the CIGAR string. In this example, regions extracted from two reads (top and middle) support a heterozygous insertion, and two reads (middle two) do not support it.

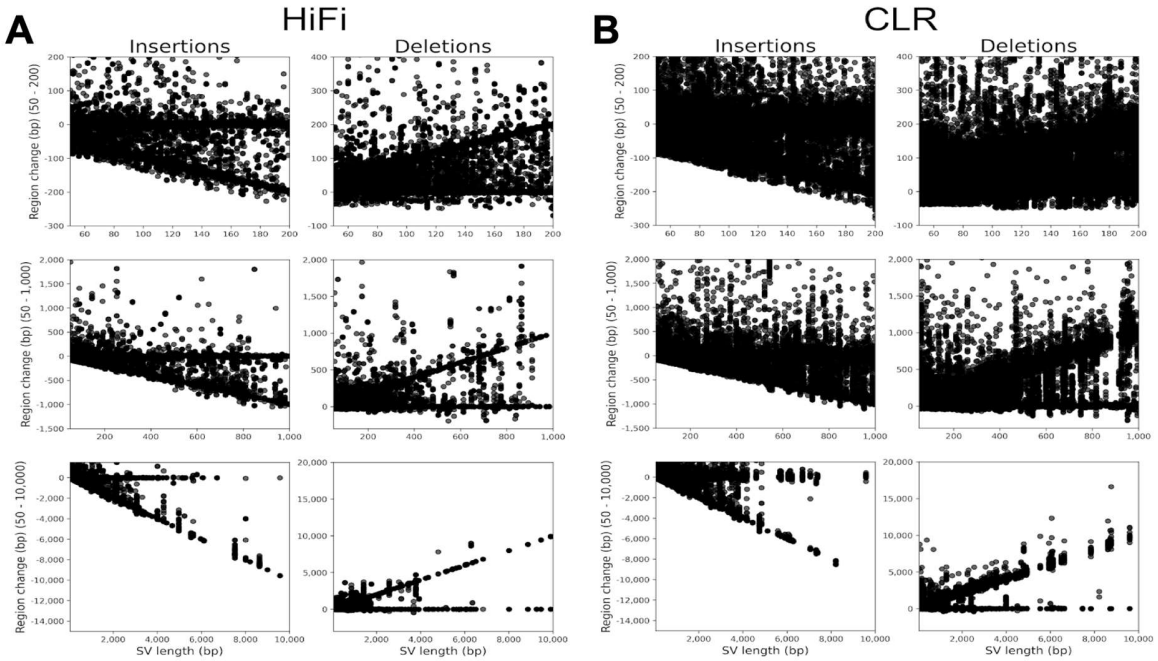


Fig. S71. Subseq region lengths deviate with SV length with less variation from HiFi

A region around each SV was extracted from raw read alignments. For any read spanning the region, the size of the region within the read was found and the expected deviation was computed assuming the SV was present. If the SV is supported, then the expected deviation must be within 50% of the SV length (±0.5 * SVLEN). We ran validations for HiFi (A) and CLR (B) for all samples (HG00733 shown). Each point in the figure is one read alignment with the SV length (x-axis) and the size deviation around the SV for a single read (y-axis). Reads supporting an SV call cluster along y=0. Reads not supporting the variant cluster along -SVLEN (insertions, alignment is shorter than expected if SV was present) and +SVLEN (deletions, alignment is longer than expected if SV was present). Figure shows SVs 50-200 bp (top), 50-1,000 bp (middle), and 50-10,000 bp (bottom). Each panel shows a random sample of 2,000 SV calls.

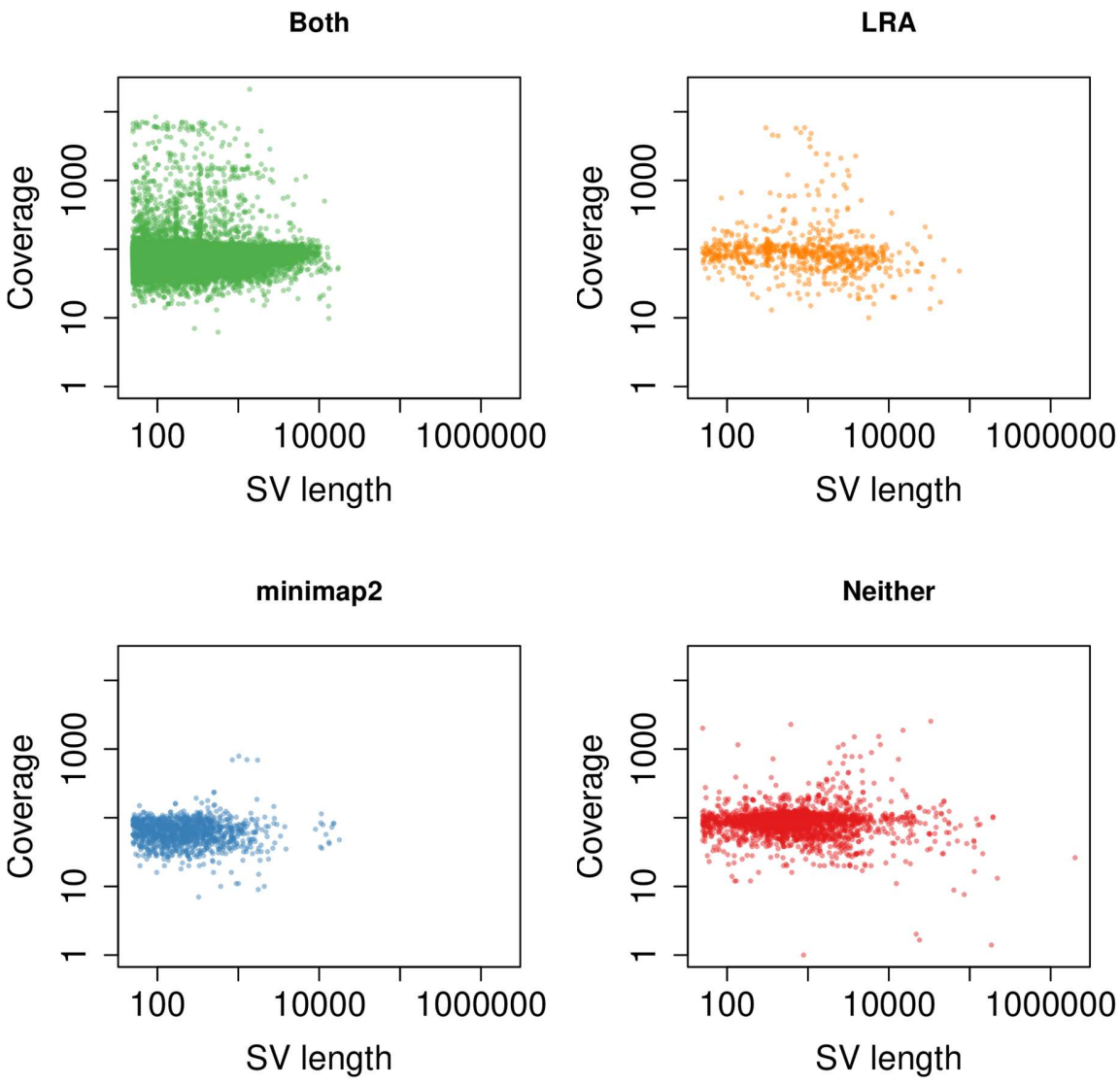


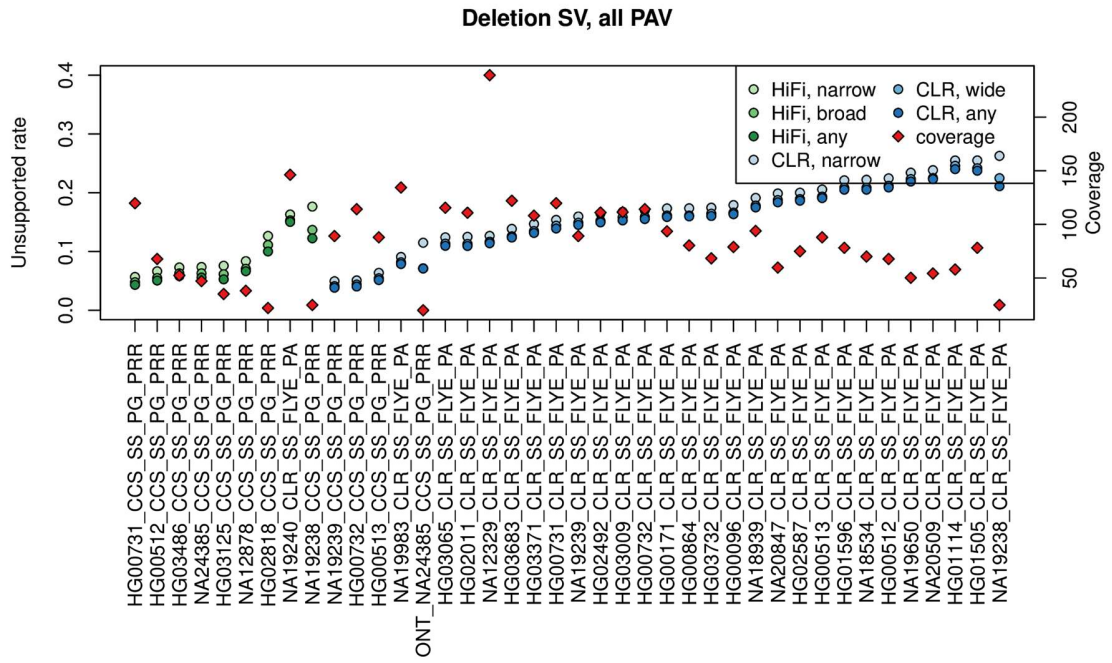
Fig. S72. SV support from raw reads

SV support from alignment of reads by LRA and minimap2, shown by length of read versus read support for SV calls made from a CLR assembly of the HG00733 genome. Top left calls with both LRA and minimap2 support. Top right calls with only LRA support. Bottom left calls with only minimap2 support. Button right calls supported by neither method. The raw-read support is used to determine the FDR of all PAV calls.

Fig. S73. Example of SV calls with and without support

Two deletion variant calls in HG00733 are shown in the region chr1:30,374,895-30,379,391 (bottom track). The left side call is a 52 base-pair deletion that has 66 alignments by minimap2, and 57 alignments by LRA supporting the call. For clarity, HiFi alignments are shown rather than CLR. The right side call is a 408 base-pair deletion that has no alignments from either LRA nor minimap2 supporting the call.

A



B

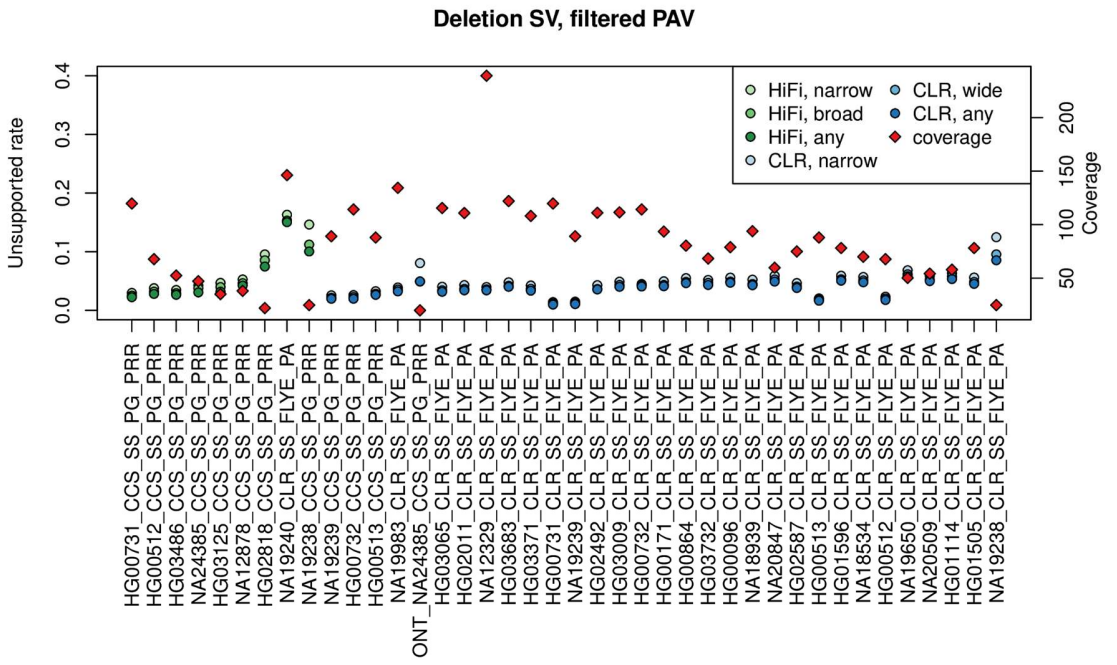
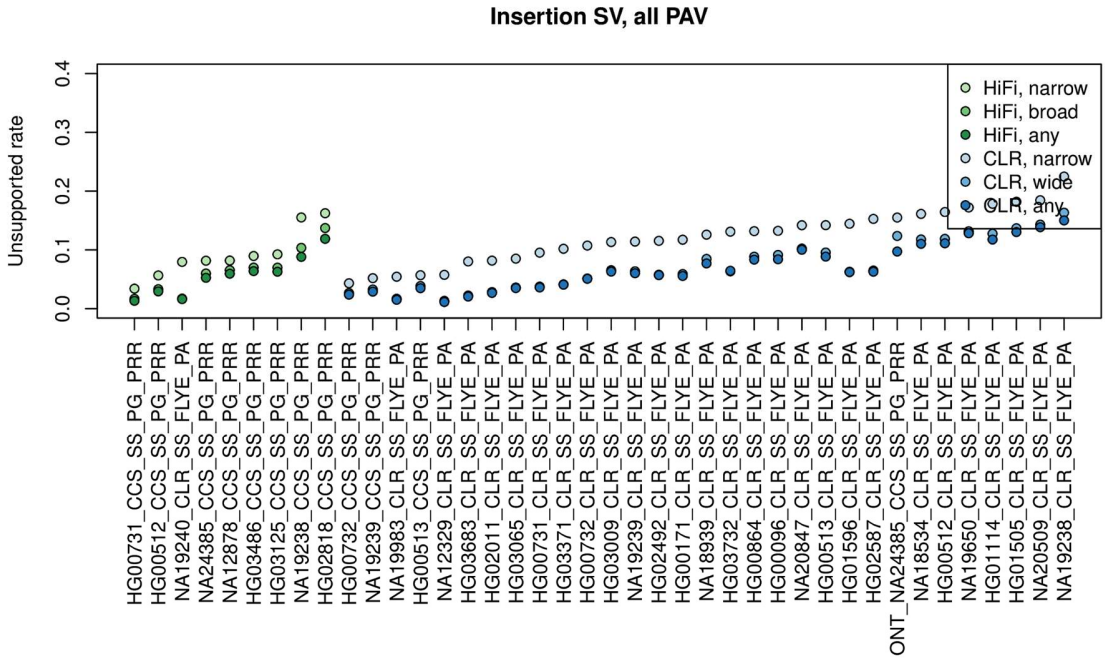


Fig. S74 (continued on next page)

C



D

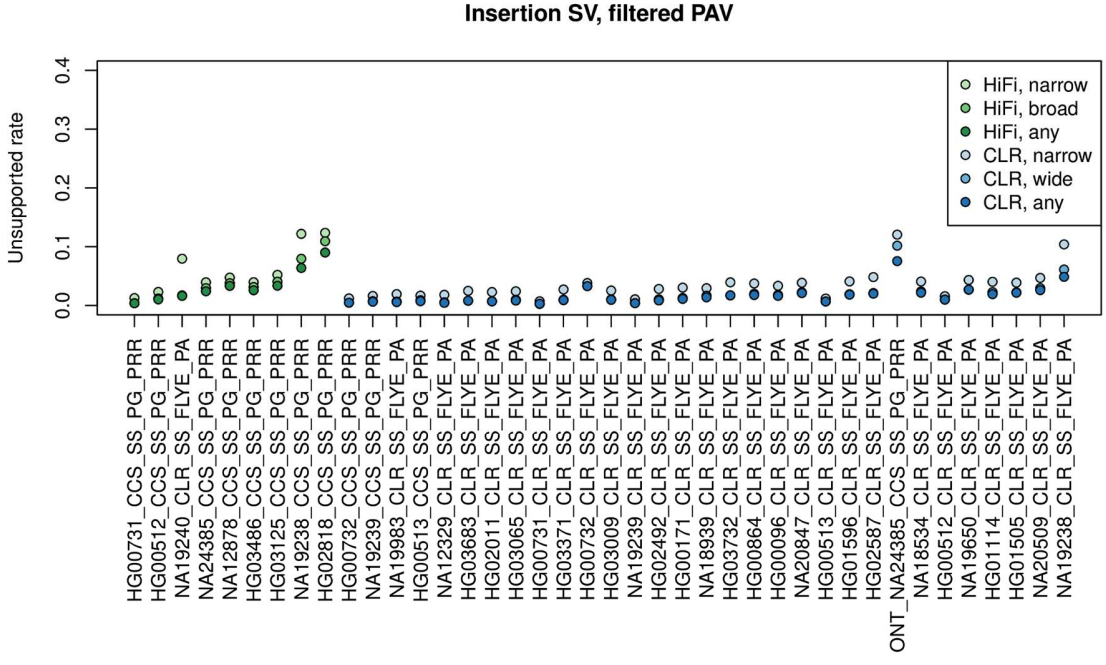


Fig. S74. Per-assembly rate of supported variants (caption next page)

Fig. S74. Per-assembly rate of supported variants

The fraction of unsupported SVs for HiFi (green) and CLR (blue) assemblies for variants detected by PAV. Each plot shows three levels of stringency for how reads supporting an SV are calculated: narrow, wide, and any. Narrow support is defined as an SV in a read of the same type within 1 kbp and of length within 50% of the length of the assembly-based SV. Wide support has the same size constraint but allows for 10 kbp distance of the SV breakpoint. Any support is within 10 kbp and any SV length (>50 bases). The latter category is not a reasonable approach to detect read-based support for SV assembly but represents the upper bound for which variants may be supported by reads. (A) Support for deletion variants detected by PAV. Red points indicate the coverage of reads aligned to the reference. The coverage correlates negatively with supported reads (HiFi r^2 = -0.739, p=0.002; CLR r^2=-0.682, p=1.24E-5); however, this is likely a combined effect of assembly quality and lower number of reads that may support a variant. (B) Support for deletion variants in the PAV calls after filtering. These variants correspond to the deletion SVs considered in this study. (C) Support for insertion variants from PAV. (D) Support for insertion variants from the filtered PAV callset. These variants correspond to the insertion SVs considered in this study.

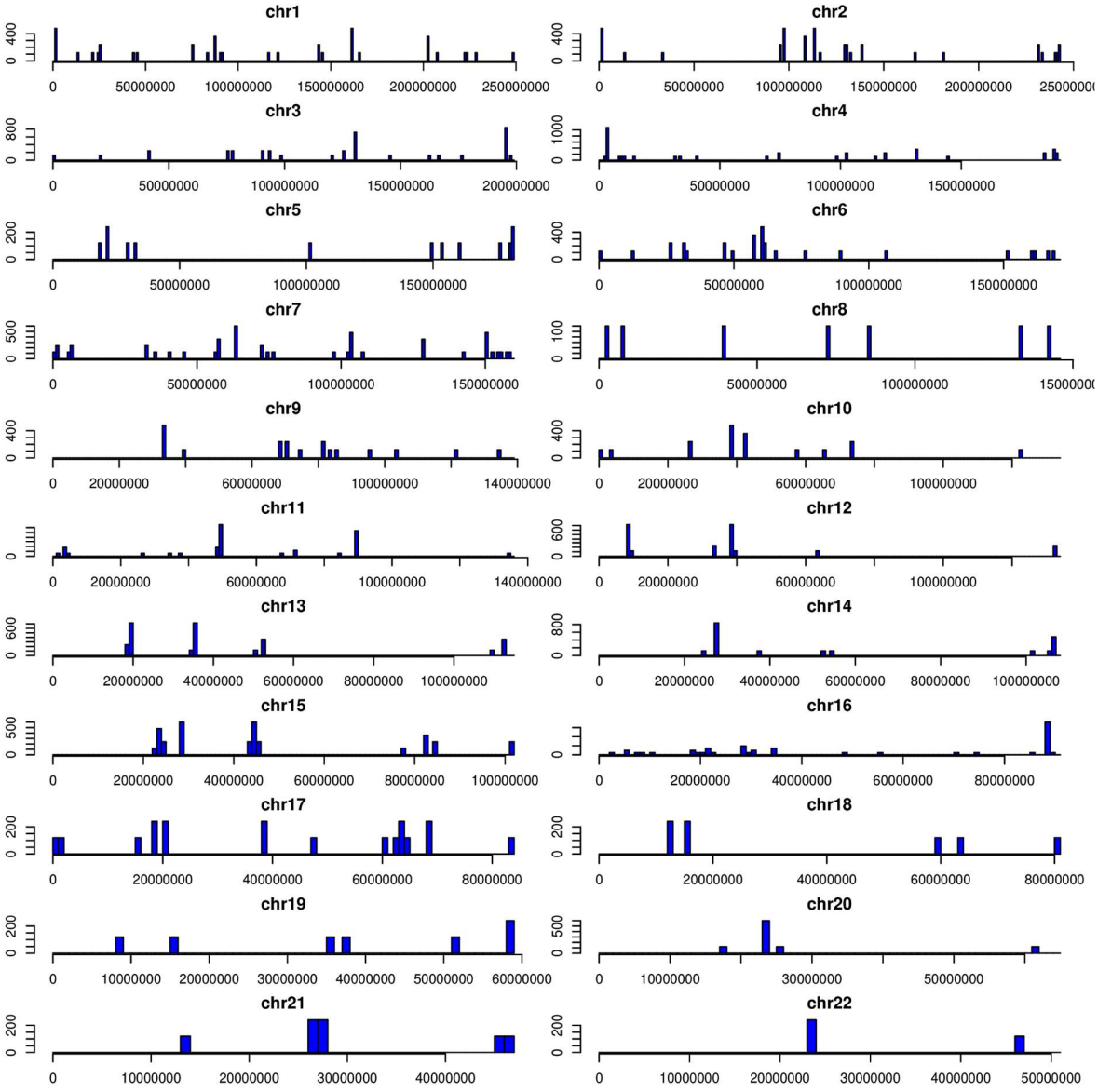


Fig. S75. Genomic distribution of unsupported variants detected by HiFi assemblies

Counts are aggregated over all assemblies and variant types for variants produced by PAV, after applying quality control filtering Each bin represents the number of unsupported SVs per 1 Mbp per genome.

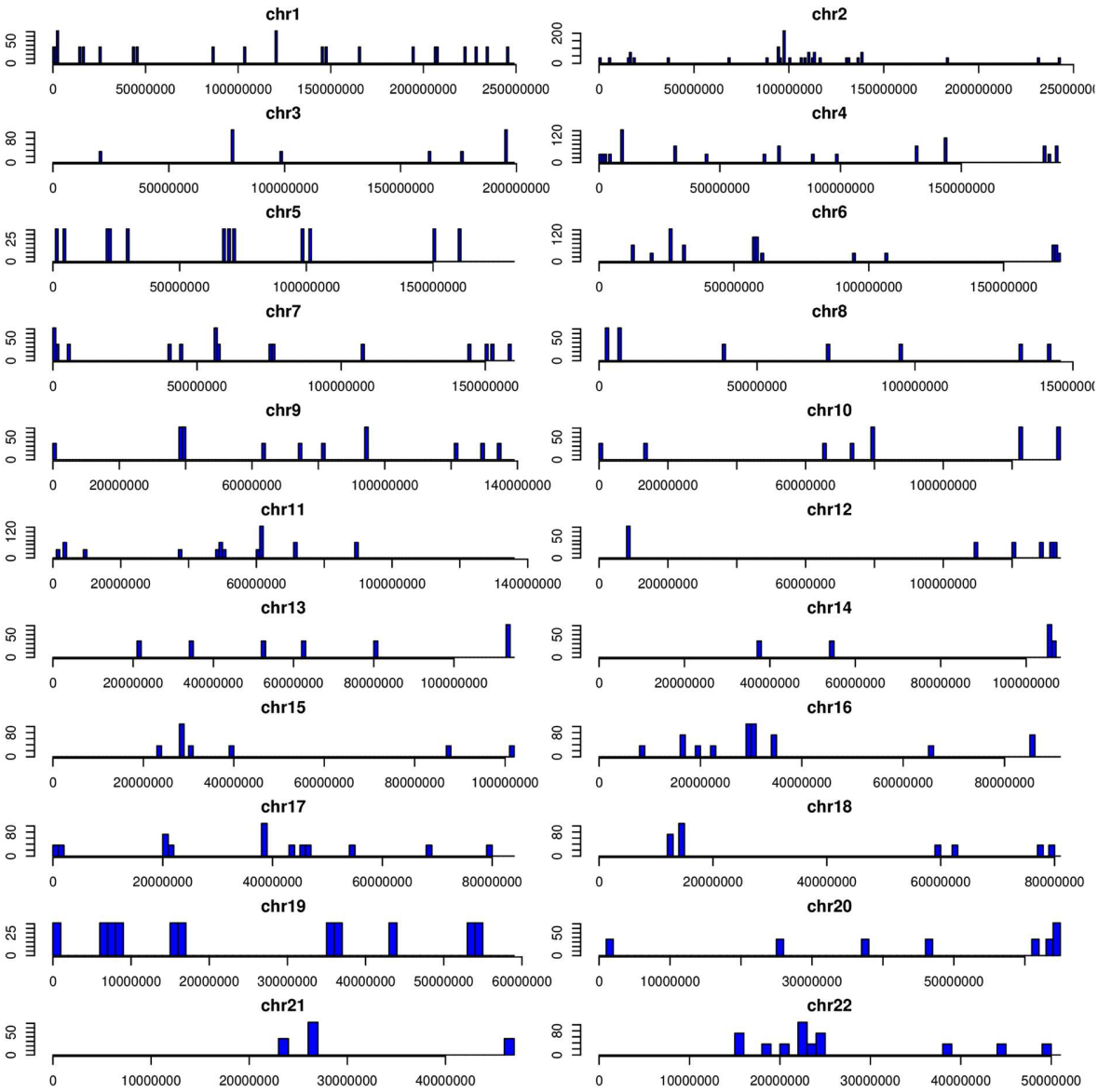


Fig. S76. Genomic distribution of unsupported variants detected by CLR assemblies

Counts are aggregated over all assemblies and variant types for variants produced by PAV, after applying quality control filtering. Each bin represents the number of unsupported SVs per 1 Mbp per genome.

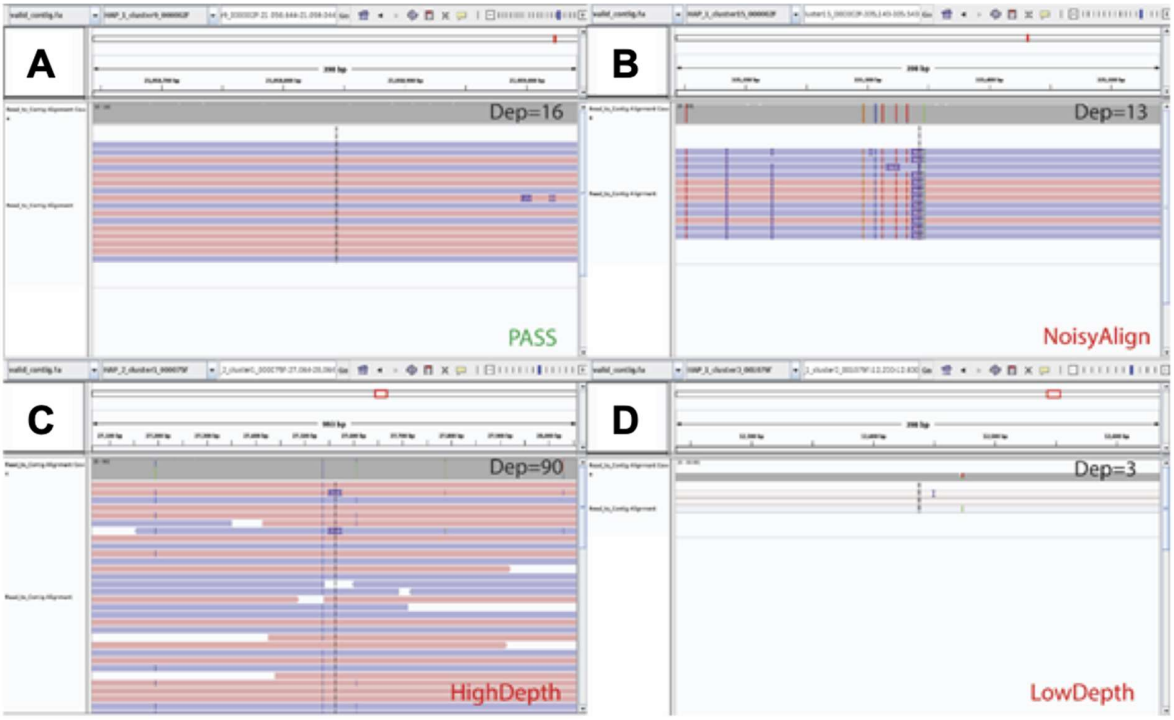


Fig. S77. Examples of inspector quality control

Examples of read_to_contig alignment patterns for four types of SVs in HG00733 (HiFi) with mean alignment depth at ~16X. (A) A “PASS” SV call with no indel or clipped alignment at its breakpoint; (B) A “NoisyAlign” deletion call with 468 bp INS signals and several mismatches in read alignment; (C) A “HighDepth” SV call with alignment depth of 90X; (D) A “LowDepth” SV call with alignment depth of 3X at its breakpoint.

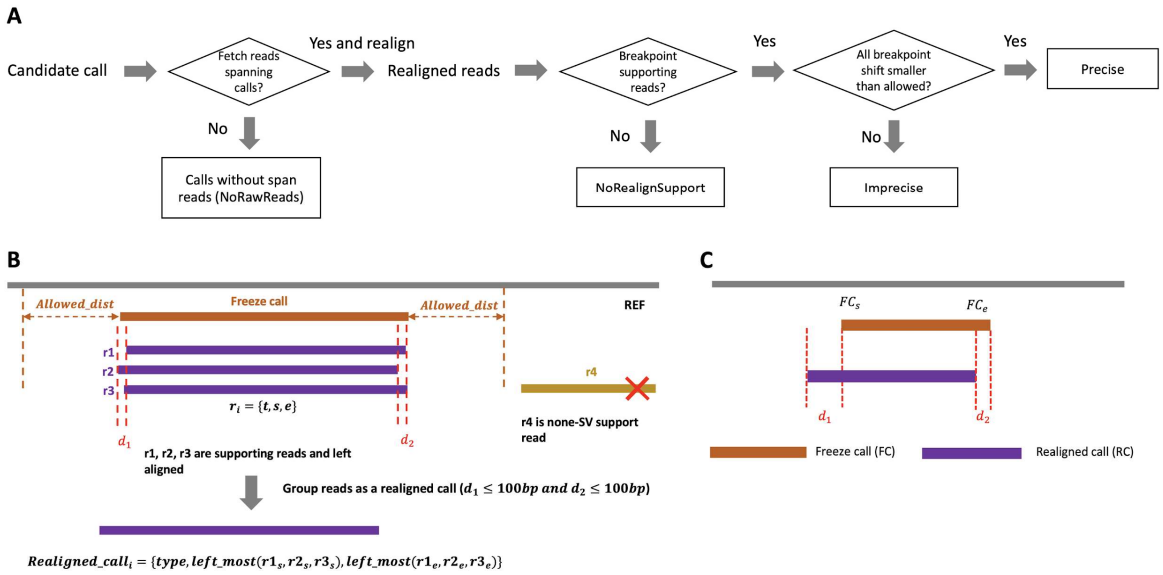


Fig. S78. Workflow of SV QC

(A) The general workflow of SV QC, classifying the raw PAV calls into “NoRawReads”, “NoRealignSupport”, “Imprecise” and “Precise”. (B) Creating realigned calls through MUMMer-based realignment. If realigned breakpoints from all SV spanning reads are outside of the allowed_dist (default is 500 bp), this SV is marked as “NoRealignSupport”. (C) The diagram shows the process of labeling “Precise” and “Imprecise” calls. For a "Precise" PAV call, the breakpoint position difference between realigned call and this PAV call (d1 and d2) should be smaller than the maximum allowed breakpoint shift threshold. Otherwise, this PAV call is labeled as "Imprecise".

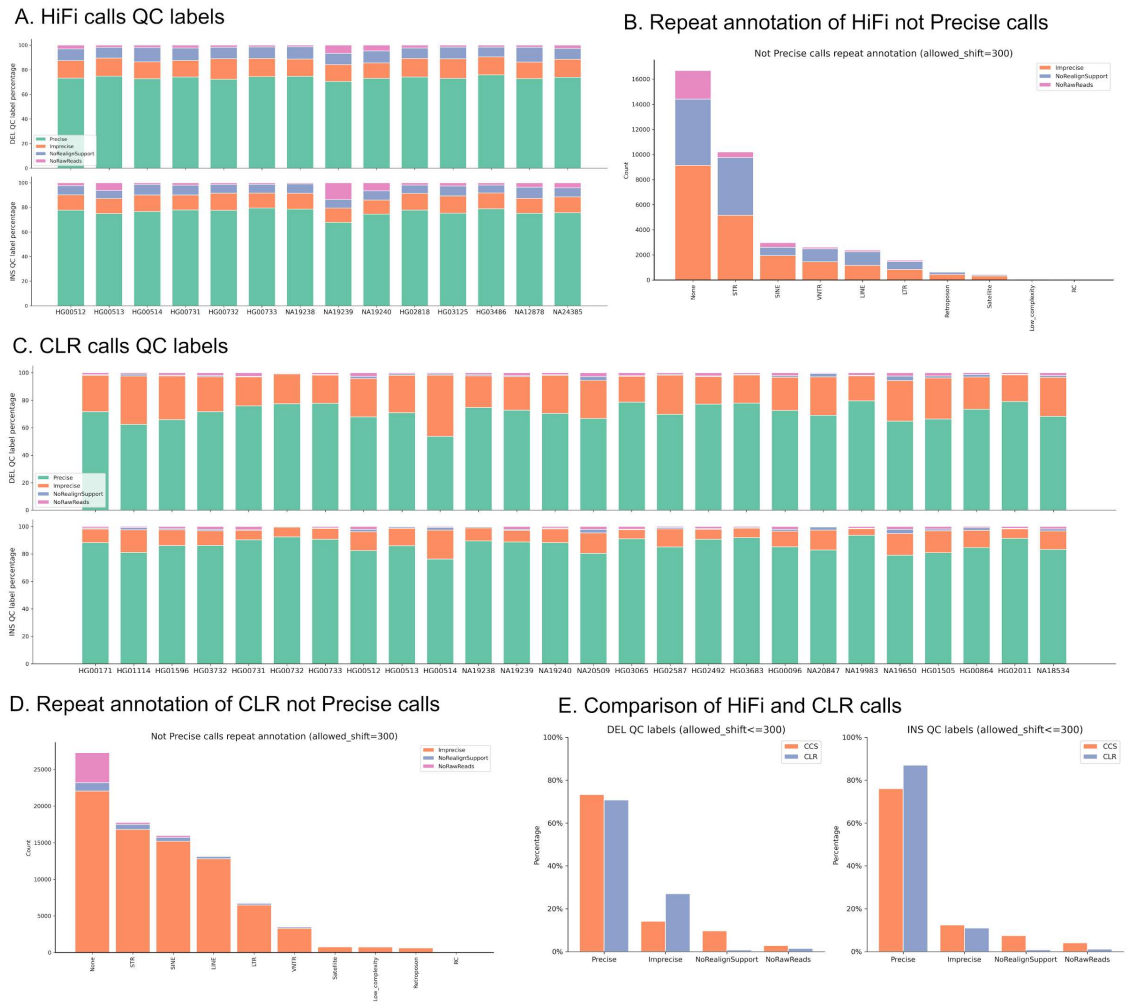
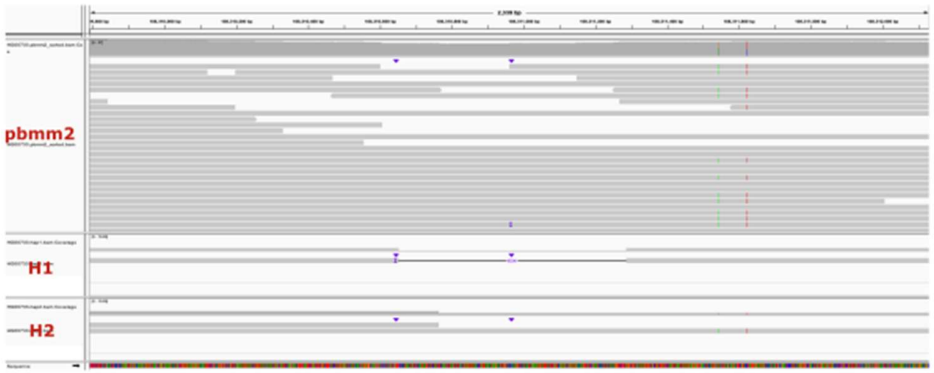


Fig. S79. The overall QC results of HiFi and CLR samples

(A-B) QC labels of HiFi samples and their repeat annotation; (C-D) QC labels of CLR samples and their repeat annotation; (E) Comparison of HiFi samples and their replicated CLR samples.

A. Example of NoRawReads INS

B. Example of NoRealignSupport DEL



C. Example of Imprecise DEL

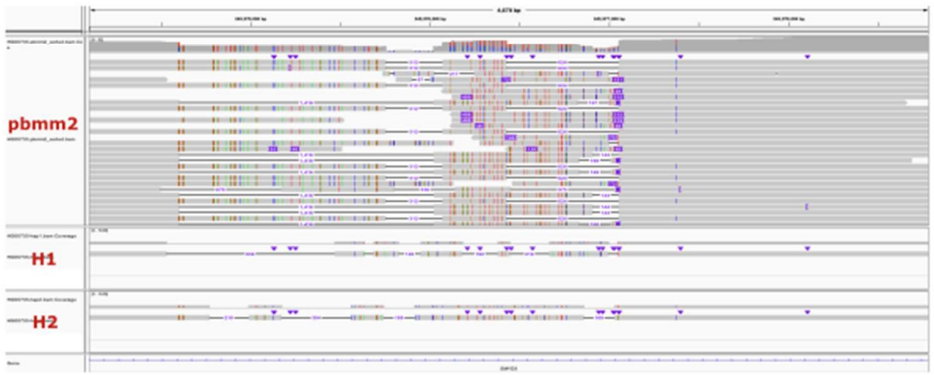


Fig. S80. Examples of QC labels

(A) Example of “NoRawReads” INS call. Note that different aligners may have different results on the label; (B-C) Examples of “NoRealignSupport” and “Imprecise” calls from sample HG00733.

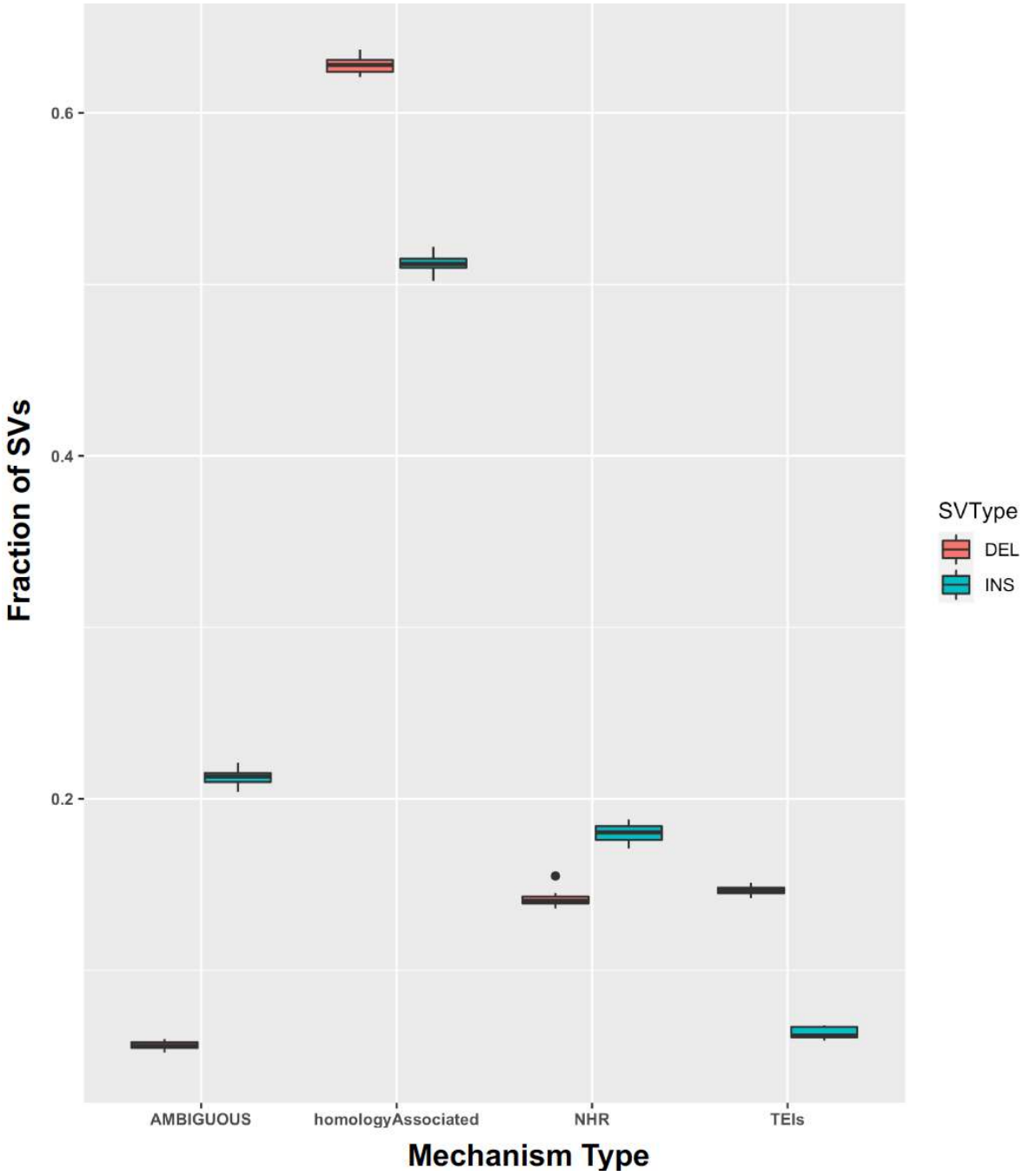


Fig. S81. Overall distribution of SVs belonging to different mechanism categories

Each data point in the box plot corresponds to the fraction of total SVs in a sample, which belong to a particular mechanism category. We performed this analysis separately for deletions (red box plot) and insertions (blue box plot).

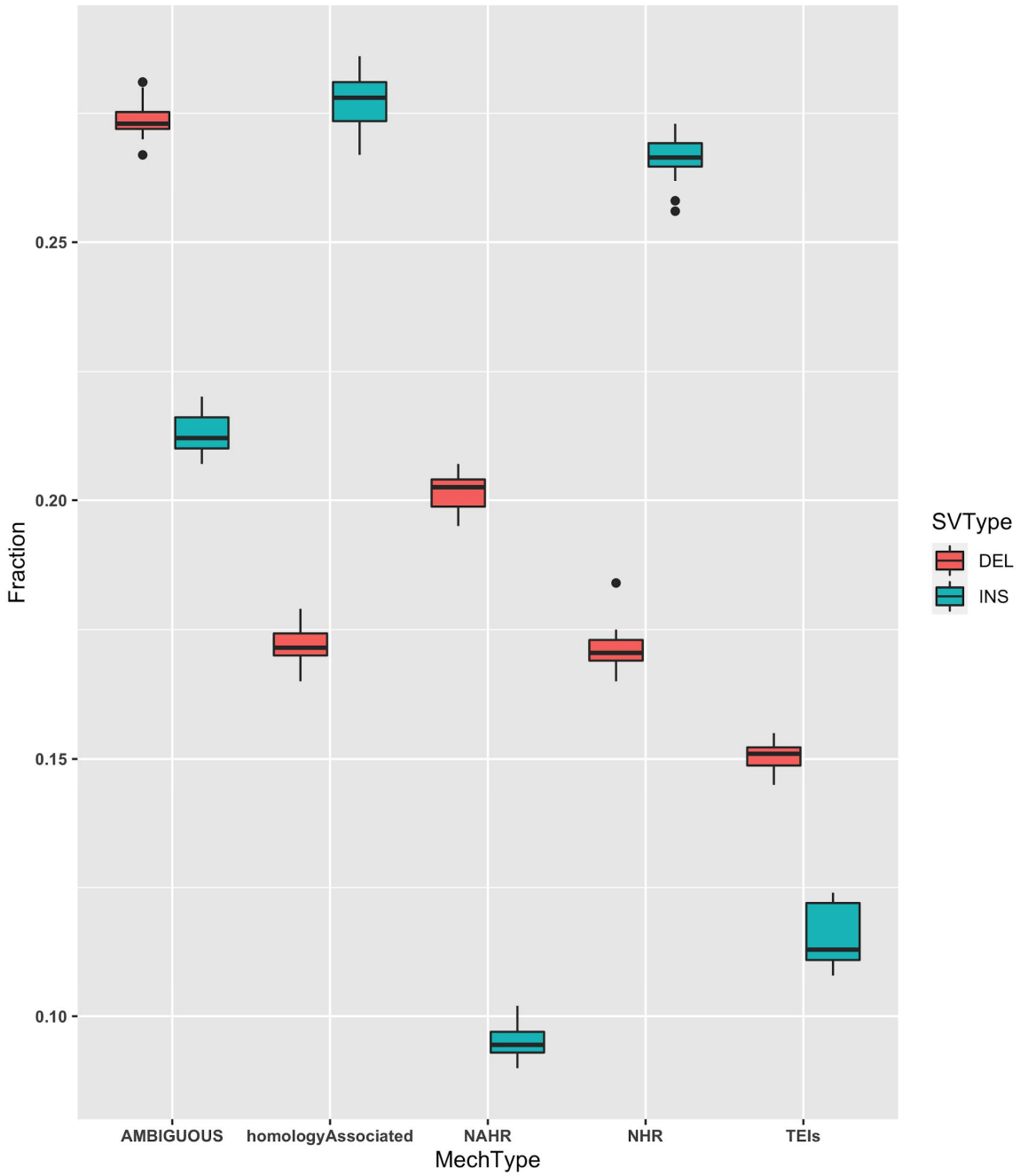


Fig. S82. Distribution of insertions and deletions belonging to different mechanism categories (based on homology length ≥200 bp)

Each data point in the box plot corresponds to the fraction of total SVs in a sample, which belong to a particular mechanism category. We performed this analysis separately for deletions (red box plot) and insertions (blue box plot).

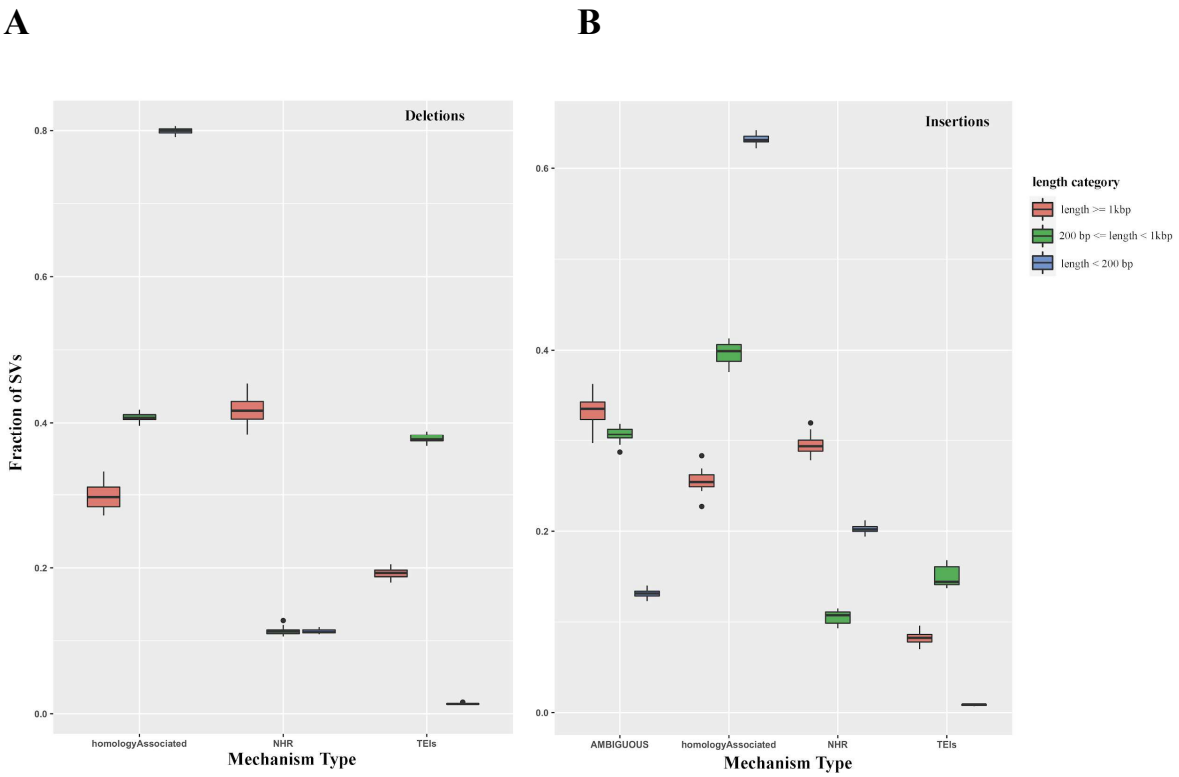


Fig. S83. Distribution of insertions and deletions of different length groups belonging to different mechanism categories

Each data point in the box plot corresponds to the fraction of total SVs in a sample, which belong to a particular mechanism category. We performed this analysis separately for deletions (A) and insertions (B). For each SV type, we divided SVs into three distinct classes based on their length: i) length <200 bp (blue), ii) ≥200 bp and <1000 bp (red), and iii) ≤1000 bp (green).

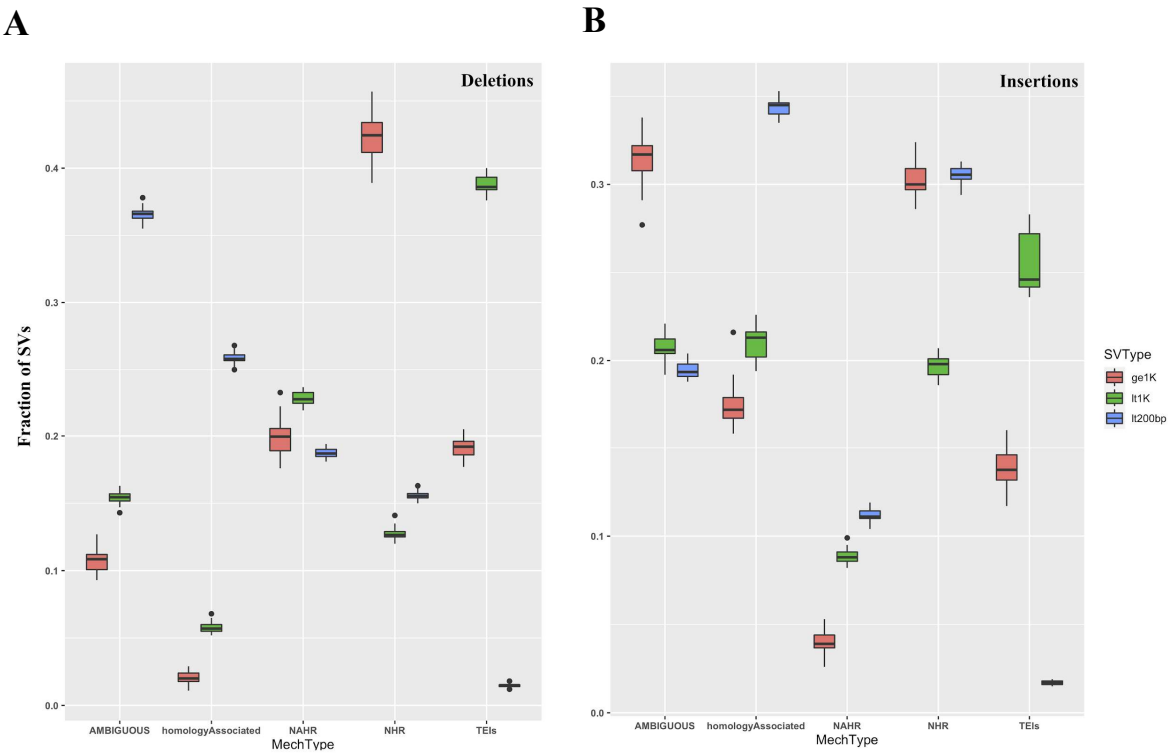


Fig. S84. Distribution of insertions and deletions of different length groups belonging to different mechanism categories (based on homology length ≥200 bp)

Each data point in the box plot corresponds to the fraction of total SVs in a sample, which belong to a particular mechanism category. We performed this analysis separately for deletions (A) and insertions (B). For each SV type, we divided SVs into three classes based on their length: i) length <200 bp (blue), ii) ≥200 bp and <1000 bp (red), and iii) ≤1000 bp (green).

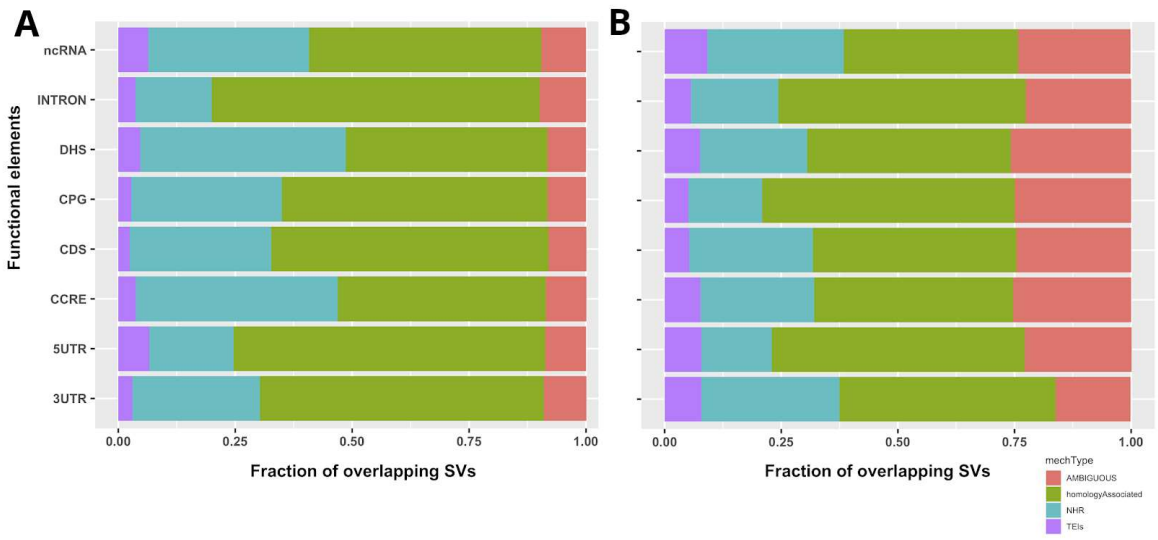


Fig. S85. Distribution of deletions and insertions overlapping with distinct functional elements belonging to different mechanism categories. Fraction of SVs (x-axis, A: deletions, B: insertions) overlapping various functional elements (y-axis) color-coded by their mechanism category. Values indicate the mechanism contribution of SVs overlapping with a given functional element. TEI: transposable element insertions; NHR: nonhomologous recombination; homology-associated: VNTR [rep. length ≥ 50 bp] plus NAHR; ambiguous: otherwise, see Section 17.6 (*18*); 3UTR: 3' UTR; 5UTR: 5' UTR; CCRE: candidate cis-regulatory element; CDS: coding sequence; CPG: CpG island; INTRON: intron; ncRNA: noncoding RNA

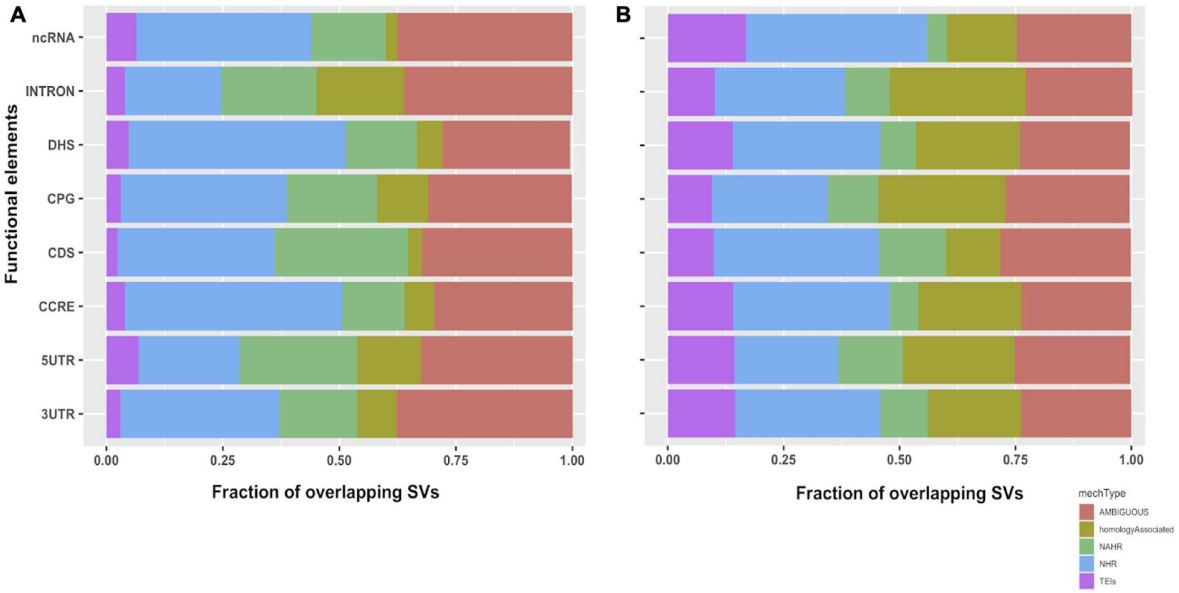


Fig. S86. Distribution of deletions and insertions overlapping with distinct functional elements belonging to different mechanism categories (based on homology length ≥200 bp). Fraction of SVs (x-axis, A: deletions, B: insertions) overlapping various functional elements (y-axis) color-coded by their mechanism category. Values indicate the mechanism contribution of SVs overlapping with a given functional element. TEI: transposable element insertions; NHR: nonhomologous recombination; homology-associated: VNTR [rep. length ≥ 50 bp and] plus NAHR; ambiguous: otherwise, see Section 17.6 (*18*); 3UTR: 3' UTR; 5UTR: 5' UTR; CCRE: candidate cis-regulatory element; CDS: coding sequence; CPG: CpG island; INTRON: intron; ncRNA: noncoding RNA

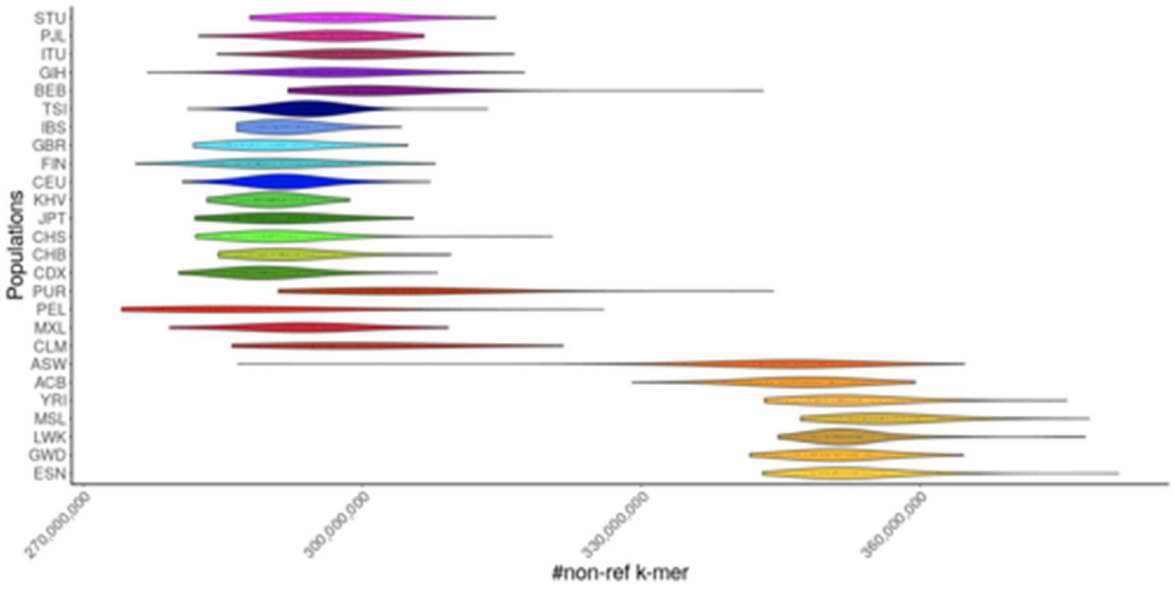


Fig. S87. Non-reference k-mer density distribution by population for 2,504 unrelated samples

Count of non-reference k-mers (k = 61, x-axis) aggregated per population (y-axis, see Fig. 1 for population color legend) for 2,504 unrelated high-coverage Illumina samples.

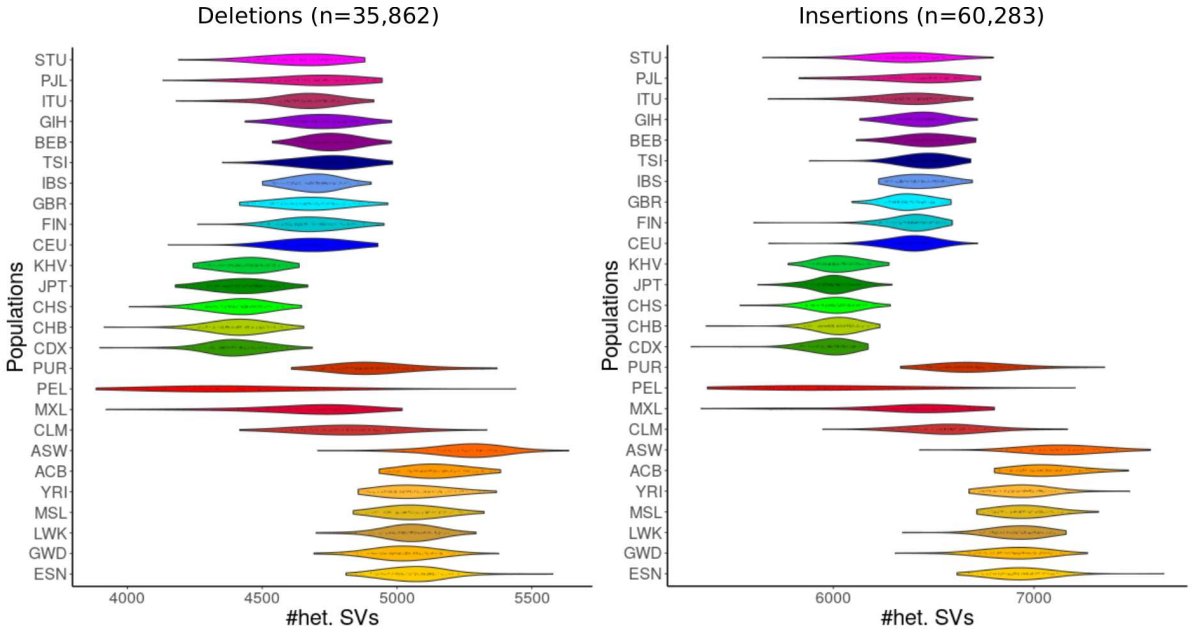


Fig. S88. Number of heterozygous SVs per population

Shown are the number of SVs typed as heterozygous by PanGenie for the different populations. The plots are based on the unfiltered callset containing all 96,145 SVs (35,862 deletions and 60,283 insertions).

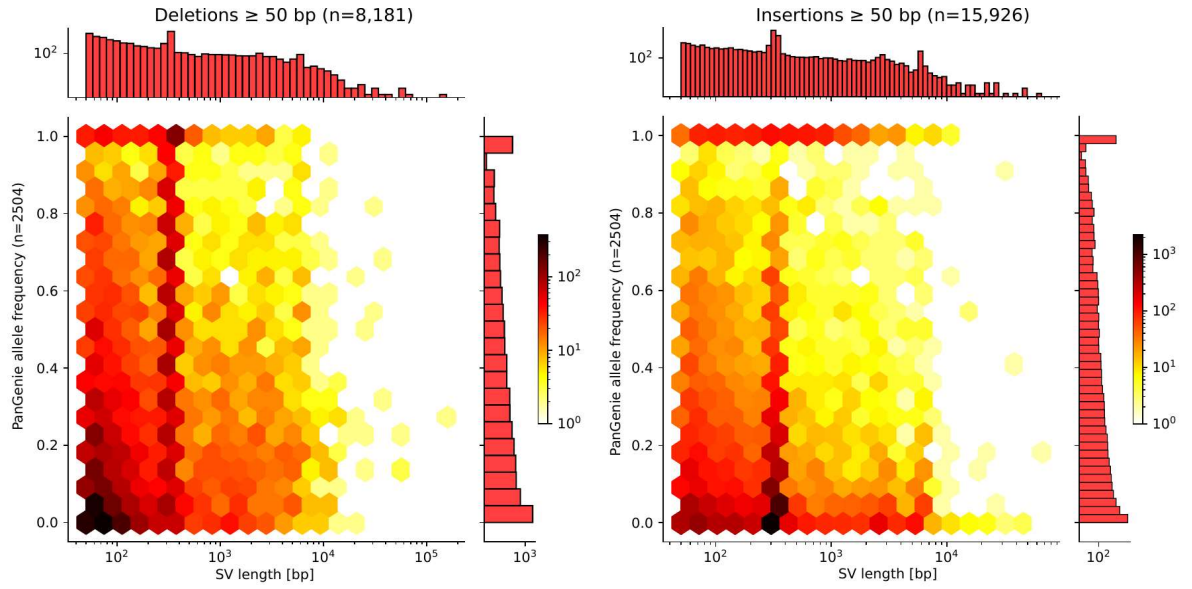


Fig. S89. PanGenie allele frequency versus SV length

PanGenie allele frequencies were computed based on the genotypes of all 2,504 unrelated samples. Only SVs contained in the filtered set (n=24,107) are considered.

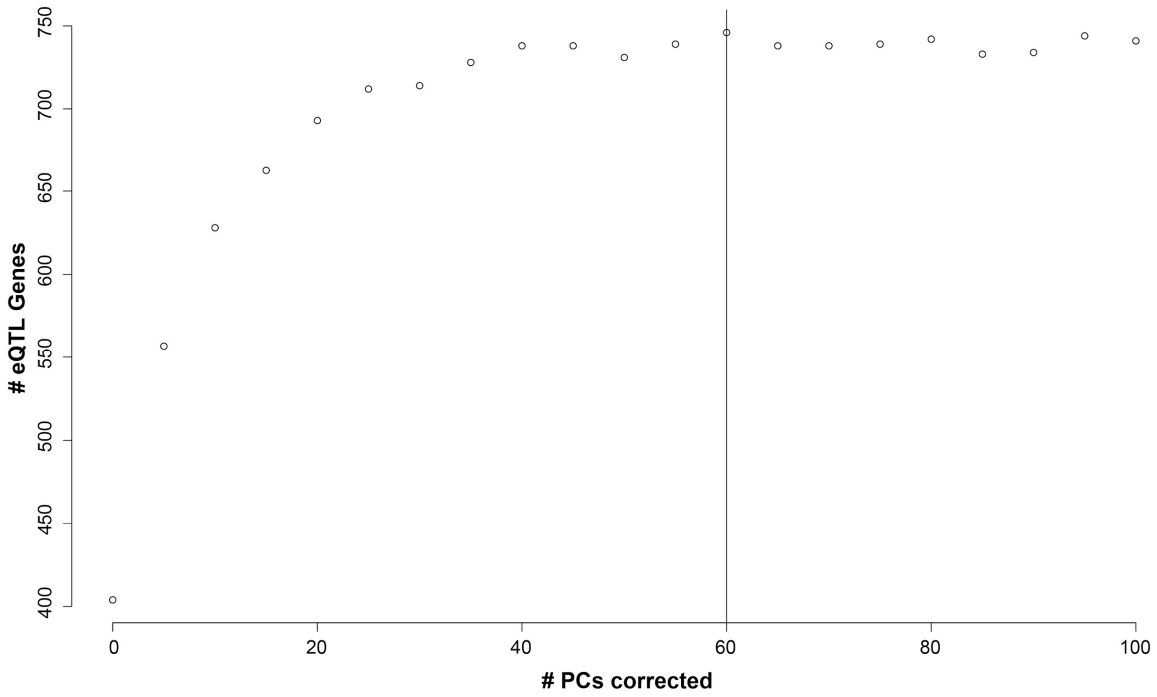


Fig. S90. Optimization of the number of principal components (PCs) to correct for when mapping eQTLs

The relation between the fraction of identified eGenes, the genes whose expression levels are associated with variation at a particular genetic variant, on chromosome 2 versus the number of PCs used to correct the expression data. The vertical bar (60) indicates the number of PCs used as fixed effect covariates in the eQTL mapping.

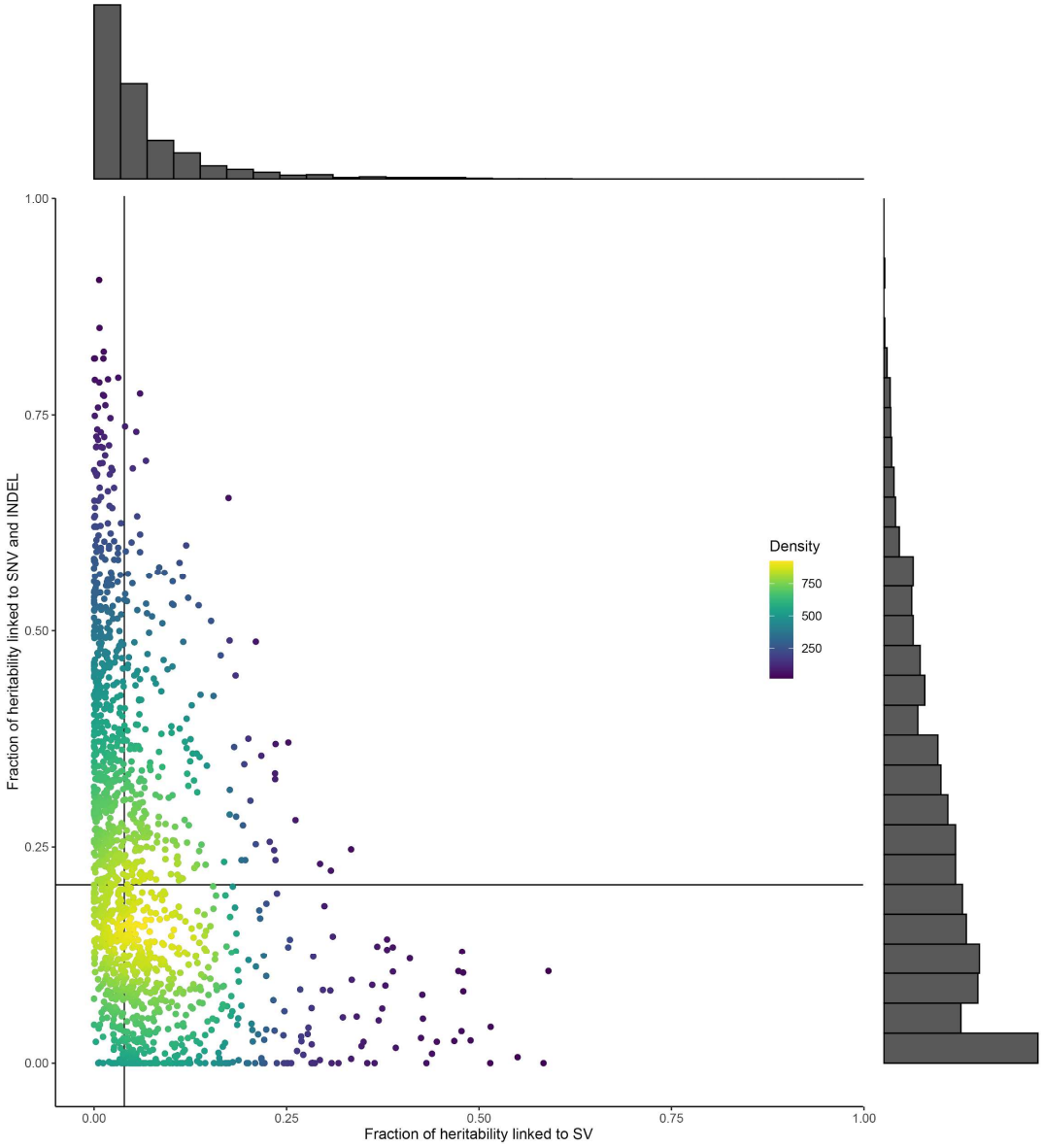


Fig. S91. Heritability of gene expression levels

Scatter plot showing the heritability of gene expression of each eQTL gene with at least one significant SV-linked eQTL. The SV heritability corresponds to the top SV per gene (x-axis) and SNV and indel heritability corresponds to the joint heritability linked to the top 1,000 SNV and indel variants per gene (y-axis). The horizontal line corresponds to the median SNV and indel heritability (20.4%); the vertical line corresponds to the average SV heritability (3.9%). The bar plots at the side depict the frequency of the eGenes, SVs along the x-axis and SNVs and indels along the y-axis. Plots are modeled after the GTEx et al. SV eQTL paper (*45*).

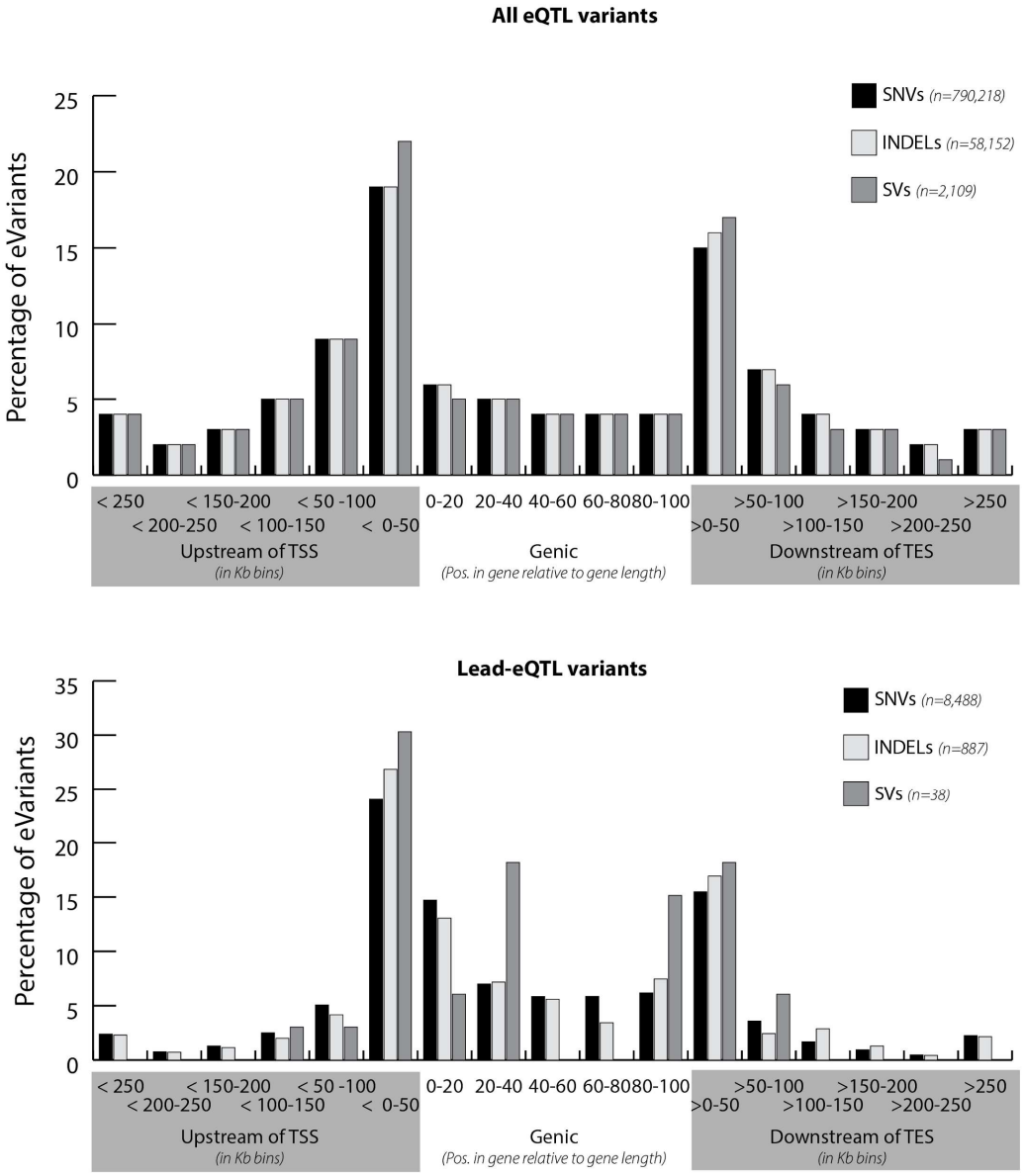


Fig. S92. Breakdown of the eQTL signals based on the distance between the gene and the eQTL variant

Shown are the proportion of eVariants (eQTL SNP/INDEL/SV) that are linked to expression variation given the distance to the gene, top all eQTL variants, bottom lead-eQTL variants. We observe high concordance between the distance of an eVariant and eGene eGene between the three classes, for SV-eQTL the eVariant is slightly more centered around the gene (i.e., closer to the gene body).

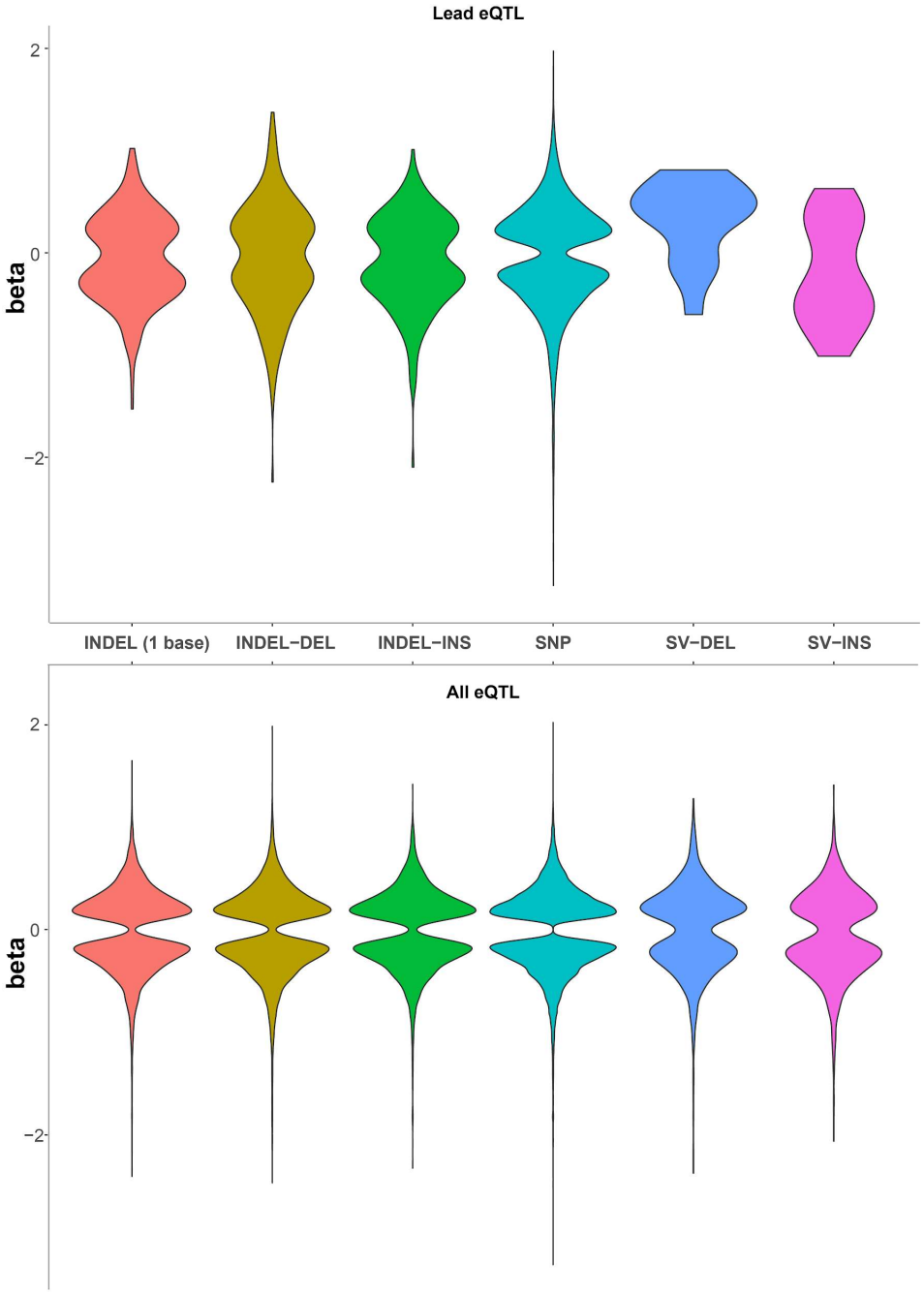


Fig. S93. Effect sizes of eQTL per eVariant class

Violin plots depicting the effects sizes and effect directions for the eQTL identified, (top lead-eQTL, bottom all eQTL). Effects are split out based on their eQTL variant (eVariant) class.

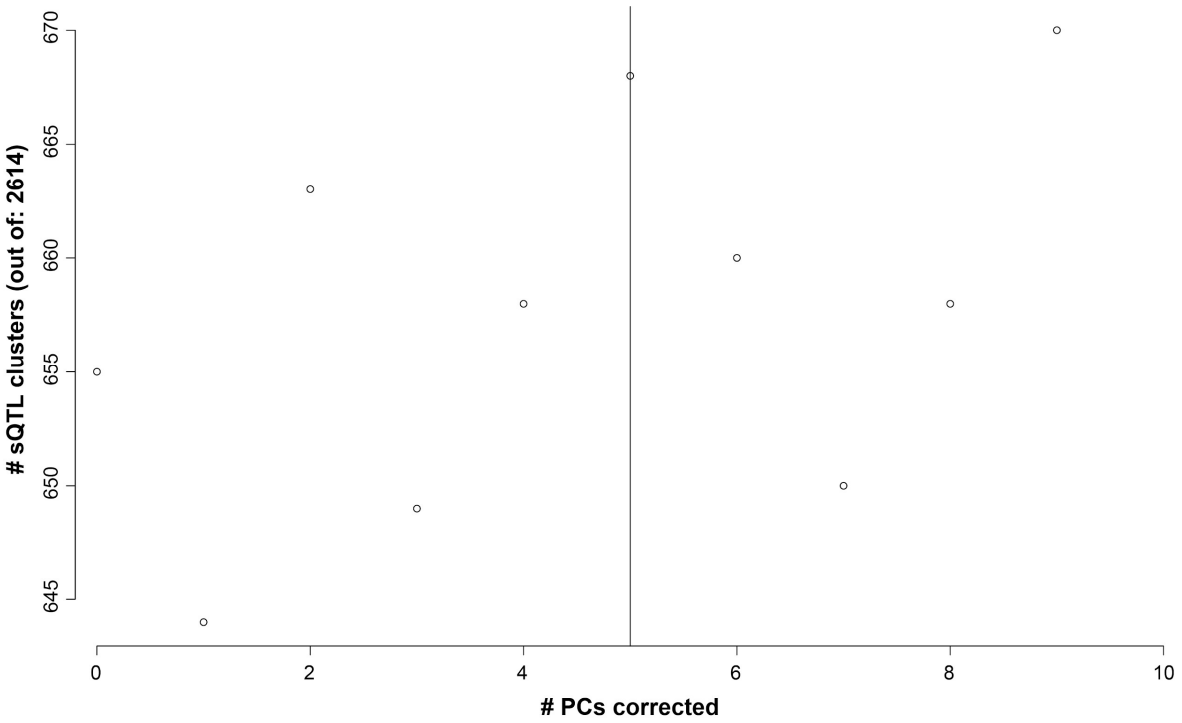


Fig. S94. Optimization of the number of principal components (PCs) to correct for when mapping sQTLs

The relation between the number of identified splice-junctions with a genetic effect (sQTL),on chromosome 2 versus the number of PCs used to correct the splicing-ratios. The vertical bar (5) indicates the number of PCs used as fixed effect covariates in the sQTL mapping.

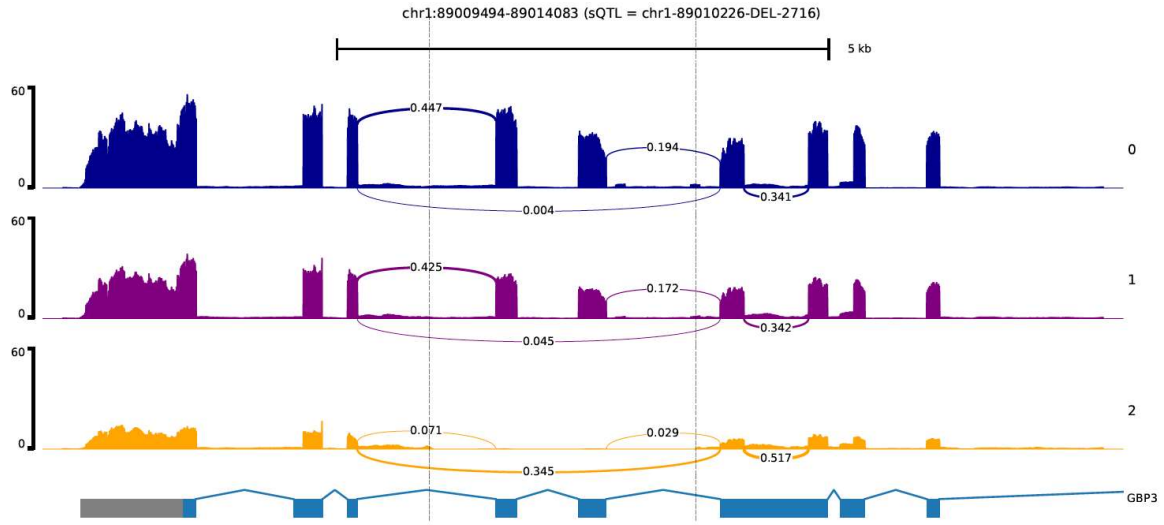


Fig. S95. SV linked sQTL on *GBP3*

Sashimi plot showing LCL RNA-seq read coverage and junction reads at the gene *GBP3*, separated by genotype for chr-89919226-DEL-2716. Transcripts from individuals homozygous for the deletion show full skipping of the middle protein-coding exons. chr-89919226-DEL-2716 is also an eQTL for *GBP3*.

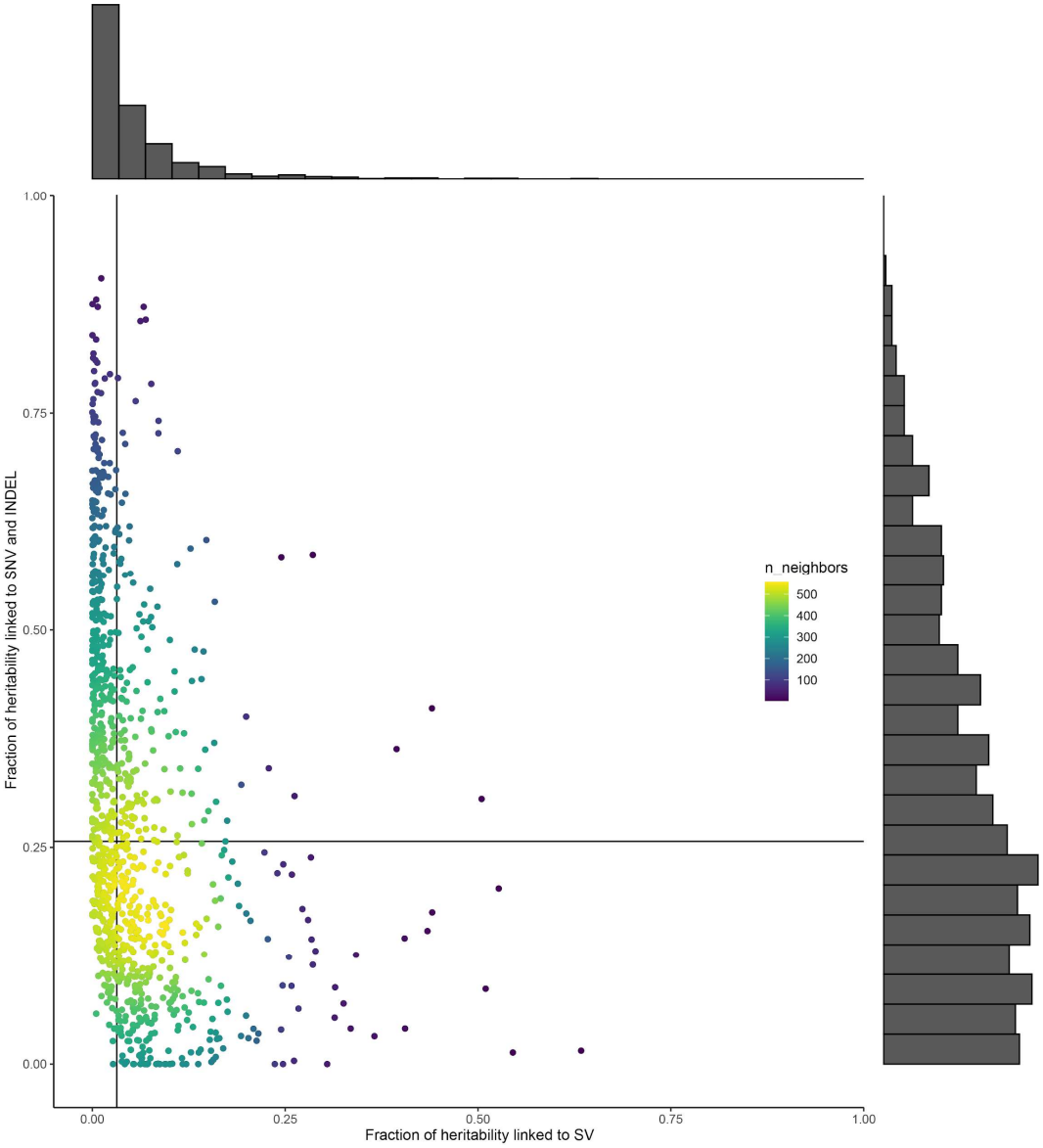


Fig. S96. Heritability of RNA splicing

Scatter plot showing the heritability of RNA splicing of each sQTL junction with at least one significant SV-linked sQTL. The SV heritability corresponds to the top SV per junction (x-axis) and SNV and indel heritability corresponds to the joint heritability linked to the top 1,000 SNV and indel variants per junction (y-axis). The horizontal line corresponds to the median SNV and indel heritability (25.6%); the vertical line corresponds to the average SV heritability (3.1%). The bar plots at the side depict the frequency of the sQTL effects, SVs along the x-axis and SNVs and indels along the y-axis. Plots are modeled after the GTEx et al. SV eQTL paper (*45*).

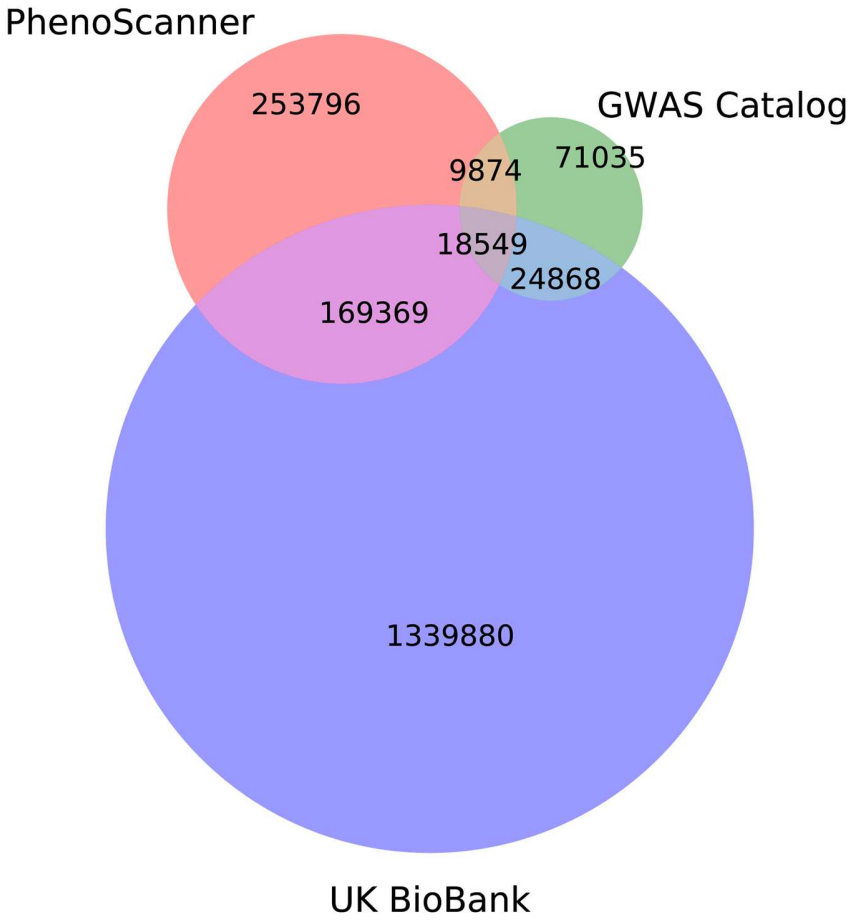


Fig. S97. Venn diagram of known GWAS associations extracted from the GWAS catalog, UK Biobank and PhenoScanner

GWAS associations reported with p-value ≤ 1e-6 are kept in this study. SNP positions in UK Biobank and PhenoScanner are lifted over to GRCh38.

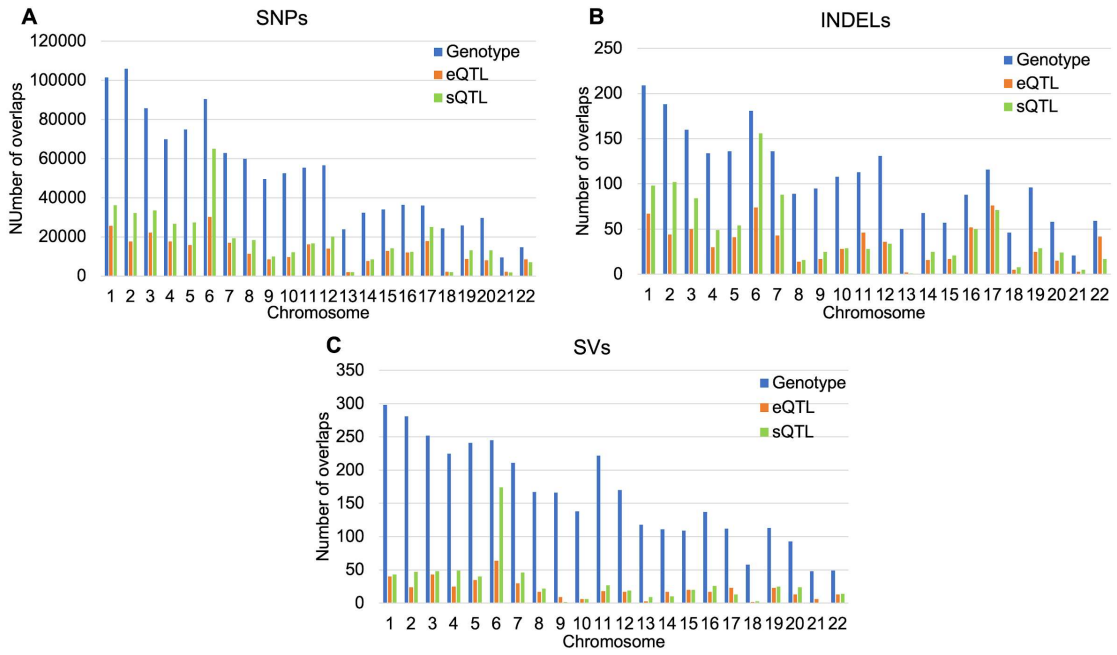


Fig. S98. Numbers of genotyped variants, eQTLs, and sQTLs of various types that overlap with known GWAS SNPs

(A) SNVs with positions matching GWAS SNPs; (B) Indels that are at least 1 bp overlap with GWAS SNPs; (C) SVs that are at least 1 bp overlap with GWAS SNPs. X-axis represents chromosomes and y-axis represents the number of overlaps.

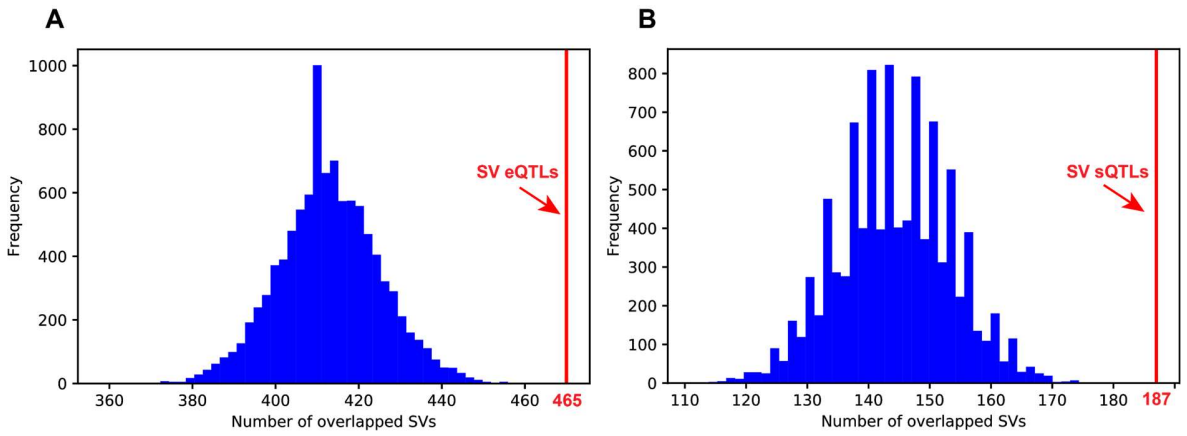


Fig. S99. Enrichment of SV eQTLs and sQTLs for overlapping with GWAS SNPs

(A) eQTL and (B) sQTL. X-axis represents the number of SVs overlapped with GWAS SNPs, and y-axis represents the frequency. The vertical line in red denotes the number of SV eQTLs or sQTLs overlapped with GWAS SNPs. The enrichment was conducted by comparing the observed overlap of SV eQTLs/sQTLs with those random SV sets in 10,000 random permutations. In each permutation, SV regions were randomly selected from genotyped SVs that are input to the eQTL/sQTL analysis pipeline, with the same distribution as SV eQTLs/sQTLs in terms of chromosome source, SV length, variant type, distance to transcription start site, and distance to transcription end site of genes.

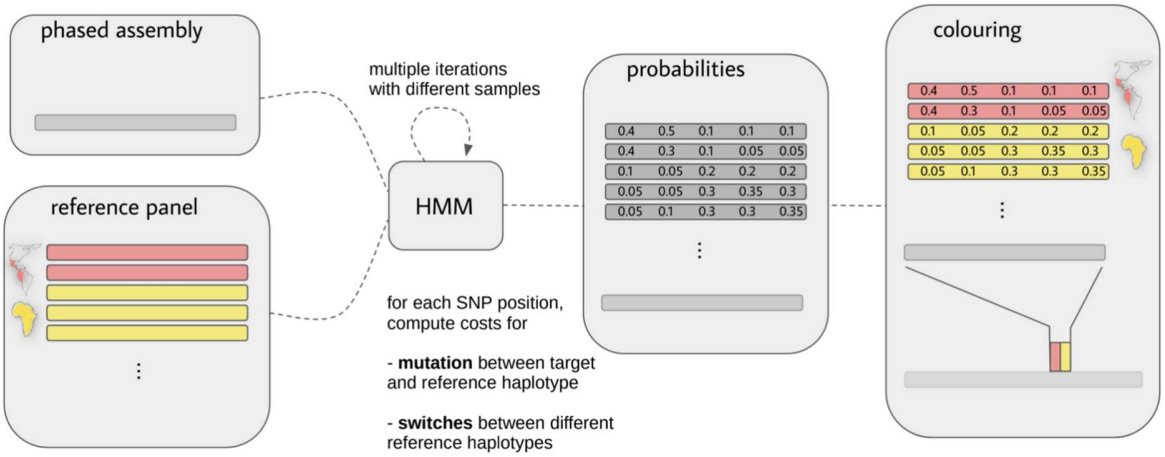


Fig. S100. Hidden Markov model (HMM) for local ancestry inference using haplotype-resolved assemblies

Schematic overview of the HMM-based method. In the first step, the assemblies as well as a reference panel from the 1000GP samples are input to the HMM. Then, the model computes a probability for each variant position and each reference haplotype by applying a forward-backward algorithm. Multiple iterations of the HMM are run using different randomly sampled sets of reference haplotypes from the panel. Last, for each variant, the probabilities of reference samples belonging to the same superpopulation are summed up to compute the likelihoods for every superpopulation. The example shows five variants and five reference haplotypes, of which two belong to the AMR population (colored in red) and three to the AFR population (colored in yellow). For the first two variants, the cumulative probability is highest for the AMR population, resulting in a red block in the output; for the last three positions, the probability for AFR is higher, which results in a yellow block.

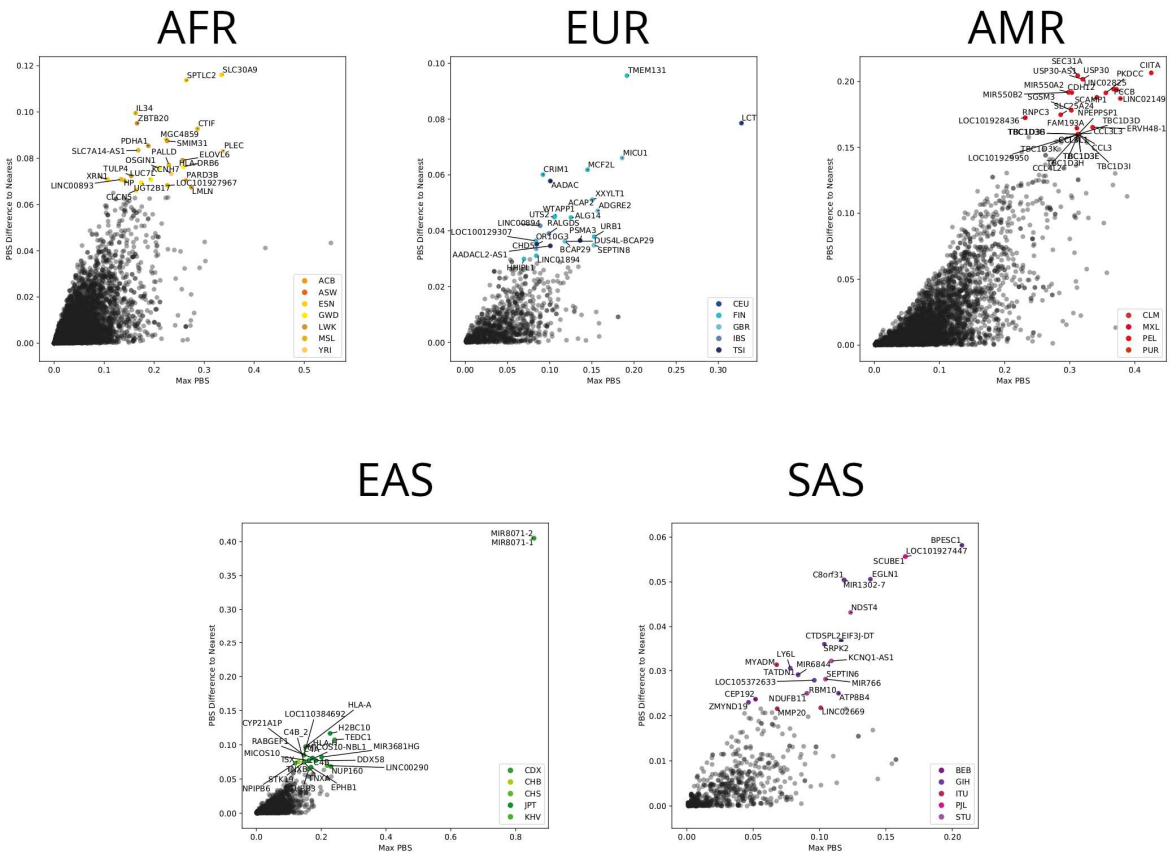


Fig. S101. Top population branch statistics (PBS) hits per superpopulation

For each gene, the maximum PBS value (horizontal axis) and the distance to the next highest PBS in another population (vertical axis) was determined. Each panel shows the PBS distribution for top hits in one superpopulation.

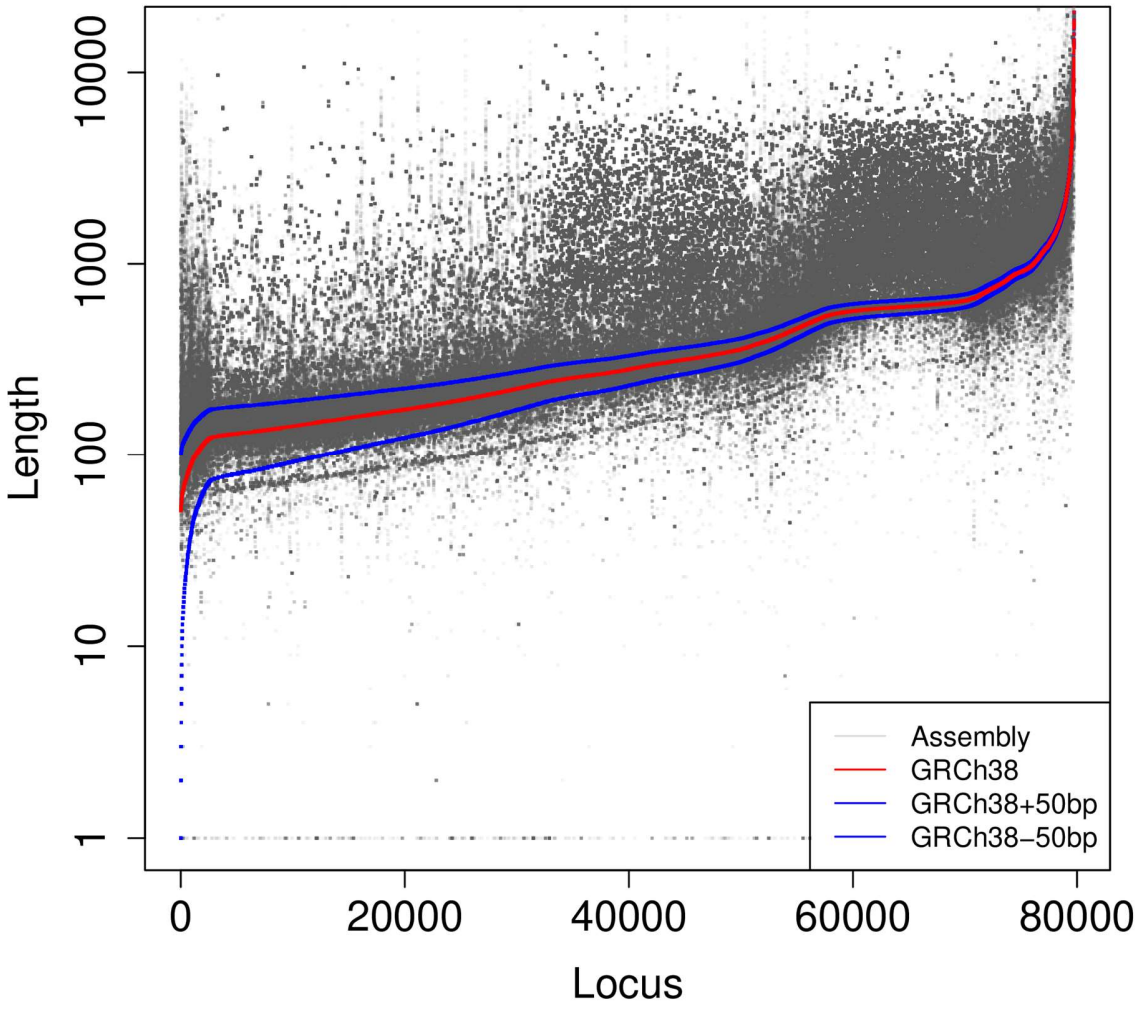


Fig. S102. Length of VNTR sequences at orthologous loci

The lengths of 79,762 VNTR loci in GRCh38 (red) and orthologous assembled haplotypes (gray). Each VNTR locus is a column, plotted in order of increasing VNTR length in GRCh38. The top and bottom blue lines represent the reference length +50 bases and -50 bases such that points above and below the line correspond to VNTR alleles that are an SV compared to the reference. There is a visual emphasis on variant alleles due to the resolution of rendering the image.

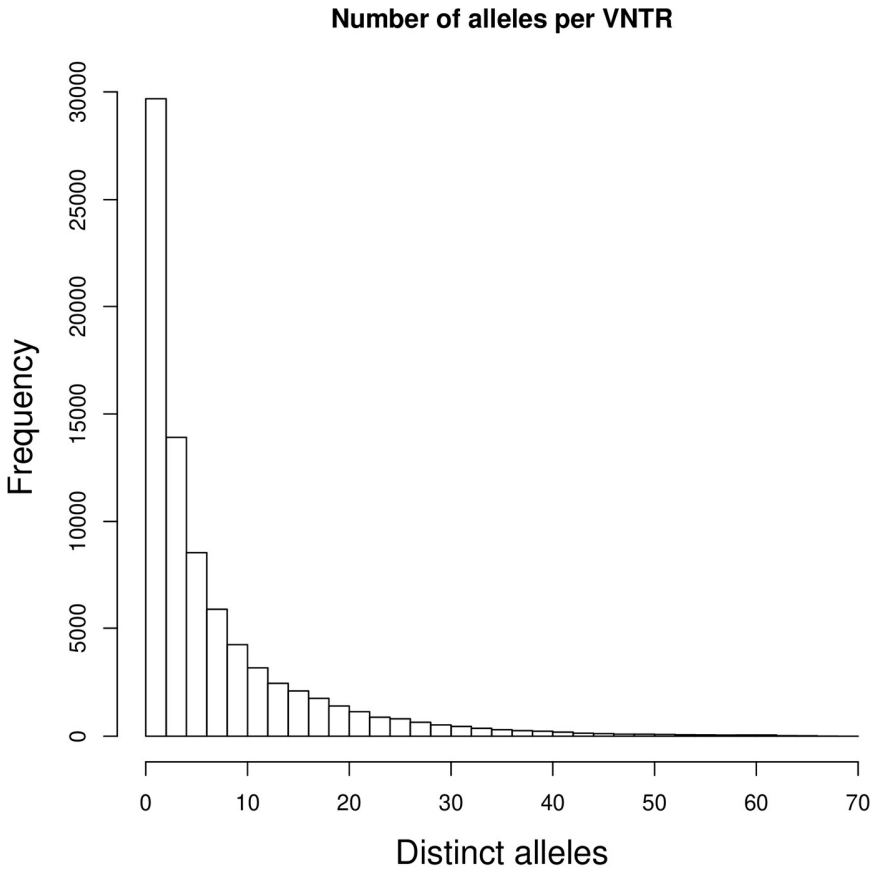


Fig. S103. The distribution of the number of distinct VNTR allele lengths

The number of distinct VNTR allele lengths is shown for 70 genomes, including the parental genomes in assembled trios. On average there are 7.3 alleles per VNTR.

List of tables available as Supplementary Material online

Table S1. Sequencing and assembly summary per sample
Summary for input long-read and Strand-seq datasets per sample used to generate phased assemblies for each sample. Length and contiguity statistics are given for the non-haplotype-resolved (“squashed”) and the haploid assemblies per sample.

Table S2. Sample summary statistics
Overview of basic attributes for samples analyzed in this study such as population of origin and sex.

Table S3. Switch error rate for phased human genomes
Phasing accuracy summary statistics (switch error rate and Hamming distance) for the children (AMR/HG00733, CHS/HG00514, YRI/NA19240) of the three family-trios part of this study.

Table S4. Variant discovery from phased assemblies
Merged callset contribution and support for each sample after QC. Calls are discovered by PAV in each sample with inversions added from Strand-seq and Bionano.

Table S5. Haploid assembly contig coverage and completeness
Haploid-assembly contig coverage for high-quality (MAPQ60) alignments relative to the GRCh38 reference (main chromosomes and ALT contigs). Assembly completeness statistics extracted from QUAST/BUSCO analysis runs.

Table S6. Sequencing QC and statistics
Long-read sequencing statistics and QC summary for all samples (HiFi and CLR) used in this study. Sequencing center “PacBio” for sequencing data added from public databases (see Table S38).

Table S7. Enrichment of assembly breaks
Enrichment summary statistics for common assembly breaks (≥10 kbp) identified at three different MAPQ thresholds for various functional annotation categories.

Table S8. Bionano Genomics hybrid scaffolding
Summary statistics for all Bionano-scaffolded assemblies (“hybrids”) indicating fraction of haploid assembly with Bionano support and number and type of misassemblies detected as part of the hybrid scaffolding process.

Table S9. Variants previously found in Audano2019
Shared variants were compared with Audano2019 (*4*) for support from any variant (Intersect) or previously reported shared variants (Shared). Intersects in coding (CDS), UTRs, promoters, enhancer-like signatures (ELS), or any of these functional categories was counted. The number of PanGenie annotations with AF=1 and that were not callable by PanGenie are also counted.

Table S10. Estimated Mendelian error rate by trios
The Mendelian error rate was estimated by merging trios only and applying the same filter +1 and +2 filters that were applied to the main callset. To estimate FDR, we computed an average weighted by the number of HiFi-supported and CLR-supported calls ("Combined"). The Han Chinese trio (CHS) is a clear outlier, which may inflate the FDR projections.

Table S11. Singletons per haplotype
The number of new singletons per haplotype and new homozygous variants per sample were found by permuting callset order and evaluating the effect of adding the final haplotype and final sample. In both cases, African samples grow the callset more than non-African samples.

Table S12. Long-read callset novelty by SV size.
Number of insertions and deletions in our merged PAV callset for various size ranges plus the fraction of the novel observations in the respective category.

Table S13. Count of SVs per sample from short-read and long-read sequences and their overlap
Summary statistics for variants (all/ALL, insertions/INS, deletions/DEL) detected from long-read/phased-assembly and short-read analysis pipelines. Counts are given for shared and technology-specific variants, and specifically for small variants (<250 bp) and small variants located in repetitive regions defined by the segmental duplication (SD) and simple repeats (SR) tracks in UCSC genome browser.

Table S14. SV hotspots detected in this study
Detected SV hotspots are listed with their genomic coordinates, number of events per hotspot, and overlapping genes (GENCODE v36 annotation). The note indicates if the hotspot is novel or has already been reported in previous studies (Sudmant2015: (*23*)).

Table S15. Haploid assembly summary statistics for MHC region
Coverage summary statistics for all haploid assemblies in the MHC region. Gaps are indicated as absolute counts and summed up as total base pairs of sequence missing (Delta, bp). Assemblies are ordered by the amount of missing sequence (Delta, pct. rank).

Table S16. Sequence-resolved HLA genes in haploid assemblies
Count of fully sequence-resolved HLA genes (out of 19 in total, GENCODE v31 annotation) per haploid assembly. HLA genes are counted as fully resolved if the locus is fully covered by MAPQ60 contig alignments, unresolved if the locus is uncovered, and partially resolved otherwise. Note: 35 individuals equals 70 haploid assemblies, plus 18 for three family trios (9 individuals) that have been assembled with both HiFi and CLR long reads.

Table S17. Frequency of SVs belonging to different mechanism class
A) This table presents frequency of SVs based on BreakSeq run with homology length cutoff >=50 bp. Percentage in brackets correspond to (= # of SVs in a mechanism category/ Total# of SVs in non-MEI SVs). MEI-associated category includes SVs overlapping with MEIs released by MEI group as well as STEI/MTEI identified by BreakSeq using RepeatMasker, B) This table presents frequency of SVs based on BreakSeq run with homology length cutoff >=200 bp. Percentage in brackets correspond to (= # of SVs in a mechanism category/ Total# of SVs in non-MEI SVs). MEI-associated category includes SVs overlapping with MEIs released by MEI group as well as STEI/MTEI identified by BreakSeq using RepeatMasker, C) This table presents the number of SVs with precise breakpoints that overlap with Illumina-based SV callsets while using homology length cutoff >=50 bp. Percentage in brackets correspond to (= # of SVs in a mechanism category/ Total# of SVs), and D) This table presents the number of SVs with precise breakpoints that overlap with Illumina-based SV callsets while using homology length cutoff >=200 bp. Percentage in brackets correspond to (= # of SVs in a mechanism category/ Total# of SVs).

Table S18. Variable coding region lengths identified by SV clustering
SVs intersecting coding sequence were clustered to identify exons with variable repeat arrays. N SV: Number of SVs intersecting the gene; INS %: Percent of SVs that are insertions (vs deletions); Frameshift %: Percent SVs predicting a frameshift in the coding sequence (i.e., variant length is not a multiple of 3); Min diff. and Max diff.: Minimum and maximum SV lengths with deletions as negative numbers; Max SD: Highest SD identity for genes within SDs or 0 or genes not in an annotated SD (SD annotations from UCSC).

Table S19. Coding and cis-regulatory SVs
SVs intersecting functional elements (“Functional element”) are summarized by the number that were not observed in chimpanzee or gorilla ("Non-ancestral") or are population stratified with Fst > 0.20 in a single superpopulation (“Superpop stratified”). Strikingly, almost all regulatory hits are of African (74.7%) and East Asian (19.3%) origin. CDS: Coding sequence, PLS: Promoter-like signature, pELS: proximal enhancer-like signature, dELS: distal enhancer-like signature. PLS, pELS, dELS, and CTCF-only are defined by ENCODE.

Table S20. Triplet repeat expansions condensed to nonredundant regions
Individual triplet repeat expansions grouped into nonredundant loci. The number of samples (N) and copy number changes relative to GRCh38 (min, max, mean) are given for each locus. Gene annotations were done by RefSeq intersect.

Table S21. Triplet repeat expansions identified in the SV calls for each sample
Triplet repeat expansion with perfect representations of the triplet repeats covering at least 95% of the SV call. Sample, ID, genotype, and SV length are derived from individual sample callsets (pre merging). RepeatMasker motif and percent identity are given for each record. Perfect match stats show the number of perfect uninterrupted repeats per record (N: Number of repeat elements, bp: Total bp change). Gene annotations were done by RefSeq intersect.

Table S22. Collection of sequence-resolved full-length L1s (FL-L1s)
This table contains information regarding all sequence-resolved FL-L1s identified at HGSVC2 dataset. This includes FL-L1s coordinates (A-C); unique cytoband identifier for active FL-L1s (D); subfamily assignation (E); boolean specifying if located on the reference or not (F); strand orientation (G); ORF content annotation (H); boolean specifying if known to be active or not (I); Hot activity status according to a Pan-cancer study (J) and in-vitro (K); contribution expressed as the % of all transductions identified that are mediated by the element in the population (L-M) and cancer genomes (N-O); number of transductions mediated by each FL-L1s in four different studies (P-S); in-vitro retrotransposition activity measured as activity % relative to L1RP (T-U); age estimations in million years based on sequence divergence for active FL-L1s (V); allele frequency expressed in percentage on the HGSVC2 dataset (W). "NA" stands for Not Available

Table S23 Compendium of FL-L1s known to be active
This table contains information regarding a set of active FL-L1s based on information collected across multiple studies. This includes FL-L1s coordinates (A-C); unique cytoband identifier (D); boolean specifying if located in the reference or not (E); strand orientation (F); boolean specifying if the FL-L1 has been sequence-resolved or not (G); Hot activity status according to a Pan-cancer study (H) and in-vitro (I); contribution expressed as the % of all transductions identified that are mediated by the element in cancer genome (J-K) and the population genetic studies (L-M); in-vitro retrotransposition activity measured as activity % relative to L1RP (N-O); number of transductions mediated by each FL-L1s in four different studies (P-S). "NA" stands for Not Available

Table S24. SVA-mediated transductions
This table contains information regarding all SVA-mediated transductions identified in the HGSVC2 dataset. This includes the transduced sequence source coordinates (A-C); insertion position coordinates (D-F); transduction length (G); boolean specifying if corresponds to a 5' (H) or 3' (I) transduction; source element cytoband identifier (J); source element coordinates (K)

Table S25. Catalogue of active SVA source loci
This table contains information regarding all the SVA source loci identified in the HGSVC2 dataset. This includes their genomic position (A-C); cytoband identifier (D); SVA subfamily (E); boolean specifying if is in the reference or not (F); strand orientation (G); boolean specifying if the source element reference or non-reference insertion has been identified or not (H); contribution expressed as the % of all transductions identified that are mediated by the element (I); total number (J), 5' (K) and 3' transductions (L) identified; transduction insertion coordinates (M). "NA" stands for Not Available

Table S26. High-confidence SVs (≥5 kbp) detected by Bionano Genomics
This table includes large structural variants (≥5 kbp) discovered by optical mapping in 30 samples. These SVs localize at clean regions where the respective sample has no more than two optical maps aligned at SV breakpoints. Among samples, calls with 80% size and position concordance are clustered. Information of the SVs include the Cluster ID (A); the sample (B); the call ID (C); the type of SV (C); genomic position (E-G); the zygosity (H); the SV size (I); the gene overlapped (J); the number of molecules supporting the call (K); the number of optical maps aligned at the SV breakpoints (L); PAV overlap information (M); the PAV dataset overlapping the call (N); population information of the sample (O-Q); gender of the sample (R). PAV overlap information only applies to insertion and deletions.

Table S27. Bionano SV clusters with calls missed by PAV
This table includes SV clusters that have calls missed in PAV in at least 1 of the 30 samples. Polymorphic variants can contribute to clusters of different sizes with similar genomic positions. Information of the clusters include the cluster ID (A); the SV type (B); genomic position (C-E); the mean size of SVs contributing to the cluster (F); number of samples contributing to the cluster (G); total number of insertions and deletion clusters in the region of the cluster (H-L); length overlapped with UCSC hg38 SuperDups (N); gene overlapped (O); sample population contributing to the cluster (P-T).

Table S28. Bionano SVs fully unresolved in the phased assemblies (N = 1,175)
This table contains the evaluation results of Bionano SVs that are missed by PacBio assembly with short read callset and sequencing depth. (I) describes the Bionano SVs fully unresolved in PAV; (II) describes additional Bionano SVs resolved but in a different sample.

Table S29. An overview of the number of variants obtained during construction of the variation graph
The first column provides the number of input variants (from PAV callsets). The second and third columns show the numbers of variants obtained after removing positions with at least 20% missing alleles in the panel and such located outside of chromosomes 1-22 or chromosome X. The last column provides the numbers of variants contained in the final graph used as input for PanGenie.

Table S30. Average number of SVs detected by long-read PacBio sequences per sample, split by concordance with Illumina calls and genomic location

This table displays the averaged number of deletions and insertions per genome in PacBio assembly callset, and their distribution across the genomic context defined based on the RepeatMasker and segmental duplication tracks from UCSC genome browser. For simplicity, we consolidated all repetitive sequence annotations into three categories: segmental duplication (SD; 5.1% of the genome), simple repeat (SR; 4.6%), and referred to all other repetitive sequence not overlapping SD/SR elements as 'repeat masked' (RM; 42.9%). The remaining 47.4% of the genome not overlapping any of these repeat categories was labeled as 'Unique' sequence.

Table S31. Top eQTL effects discovered in the joint LCL eQTL mapping.

Summary statistics of cis-eQTL mapping (lead only) in both the deep 1000GP RNA seq samples and GEUVADIS. Beta and beta_se obtained using a linear mixed model with P-values obtained from a likelihood ratio test, empirical p-values are derived from gene-level permutations and global P-values are corrected for the number of features tested using Storey's Q value.

Table S32. Top sQTL effects discovered in the joint LCL sQTL mapping

Summary statistics of cis-sQTL mapping (lead per splice junction only) in both the deep 1000GP RNA seq samples and GEUVADIS. Beta and beta_se obtained using a linear mixed model with P-values obtained from a likelihood ratio test, empirical P-values are derived from splice cluster level permutations and global P-values are corrected for the number of splice clusters tested tested using Storey's Q value.

Table S33. Break down of sQTL and eQTL enrichment for Indel and SV

Detailed analysis of different genetic variant types, SV, SV-deletion, SV-insertion, indel, indel-deletion, indel-insertion, and the likelihood of driving an QTL effect, both splicing and expression QTL. The enrichments and p-values are derived from Fisher exact tests, testing for an enrichment in the number of observed QTLs for a genetic variant type versus the expected number of QTLs for the variant type.

Table S34. Top Population Branch Statistics (PBS) values per gene

Population branch statistics (PBS) were computed from PanGenie genotypes with a 2 kbp window around genes (by RefSeq annotation) and include the total number of SVs in the gene window (N SV in gene). Top hits were manually inspected and assigned a Type (broad SV type), Class (further qualified type), and Gene effect (direct effect on the gene). Based on the SV and its genomic context (e.g., gene, regulatory element, functional annotations, SV confidence), we attempted to assign a functional effect and confidence to each record. We quantify the highest and second highest PBS hits for each record including the population and superpopulation for each. Finally, we included a prediction of the SV age inferred using 1 Mbp sequences centering at the SV of interest and Relate (*202*).

Table S35. Top Gene-SV outliers by PBS
Population branch statistics (PBS) were computed from PanGenie genotypes with a 2 kbp window around genes (by RefSeq annotation). The highest and second highest PBS values were included with a PBS z-score based on all PBS calculations. Variant discovery information (SV ID, SV Type, SV Length, and GC) are included with intersects from previous studies (Audano2019: (*4*), Chaisson2019: (*1*)), Illumina SVs from the same samples using high-coverage 1000GP data, and the PanGenie confidence level (PG Conf.; 4 is highest, 1 is lowest). Superpopulation fixation (Fst) and allele frequency (AF) statistics were computed from PanGenie genotypes.

Table S36. Sample list for each technology used in this study
Overview table indicating available data types per sample analyzed in this study.

Table S37. FTP links or accessions for key data sets
Summary table for HGSVC2 data access and repository accessions.

Table S38. NCBI accession numbers for samples sequenced by others
List of repository accessions for sequencing datasets that were not produced as part of the HGSVC2 effort.

Table S39. RNA-seq QC statistics
RNA-seq summary QC statistics for all datasets produced as part of the HGSVC2 effort.

Table S40. List of excluded Strand-seq libraries
List of Strand-seq libraries not used for clustering in the assembly pipeline due to quality concerns (“blacklist”).

Table S41. Segmental Duplication Assembler (SDA) assembly collapses
Summary statistic of collapsed sequence per haploid assembly identified by the SDA assembler.

Table S42. DeBreak summary statistics
Number of deletions, insertions, duplications, inversions and total (all SV types combined) detected by DeBreak per sample. SV counts reported in this table refer to SVs passing all DeBreak QC filters.

Table S43. Bionano SV clusters across samples
This table includes the clusters of SVs (≥5 kbp) discovered by optical mapping in the 30 samples. These SVs localize at clean regions where the respective sample has no more than two optical maps aligned at SV breakpoints. Information of the clusters include the Cluster ID (A); SV type (B); genomic position (C-E); the mean size of SVs contributing to the cluster (F); number of samples contributing to the cluster (G); total number of insertions and deletion clusters in the region of the cluster (H-L); length overlapped with UCSC hg38 SuperDups (N); gene overlapped (O); sample population contributing to the cluster (P-T).

Table S44. Bionano unique clusters at segdup and gene
This table includes the insertion and deletion clusters (≥5 kbp) that are fully unresolved in PAV and localize at regions overlapping UCSC hg38 SuperDups and genes. Information of the clusters include the Cluster ID (A); SV type (B); genomic position (C-E); the mean size of SVs contributing to the cluster (F); number of samples contributing to the cluster (G); total number of insertions and deletion clusters in the region of the cluster (H-L); length overlapped with UCSC hg38 SuperDups (N); gene overlapped (O); sample population contributing to the cluster (P-T).

Table S45. Population distribution of the haplotypes observed at chromosome region 3q29
Listing of sample haplotypes corresponding to the haplotype configurations identified by Bionano at the 3q29 locus.

Table S46. Summary statistics of assembly contig coverage for chromosome region 3q29
Coverage summary statistics for all haploid assemblies in the 3q29 region. Gaps are summed up as total base pairs of sequence missing (Delta, bp). Assemblies are ordered by the amount of missing sequence (Delta, pct. rank).

Table S47. FDR estimates by orthogonal support
FDR by Subseq, Chaisson, DeBreak, Inspector and PAV-LRA are shown. “Support +1” and “Support +2” is an FDR estimate requiring support from at least one or two methods, respectively. The callset was filtered by “+1” support, so it’s FDR is 0. To obtain a less biased estimation of FDR, we recalculate “+1” and “+2” support excluding PAV-LRA.

Table S48. VNTR allele rank. Top 1% of VNTR alleles ranked by variance of allele length that are also within 10kb of a gene
VNTR alleles are defined by the span between coordinates of orthologous flanking sequences of VNTR loci detected in GRCh38 and a collection of *de novo* human assemblies. After selecting the top 1% of alleles by length variance, loci are reported that are within 10kb of a gene, along with the most proximal gene and the distance to the gene.

Table S49. Support for SVs from read alignments according to genomic location and support metric

Structural variant (SV) support is defined using various combinations of aligners and stringency of support from raw reads. Aligner classes include either minimap2 (mm2) or LRA alignments, and support by individual aligners. Support is measured using a standard, wide, and broad support range; standard support is a call within 50% of length, within 1 kbp, wide is 50% length within 5 kbp, and broad is any SV of the same type, within 10 kpb. Broad support is intended to represent an upper bound of agreement between methods. Call support is further classified in tandem repeat (TR) and segmental duplications (SDs). Each column specified by Rat is the number of supported calls/number of calls in each category. Support is defined for PAV calls prior to filtering the final callset.

Table S50. Callsets used for sensitivity measurements

SV callset sizes in phase 1 and phase 2. Two different callsets are given for phase 1: long-read only (PhasedSV+MSPAC+fill-in), and a multi-technology set, MNR (merged, nonredundant). The MNR callset includes calls produced by long-read approaches as well as short-read and Bionano Genomics optical map calls.

Table S51. Calls unique to HGSVC phase 1

A comparison of calls generated from assemblies of genomes with variant calls from the phase-1 of HGSVC: HG00514, HG00733, and NA19240. Calls are unique to a dataset if they are more than 1kb away from a similar sized call (50% relative size). Stratification is provided for calls that overlap tandem repeats and segmental duplications. The number of calls from HGSVC phase 1 genomes that do not overlap assemblies generated in HGSVC phase 2 are given in the asm-miss columns.

Table S52. Callset statistics for the lenient set

The table lists the number of variants (≥50 bp) contained in the unfiltered, strict and lenient SV callsets. The machine-learning-based process to define the lenient set can be tuned by setting different cutoffs separating high- from low-quality SVs. Columns C/F state Pearson correlation between PanGenie (2,504 unrelated samples) and PAV genotype allele frequencies. Columns D/G state average genotype concordance for all assembly haplotypes estimated via leave-one-out experiments.

Table S53. Numbers of genotyped SNVs/indels/SVs, eQTLs and sQTLs directly overlapped or in high LD (r2 ≥ 0.8) with GWAS signals

We counted genotyped variants that are in the union GWAS dataset for SNVs, and at least 1bp overlap with any GWAS SNP for indels and SVs respectively. LD was calculated using the genotype of SNVs/indels/SVs and genotypes of nearby GWAS SNPs that are located 1 Mbp upstream and downstream from the variant (SNV) or midpoint of the variant (SV/indel). We reported the number of variants that are in high LD (r2 ≥ 0.8) with GWAS SNPs. For SNV/indel/SV eQTLs and sQTLs, we did the overlapping and LD analysis similar to that of variants respectively, and presented the counts in this table.

Table S54. eQTL and GWAS co-localization results using SMR

We performed the SMR test based on the collection of summary statistics of SNV/indel/SV eQTLs and those of GWAS SNPs that are located 1 Mbp upstream and downstream of the midpoint position of the tested gene. We identified 5,296 eGenes that passed the SMR test with an FDR of 5%.

Table S55. sQTL and GWAS co-localization results using SMR

We performed the SMR test based on the collection of summary statistics of SNV/indel/SV sQTLs and those of GWAS SNPs that are located 1 Mbp upstream and downstream of the midpoint position of the tested gene. We identified 2,826 genes associated with sQTLs that passed the SMR test with an FDR of 5%.

Table S56. PAV vs dipcall and svim-asm

PAV calls were intersected with dipcall and svim-asm on matching samples where SVs were matched if they were within 1 kbp. The "All" section is for all SV calls and gives an FDR estimate within PAV-trimmed regions (Other only trim). The "PAV QC pass" section gives the number of PAV calls passing QC that were not seen by other callers. Section "PAV QC dropped" shows the number of PAV calls that were dropped in QC but have support from the other caller.