
Supplementary information

Complex genetic variation in nearly complete human genomes

In the format provided by the
authors and unedited

Supplementary Text and Figures

Note on contributing authors	5
Sample selection	5
Methods	5
Data production and organization	5
Methods	5
Assembly production and quality control	8
Methods	9
Results	18
Variant discovery and callset development	19
Methods	19
Results	25
Mobile elements	25
Methods	26
Results	28
Inversions	28
Methods	29
Results	29
Segmental duplications	30
Methods	30
Results	31
STR and VNTR annotation	32
Methods	32
Results	32
Y chromosome	33
Methods	33
SVs disrupting genes	34
Methods	34
Results	35
Phased transcripts, isoforms, and effects of SVs	36
Methods	36
Results	37
Functional effects of SVs	39
Methods	39
Results	41
Genotyping	43
Methods	43
Major histocompatibility complex	53
Methods	53

Results	54
Complex structural polymorphisms	58
Methods	58
Results	61
Centromeres	63
Methods	63
Supplementary Figures	64
Fig. 1	64
Fig. 2	66
Fig. 3	67
Fig. 4	68
Fig. 5	69
Fig. 6	70
Fig. 7	71
Fig. 8	72
Fig. 9	73
Fig. 10	74
Fig. 11	76
Fig. 12	77
Fig. 13	78
Fig. 14	79
Fig. 15	80
Fig. 16	81
Fig. 17	82
Fig. 18	83
Fig. 19	84
Fig. 20	85
Fig. 21	86
Fig. 22	87
Fig. 23	88
Fig. 24	89
Fig. 25	90
Fig. 26	92
Fig. 27	93
Fig. 28	94
Fig. 29	96
Fig. 30	98
Fig. 31	100
Fig. 32	101
Fig. 33	103
Fig. 34	105

Fig. 35	107
Fig. 36	108
Fig. 37	109
Fig. 38	111
Fig. 39	112
Fig. 40	113
Fig. 41	114
Fig. 42	115
Fig. 43	116
Fig. 44	117
Fig. 45	118
Fig. 46	120
Fig. 47	123
Fig. 48	125
Fig. 49	126
Fig. 50	127
Fig. 51	129
Fig. 52	130
Fig. 53	131
Fig. 54	132
Fig. 55	133
Fig. 56	135
Fig. 57	136
Fig. 58	137
Fig. 59	139
Fig. 60	140
Fig. 61	142
Fig. 62	143
Fig. 63	144
Fig. 64	145
Fig. 65	146
Fig. 66	148
Fig. 67	149
Fig. 68	150
Fig. 69	151
Fig. 70	152
Fig. 71	153
Fig. 72	154
Fig. 73	155
Fig. 74	157
Fig. 75	158

Fig. 76	159
Fig. 77	160
Fig. 78	161
Fig. 79	162
Fig. 80	163
References	164

Note on contributing authors

Contributing authors are listed in alphabetical order by surname.

Sample selection

Contributing authors: Evan E. Eichler, Pille Hallast, Charles Lee, Katherine M. Munson, Tobias Rausch, Arvis Sulovari

Methods

65 diverse humans were selected for inclusion in our study according to four criteria: (i) Individuals ($n=32$) for which sequencing data had previously been generated⁸ were updated with additional Pacific Biosciences (PacBio) high-fidelity (HiFi) and Oxford Nanopore Technologies (ONT) long-read sequencing coverage. (ii) Additional individuals were chosen from the Reference Genome Improvement project (<https://genome.wustl.edu/items/reference-genome-improvement/>; $n=8$), representing the 1kGP populations ASW, YRI, GWD, MSL, ESN, and ACB, for which deep Illumina coverage exists²⁹ ($n=7$). (iii) We calculated population-specific k -mers⁶⁶ for each 1kGP individual with deep-coverage Illumina data available²⁹ to quantify the distance of each individual relative to its respective population cluster's centroid (**Code Availability**). Briefly, we conducted a principal components analysis using all of the autosomal variants from each individual using the SNPreLate R package¹³⁵ (v1.26.0). The ideal number of clusters (k value) was determined using the gap statistic implemented within the factoextra R package (github.com/kassambara/factoextra; v1.0.7). The optimal number of clusters for EUR, AFR, EAS, SAS, and AMR individuals was found to be 3, 7, 1, 4 and 6, respectively using PC1 and PC2 eigenvalues; this allowed us to compute the distance to the nearest centroid for each individual (**Code Availability**). We then chose additional individuals ($n=13$) either closest or farthest from the centroid of their population cluster. Individuals previously chosen for other long-read sequencing efforts (e.g., by the Human Pangenome Reference Consortium, HPRC) were not considered. (iv) Lastly, three individuals (NA18989, HG01890, HG00358) were included to represent specific Y-chromosomal haplotypes²³.

Data production and organization

Contributing authors: Peter A. Audano, Olanrewaju Austine-Orimoloye, Scott E. Devine, Peter Ebert, Pille Hallast, Patrick Hasenfeld, Kendra Hoekzema, Sarah E. Hunt, Youngjun Kwon, Katherine M. Munson, Likhitha Surapaneni, Feyza Yilmaz, Qihui Zhu

Methods

Data generated as part of this project were derived from lymphoblastoid cell lines available from the Coriell Institute for Medical Research for research purposes (<https://www.coriell.org/>). Regular checks for mycoplasma contamination are performed at the Coriell Institute.

PacBio HiFi sequence production

University of Washington

Data were generated as described previously²³. For additional coverage, high-molecular-weight (HMW) genomic DNA (gDNA) from seven individuals (HG00514, HG00731, HG00732, NA19238, NA19239, NA19240, and NA18939) were processed with the same shearing and size selection parameters but using the SMRTbell Prep Kit v3.0 (PacBio P/N 102-182-700) and SMRTbell Barcoded Adapter Plate 3.0 (PacBio P/N 102-009-200) according to the manufacturer's recommendations. Additional sequencing was performed on either the PacBio Sequel II platform using chemistry v3.2 (PacBio P/N 102-333-300) with a 30-hour movie and 2-hr pre-extension times, or the PacBio Revio platform with chemistry v1 (PacBio P/N 102-817-900) on collection software SMRT Link v12.0 and v13.0, using 30-hour movies and 2- or 1.417-hr pre-extension times. PacBio Revio datasets were post-processed onboard through ccs analysis and DeepConsensus v1.0 or v2.0.

The Jackson Laboratory

Data were generated as described previously²³. To generate additional coverage for three individuals (NA19238, NA19240, and HG03732), HMW gDNA was extracted using the Monarch HMW DNA Extraction Kit for Cells & Blood (NEB T3050L). Libraries were prepared using the SMRTBell Express Template Prep Kit 3.0 (PacBio) with 5 µg of DNA, which had been sheared using the Megaruptor 3.0 (Diagenode). Libraries were sequenced on either PacBio's Sequel II (NA19238 and NA19240) or Revio (HG03732).

University of Maryland Institute for Genome Sciences (UMIGS)

Data were generated essentially as described²³. HMW gDNA from four individuals (NA19434, NA19836, HG03520, and HG02282) was isolated using the Circulomics¹³⁶ DNA preparation kit following the manufacturer's protocols. Four samples were sequenced by shearing the HMW gDNA and isolating an average fragment size of 18 kbp. PacBio libraries were prepared from the sheared 18 kbp fragments and sequenced using PacBio's Sequel II platform with a target coverage of 30X coverage (actual coverages = 27.8X, 32.4X, 47.1X, 43X, respectively for NA19434, NA19836, HG03520, and HG02282).

ONT-UL sequence production

University of Washington

Data were generated as described previously²³. HMW gDNA was extracted from 2 aliquots of 30 million frozen pelleted cells using the phenol–chloroform approach as described previously¹³⁷. Libraries were prepared using the Ultra-Long DNA Sequencing Kit (SQK-ULK001, ONT) according to the manufacturer's recommendations. In brief, DNA from around 10 million cells was incubated with 6 µl of fragmentation mix at room temperature for 5 min and 75°C for 5 min. This was followed by an addition of 5 µl of adapter (RAP-F) to the reaction mix and incubated for 30 min at room temperature. The libraries were cleaned up using Nanobind disks (Circulomics) and long fragment buffer (SQK-ULK001, ONT) and eluted in elution buffer. Libraries were sequenced on the flow cell R9.4.1 (FLO-PRO002, ONT) on a PromethION (ONT)

for 96 h. A library was split into 3 loads, with each load running for 24 h followed by a nuclease wash (EXP-WSH004, ONT) and subsequent reload.

The Jackson Laboratory

Data were generated as described previously²³. HMW gDNA was extracted from 60 million frozen pelleted cells using the phenol–chloroform approach as previously described¹³⁸. Libraries were prepared using the Ultra-Long DNA Sequencing Kit (SQK-ULK001, ONT) according to the manufacturer's recommendations. In brief, 50 µg of DNA was incubated with 6 µl of fragmentation mix at room temperature for 5 min and 75°C for 5 min. This was followed by an addition of 5 µl of adapter (RAP-F) to the reaction mix and incubated for 30 min at room temperature. The libraries were cleaned up using Nanodisks (Circulomics) and eluted in an elution buffer. Libraries were sequenced on the flow cell R9.4.1 (FLO-PRO002, ONT) on a PromethION (ONT) system for 96 h. A library was generally split into 3 loads with each loaded at an interval of about 24 h or when pore activity dropped to 20%. A nuclease wash was performed using the Flow Cell Wash Kit (EXP-WSH004) between each subsequent load.

ONT basecalling

For genome analysis including genome assembly, ONT reads were basecalled with Guppy 5.0.11 with configuration “dna_r9.4.1_450bps_sup_prom.cfg” distributed with the Guppy software.

For ONT-based methylation analyses, reads were basecalled a second time with Guppy 6.5.7 with configuration “dna_r9.4.1_450bps_modbases_5mc_cg_sup_prom.cfg” distributed with the Guppy software.

Bionano Genomics optical genome maps production

Optical genome mapping data for 17 individuals were generated at Bionano Genomics as described previously²³ (**Supplementary Table 3**). We complemented our analysis set with data from previously published HGSVC efforts ($n=13$, PRJEB58376²³; $n=32$, PRJEB41077⁸) and using publicly available data for HG02818 (PRJNA339722) and NA24385/HG002 (**Supplementary Table 3**). Optical genome maps of the seventeen samples were *de novo* assembled using the Bionano Solve v3.5 (<https://bionanogenomics.com/support/software-downloads/>) assembly pipeline, with default settings as described previously⁸.

```
python2.7 Solve3.5.1_01142020/Pipeline/1.0/pipelineCL.py -T 64
-U -j 64 -jp 64 -N 6 -f 0.25 -i 5 -w -c 3 -y -b ${bionano_bnx}
-l ${output_dir} -t Solve3.5.1_01142020/RefAligner/1.0/ -a
Solve3.5.1_01142020/RefAligner/1.0/optArguments_haplotype_DLE1_
saphyr_human.xml -r ${reference_genome}
```

A pairwise comparison of DNA molecules (min. 250 kbp) was generated to produce the initial consensus genome maps. During an extension step, molecules were aligned to genome maps, and maps were extended based on the molecules aligning past the map ends. Overlapping genome maps were then merged. Extension and merge steps were repeated five times before a final refinement of the genome maps. Clusters of molecules aligned to genome maps with

unaligned ends >30 kbp in the extension step were re-assembled to identify all alleles. To identify alternate alleles with smaller size differences from the assembled allele, clusters of molecules aligned to genome maps with internal alignment gaps of size <50 kbp were detected, and the genome maps were converted into two haplotype maps. The final genome maps were aligned to the reference genome, GRCh38.p12.

Strand-seq data generation and data processing

Data for nine individuals were generated as described previously, with Strand-seq libraries generated separately for each cell line following BrdU incorporation²³. We complemented our analysis set with previously published data from HGSVC efforts ($n=9$, PRJEB12849¹³; $n=34$ ⁸, $n=3$ ²³, and $n=9$ ¹⁶, PRJEB39750) and publicly available data for NA24385/HG002 (AWS:S3:s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=NHGRI_UCSC_panel/HG002/hpp_HG002_NA24385_son_v1/Strand_seq/2019-12-16-HWVTJAFXY/).

Hi-C data production

In total, Hi-C data for 63 individuals were collected for analysis (**Supplementary Tables 4, 5**). Specifically, Hi-C data for 20 individuals were generated in this study using Proximo Hi-C kits v4.0 (Phase Genomics), followed by sequencing at the Jackson Laboratory for Genomic Medicine as described previously²³. We complemented our analysis set by adding data from previous HGSVC efforts ($n=30$ using Proximo Hi-C kits v3.0, PRJEB39684⁸; $n=10$, PRJEB58376²³) and from one external source for the three individuals NA19238, HG00513 and HG00732 (PRJNA528584¹³⁹).

Iso-Seq data production

Iso-Seq data were generated at The Jackson Laboratory for 12 (PRJEB75191) individuals as described previously²³ (**Supplementary Table 6**).

RNA-seq data production

RNA-seq data were generated at The Jackson Laboratory for 21 individuals (PRJEB75192) as described previously²³, with the exception that 1 million cells were used as a starting material, and paired-end 150 bp reads were sequenced on an Illumina NovaSeq 6000 platform. We complemented our analysis set with previously published data from HGSVC efforts ($n=33$, PRJEB39684⁸; $n=10$, PRJEB58376²³; **Supplementary Table 7**).

Assembly production and quality control

Contributing authors: Haoyu Cheng, Zechen Chong, Peter Ebert, Evan E. Eichler, William T. Harvey, Mir Henglin, Sergey Koren, Youngjun Kwon, Tobias Marschall, Adam M. Phillippy, Mikko Rautiainen, Yuwei Song

Methods

Verkko hybrid genome assembly

Sequencing data from all individuals were uniformly processed with a Snakemake v7.19.1¹⁴⁰ workflow implementing the hybrid genome assembly using Verkko⁹ (v1.4.1; see **Code Availability**). Verkko was first executed with all PacBio HiFi and ONT data per individual to generate an unphased whole-genome assembly. The resulting assembly graph in GFA format was then forwarded into the previously published Graphasing¹² (v0.3.1-alpha) phasing pipeline that used all available Strand-seq data per individual to generate phasing paths through the assembly graph. The phase information was further processed by the Verkko companion tool Rukki⁹ to generate a Verkko-compatible input file. This input file was then used in addition to the read input data for the second Verkko run resulting in fully phased assemblies for each individual. The Verkko pipeline itself executed the following tools as part of the assembly process: MBG¹⁴¹ (v1.0.15), GraphAligner¹⁴² (v1.0.17), and MashMap⁸⁰ (v3.0.6) for all individuals except NA19240, HG00514, HG00732, HG00096 and NA19239. For this last set of individuals, the Verkko pipeline tools were necessarily updated to MBG¹⁴¹ (v1.0.16) and GraphAligner¹⁴² (v1.0.18) to solve a bug originally discovered in the HG00096 assembly process, which is documented in the Verkko GitHub issue #189. The dataset for HG00512 contained one ONT file that was later discovered to be a contaminant (library id 23-lee-007_PCA100115_3F-run15, **Supplementary Table 2**). We examined the issue with the Verkko developers (*personal communication*) and decided not to rerun the assembly because Verkko's process of aligning the ONT reads against an initial PacBio HiFi-based assembly graph is inherently robust against such contamination.

Verkko command to produce unphased assemblies:

```
verkko --hifi ALL-HIFI-READS.fastq --nano ALL-ONT-READS.fastq  
--screen human -d WORK-DIR-UNPHASED
```

Verkko command to produce phased assemblies:

```
verkko --hifi ALL-HIFI-READS.fastq --nano ALL-ONT-READS.fastq  
--screen human -d WORK-DIR-PHASED --assembly WORK-DIR-UNPHASED  
--paths STRANDSEQ-PHASING-PATHS.gaf
```

The phased assembly output of Verkko consisted of the following assembly units:

1. haplotypes 1 and 2: the main haploid assemblies for each genome
2. unassigned: a set of unphased contigs, i.e., contigs that are not assigned to a specific haplotype.
3. disconnected: a set of usually short assembled sequences that could not be connected to any other component in the assembly graph.
4. rDNA: the set of sequences identified by Verkko as rDNA sequences (see option "--screen")
5. EBV: the set of sequences identified by Verkko as Epstein-Barr viral sequences
6. mito: the set of sequences identified by Verkko as mitochondrial sequences (see option "--screen")

If not explicitly stated otherwise, the unassigned, disconnected, rDNA, EBV and mitochondrial sequences were omitted from any downstream analysis.

We emphasize here that, due to the characteristics of the Strand-seq technology, it is not possible to *a priori* determine the parent of origin for a given assembled haplotype of a genome. Hence, the common denominations haplotype 1 (hap1/H1) and haplotype 2 (hap2/H2) are in principle interchangeable and, in particular, have not been harmonized - in terms of the parent of origin - across individuals.

Hifiasm hybrid genome assembly

Hifiasm (UL)¹⁰ (v0.19.6) assemblies were generated in three steps. First, a pseudo-phased assembly was generated using both the PacBio HiFi and ONT reads:

```
hifiasm -o {output.prefix} -t {threads} --ul {input.ont_fastq}
{input.hifi_fastq}
```

The resulting assembly graph in GFA format was then forwarded into the previously published Graphasing phasing pipeline¹² as for Verkko above. The resulting phasing paths and sequences were then used to generate the yak (github.com/lh3/yak) databases for phasing the assemblies. With these yak databases, the command was as follows:

```
hifiasm -o {output.prefix} -t {threads} -1 {input.yak_h1} -2
{input.yak_h2} --ul /dev/null /dev/null
```

This process uses the same error corrected reads as the initial run but applies the phasing information to the resulting assembly graph.

Standardized evaluation of genome assemblies

All phased assemblies were subjected to the same evaluation pipeline implemented in a Snakemake¹⁴⁰ (v7.19.1) workflow (**Code Availability**). The input to the evaluation pipeline consisted of FASTA files representing the different assembly units, e.g., haplotypes 1 and 2 or the unphased (“unassigned”) sequences for Verkko assemblies (see “Verkko hybrid genome assembly” above), plus the same PacBio HiFi and ONT reads that were used for the assembly.

Identification of contaminants

We employed NCBI’s Foreign Contamination Screening (FCS) tool⁷¹ (v0.4.0) with database release 2023-01-24 to screen all assembly units for remaining sequencing adapters or common contaminants such as mycoplasma. Sequences flagged as contaminants of any type were removed from the assembly sequence files - with the exception of Verkko’s “EBV” output - and the cleaned-up assemblies were then further characterized as explained in the following steps.

FCS-GX command line to screen for foreign organism contamination:

```
python3 fcs.py screen genome --fasta ASSEMBLY.fasta --out-dir OUT_DIR  
--gx-db r2023-01-24/all --tax-id 9606
```

FCS-adaptor command line to screen for adaptor and vector contamination:

```
run_fcsadaptor.sh --fasta-input ASSEMBLY.fasta --output-dir OUT_DIR  
--euk --container-engine singularity --image fcs-adaptor.sif
```

Assembly contig-to-reference alignment

All assembled contigs were aligned against two human reference assemblies: a variant of GRCh38 excluding the ALT sequences (GRCh38-NoAlt)⁸ and the T2T-CHM13v2.0 assembly (T2T-CHM13, UCSC hs1)⁴, which includes a complete Y chromosome of the individual HG002/NA24385. Assemblies from female individuals were aligned against a T2T-CHM13 reference with N-masked PAR regions of the Y chromosome. All alignments were performed with minimap2^{78,79} (v2.26) with the following command: `minimap2 -a -x asm20 --cs --eqx -t THREADS -R READGROUP REF-ASSEMBLY.fasta INPUT-ASSEMBLY.fasta`. Alignments were filtered to exclude unmapped sequences (flag 1540) with SAMtools⁹⁴ (v1.17). The contig-to-reference alignments were used to derive a reference chromosome assignment for the assembled contigs by selecting the chromosome with the highest number of matches to the assembled contig as reported by minimap2. The sex (male/female) for each assembled haplotype of the male individuals was determined in the same way.

Read-to-assembly alignment

All PacBio HiFi and ONT reads used for the assembly were aligned against the complete assembly for the purpose of quality control. All alignments were performed with minimap2^{78,79} (v2.26) with the following command: `minimap2 -a -x READ-PRESET --MD --eqx --cs -t THREADS -R READGROUP INPUT-ASSEMBLY.fasta ALL-READS.fastq`. Alignments were not filtered by default. The minimap2 parameter presets “map-hifi” and “map-ont” were used to align the PacBio HiFi and the ONT reads, respectively.

Flagging regions of putative assembly collapses with NucFreq

We used an adapted variant of the tool NucFreq⁷³ (fork NucFreqTwo / branch “split-two-phases” / commit bd080aa) to identify positions where the PacBio HiFi read-to-assembly alignments (see section “Read-to-assembly alignment” above, filtered here with flag 3844) showed an elevated count of the second most common base in the alignments for the respective position in the assembly. We applied the following filtering criteria to identify those positions: (1) the second most common base must be observed in at least two alignments at that position; (2) the second most common base must be observed in at least 10% of all alignments at that position. Next, regions with an increased density of such discordant alignment positions were flagged if at least five discordant positions occurred in a window of at most 500 bp. A common interpretation of NucFreq flagged regions is that they represent potential assembly collapses if they coincide with regions that exhibit an elevated read depth. We hence annotated all NucFreq regions with the

median PacBio HiFi read depth normalized to a percentage value relative to the global median PacBio HiFi depth for the respective individual for comparability using custom scripts. In the integrative QC analysis (see section “Integrative analysis of QC annotations” below), all regions flagged by NucFreq were labeled “NUCFRQ”.

The adapted version of NucFreq we used (termed “NucFreqTwo”) is quasi-identical to the original NucFreq code but separates the computations for identifying the flagged regions from the plotting output of said regions. These minor changes were necessary because the original NucFreq code broke for more fragmented assemblies, i.e., in cases where a lot of flagged regions were plotted automatically. The following command was executed to identify the regions flagged for potential assembly collapses:

```
NucFreqTwo.py          --infile          HIFI-READ-TO-ASSEMBLY.bam
--flag-discordant-pct 10 --flag-discordant-abs 2 --flag-min-interval
500 --flag-num-hets 5 --min-region-size 500000
```

```
# last parameter was introduced to run NucFreq[Two] genome-wide as
opposed
```

```
# to the original NucFreq which is commonly executed with a list
```

```
# of regions (BED file) selected for flagging/plotting. The above
filter
```

```
# skips over regions smaller than 500000 in the BAM file
```

Flagging regions of putative assembly errors with Flagger

We used Flagger¹ (v0.3.3), a read-alignment-based pipeline, to detect putative assembly errors in the diploid assemblies. PacBio HiFi reads were aligned to the phased assemblies, i.e., combining haplotype 1 and 2, in a haplotype-aware manner. For improved mapping in repeat regions, Meryl⁷² (v1.0) and Winnowmap2^{143,144} (v2.03) were used to generate repeat-aware read alignments:

```
meryl count k=15 output MERYL_DIR ASSEMBLY.fasta
meryl print greater-than distinct=0.9998 MERYL_DIR > REP_K15.txt
winnowmap -W REP_K15.txt -t THREADS -I 10G -Y -ax --MD --cs -L --eqx
ASSEMBLY.fasta HIFI_READS.fastq.gz
```

Flagger was executed using a WDL pipeline. The command line used for execution is as follows:

```
java -jar cromwell.jar run flagger_end_to_end.wdl --inputs INPUT.json
--metadata-output OUTPUT.json
```

The configuration options in the INPUT.json file were set as follows:

```
{
  "FlaggerEndToEnd.refSDBed": REF_SD.bed,
  "FlaggerEndToEnd.preprocess.variantCallingMemory": 48,
  "FlaggerEndToEnd.sampleName": SAMPLE_NAME,
  "FlaggerEndToEnd.fai": CONCAT_ASSEMBLY.fasta.gz.fai,
  "FlaggerEndToEnd.stats.threadCount": 4,
  "FlaggerEndToEnd.refBiasedRegionFactorArray": [1.25, 0.75],
  "FlaggerEndToEnd.flagger_alt_removed.isDiploid": true,
  "FlaggerEndToEnd.stats_alt_removed.threadCount": 4,
  "FlaggerEndToEnd.flagger.isDiploid": true,
  "FlaggerEndToEnd.stats_alt_removed.memSize": 4,
  "FlaggerEndToEnd.maxReadDivergence": 0.02,
  "FlaggerEndToEnd.refCntrCtBed": REF_CENTROMERE_TRANSITION.bed,
  "FlaggerEndToEnd.refSexBed": REF_SEX_CHROMS.bed,
  "FlaggerEndToEnd.variantCaller": "dv",
  "FlaggerEndToEnd.stats.memSize": 4,
  "FlaggerEndToEnd.secphaseOptions": "--hifi",
  "FlaggerEndToEnd.project.isAssemblySplit": false,
  "FlaggerEndToEnd.assemblyFastaGz": CONCAT_ASSEMBLY.fasta.gz,
  "FlaggerEndToEnd.suffix": "",
  "FlaggerEndToEnd.refBiasedBlocksBedArray": [REF_HIFI_R1_BIASED.bed, REF_HIFI_R2_BIASED.bed],
  "FlaggerEndToEnd.refCntrBed": REF_CENTROMERIC_SATELLITES.bed,
  "FlaggerEndToEnd.refBiasedRegionNameArray": ["hifi_biased_high", "hifi_biased_low"],
  "FlaggerEndToEnd.refName": "hs1",
  "FlaggerEndToEnd.readAlignmentBam": HIFI_READS_TO_ASSEMBLY.bam,
  "FlaggerEndToEnd.hap1ToRefBam": HAP1_ASSEMBLY_TO_REF.bam,
  "FlaggerEndToEnd.hap2ToRefBam": HAP2_ASSEMBLY_TO_REF.bam
}
```

In cases where the primary alignment is deemed inaccurate due to assembly errors, Secphase (github.com/mobinasri/secphase; v0.4.3; integrated in Flagger) identifies the correct haplotype among the secondary alignments, ensuring more accurate read assignment to each haplotype.

For the integrative QC analysis (see section "Integrative analysis of QC annotations" below), all non-haploid Flagger labels were used. The Flagger labels are all prefixed with "FLG" and were shortened as follows: error (-ERR), unknown (-UNK), collapsed (-COL), duplicated (-DUP).

Variant-based QV estimation with DeepVariant

We used DeepVariant⁷⁵ (v1.6.0) to call short variants in the PacBio HiFi read-to-assembly alignments (see section "Read-to-assembly alignment" above, filtered here with flag 3844) that could indicate small assembly errors. The resulting callset was quality-filtered with bcftools⁹⁴ (v1.17) and all remaining variants were counted as putative assembly errors:

```
/opt/deepvariant/bin/run_deepvariant --model_type PACBIO --ref
INPUT-ASSEMBLY.fasta --reads HIFI_READS.bam --num_shards THREADS
--output_vcf DV_CALLS.vcf --noruntime_report --novcf_stats_report
--sample_name SAMPLE --intermediate_results_dir TMP_RUN
```

```
bcftools view -f PASS -i 'QUAL>=10 && FORMAT/DP>=5' DV_CALLS.vcf
```

All variant calls remaining after filtering were counted in base pairs as putative assembly errors and a sequence QV estimate was derived as described previously⁸ after adjusting the number of error-free base pairs by subtracting the total length of all gaps (stretches of Ns) in the respective sequence. In the integrative QC analysis (see section "Integrative analysis of QC annotations" below), all regions flagged by DeepVariant were labeled as "DEEPVR".

k-mer based analysis of assemblies with Merqury

We used Meryl⁷² (v1.0) and Merqury⁷² (v1.0) to estimate the QV values of the assemblies. The QV values were calculated as phred-scaled scores¹⁴⁵, based on the error rates determined by comparing the *k*-mer databases generated from the sequencing datasets with the assembled sequences. Illumina sequencing data, which has high base accuracy, was used to construct the *k*-mer databases.

To generate the *k*-mer database from the sequencing dataset, the following command was executed:

```
meryl k=21 count ILLUMINA_READS.fastq.gz output READS_DB.meryl
```

To compute the QV value for the entire diploid assembly, as well as the QV values for each haplotype (hap1 and hap2), the following command was executed:

```
merqury.sh READS_DB.meryl HAP1_ASSEMBLY.fa.gz HAP2_ASSEMBLY.fa.gz
MERQURY_OUTPUT
```

Assessing gene completeness of the assemblies with compleasm

We used compleasm⁷⁶ (v0.2.5) to assess the gene completeness of the assembled haplotypes in terms of the presence/absence of known single-copy orthologs in the OrthoDB⁷⁷ (v10) primate database.

```
compleasm run --mode busco -L DB_DIR/ -l primates_odb10 --threads  
THREADS -o OUT_DIR/ -a INPUT-ASSEMBLY.fasta
```

The results of compleasm includes the information of single-copy genes that exhibit a fragmented alignment to the respective assembly or that show additional copies. That information was incorporated into the integrative QC analysis (see section "Integrative analysis of QC annotations" below) by labeling the respective regions as "BSCFRG" (fragmented) and "BSCDUP" (duplicated), respectively.

Identifying complete chromosome assemblies

We labeled chromosomes as T2T by applying the following set of criteria:

1. Telomeres were assembled on both arms of the chromosome.
2. The chromosome spans the respective T2T-CHM13 reference chromosome approximately from end to end.
3. The QV estimate for the assembled sequence is above 50 (see sections "Variant-based QV estimation with DeepVariant" and "K-mer based analysis of assemblies with Merqury" above).

The required input data to check criteria 1 and 2 were computed as follows:

```
(1-telomere): seqtk telo INPUT-ASSEMBLY.fasta > ASSEMBLY.telo.tsv
```

```
(2-ref span): mashmap -f one-to-one --pi 99 --seqLength 100000  
--dense -r REF-ASSEMBLY.fasta -q INPUT-ASSEMBLY.fasta --output  
APPROX-ALIGN.paf
```

The adapted contig-to-reference alignment strategy using MashMap⁸⁰ (v3.1.3) resulted from the observation that base-level alignments occasionally could not traverse complex regions, e.g., on chromosome 16. This resulted in fragmented alignments of contiguously assembled chromosomes, which was avoided by resorting to the approximate, i.e., not base-level accurate alignment strategy of mashmap. Mashmap was executed with a stringent setting for reporting alignments only above a threshold of 99% sequence identity to then identify a high-confidence anchor region, i.e., the largest reported alignment. The span of the assembled chromosome relative to the reference was then computed relative to this anchor region. An assembled sequence was considered reference spanning if it covered more than 95% of the reference sequence length. This threshold was lowered to 60% for the Y chromosome due to the recently reported high variation in total length²³, and to 90% for the T2T-CHM13 chromosome 9 to account for the multi-megabase HSat3 duplication in the CHM13 cell line⁴.

We note here that for HG00732, the X chromosome assembly in one haplotype was deteriorated to a degree that we labeled it as missing (**Fig. 1e**). Based on the progressive loss of read depth on the X chromosome across the three PacBio HiFi sequencing batches for that individual (**Supplementary Fig. 6**), we assume that this is an artifact caused by the cell line.

Detection of phasing inconsistencies

The Graphasing pipeline¹⁴⁶ assumes that all unitigs in the input assembly graph are specific to a single haplotype. Consequently, haplotype switches present in the input assembly graph will be propagated into the final assembly. We developed the following method to localize these events (referred to as “breakpoints” for brevity) and applied it to the Verkko assemblies.

Strand-seq based (phasing/switch-error) breakpoints were detected using a method similar to that used in the breakpointR R package¹⁴⁷. The method leverages the haplotype-informative signal inferred from the alignment orientation of the Strand-seq reads, i.e., either in ‘Crick’ (C) or in ‘Watson’ (W) direction relative to the assembly sequence. Consequently, a change in the predominant alignment orientation, e.g., from mostly Crick at the beginning of an assembled contig to mostly Watson at its end, indicates a putative phasing error introduced by the assembler that is not supported by the Strand-seq signal.

Our method scans each unitig in windows of varying size by fixing the number of read alignments per window. For each window, the following two-sample binomial proportion test statistic Z is calculated:

$$Z = \frac{p_1 + p_2}{\sqrt{p(1-p)\left(\frac{1}{W_1+C_1} + \frac{1}{W_2+C_2}\right)}}$$
$$p_1 = \frac{W_1}{W_1+C_1}, p_2 = \frac{W_2}{W_2+C_2}, p = \frac{W_1+W_2}{W_1+C_1+W_2+C_2}$$

With W_1 and C_1 equal to the number of Watson and Crick reads in the back half of the sliding window respectively, and W_2 and C_2 equal to the number of Watson and Crick reads in the front half of the sliding window respectively. P-values were calculated from the Z statistics, after which a Bonferroni correction is applied at the unitig level to account for the multiple testing problem in this setting. Finally, breakpoints can be identified as locations with p-values lower than a given threshold, with adjacent locations merged into “peaks”. The starting and ending coordinate of each peak is reported, giving a high-confidence window in which the breakpoint exists. For breakpoint calling, we used a window size of 300, and a p-value threshold of 10^{-8} . To filter out small noise peaks, peaks were further filtered to those wider than 75 continuous locations below the p-value threshold. Breakpoints are calculated during the Graphasing pipeline if `calc_breakpoints: True` is set in the configuration file. Besides the Strand-seq reads, the input for Verkko assemblies are the homopolymer compressed assembly graph (“assembly.homopolymer-compressed.gfa”) and the assembly PacBio HiFi coverage (“assembly.hifi-coverage.csv”), while for hifiasm only the unitigs graph (“*p_utg.gfa”) is needed. In both cases, the paths to these files, as well as the “assembler” flag need to be specified in the sample sheet listed in the configuration file.

```
snakemake -d ../wd/ --configfiles config/config.yaml
// see text above for exec details Verkko vs hifiasm
```

Assembly analysis with Inspector

Inspector⁷⁴ (v1.2) was used to evaluate the assembly errors of 65 assemblies, which were recorded as BED files. Assembly completeness was also checked using inspector reference mode aligned against the T2T-CHM13 genome.

```
inspector.py -c INPUT-ASSEMBLY.fasta -r HiFi_RAWREADS.fastq.gz --ref  
REF-Genome.fasta -o OUTPUT_DIRECTORY/ -t8 --datatype hifi
```

```
inspector.py -c INPUT-ASSEMBLY.fasta -r ONT_RAWREADS.fastq.gz --ref  
REF-Genome.fasta -o OUTPUT_DIRECTORY/ -t8 --datatype nanopore
```

For each individual, 28 metrics were evaluated using a reference-based approach for both Verkko and Hifiasm assemblies, employing either PacBio HiFi raw reads or ONT raw reads with the T2T-CHM13 reference (**Supplementary Tables 12, 13, 21, 22**). Basic contig and alignment statistics were calculated, and assembly errors were classified into structural assembly errors (>50 bp) and small-scale errors (≤50 bp). The types of assembly errors considered included base substitution, collapse, expansion, inversion, and haplotype switch. Overall assembly quality (QV) was measured as the ratio of the total bases affected by errors plus the number of inversion errors to the total assembly length. All evaluated parameters are depicted in (**Supplementary Figs. 4, 5, 7, 8**). In the integrative QC analysis (see below), the Inspector flagged regions were labeled “ISPC” plus the suffix “HF” or “ON” to indicate the input read type, i.e., PacBio HiFi or ONT, respectively.

Determining the status of previously reported assembly gaps

The evaluation of the status of a set of $n=592$ previously reported assembly gaps observed in PacBio HiFi-only assemblies² was implemented by combining the base-level alignments generated with minimap2 (see section “Assembly contig-to-reference alignment” above) with the approximate alignments generated with mashmap (see section “Identifying complete chromosome assemblies” above) in custom code (see **Code availability**). The output of both aligners was combined to better handle the substantial size variation of the reported assembly gaps (1 bp min, ~30 Mbp max, ~3.6 kbp median). As an example, the largest gaps are located on chromosome Y (q12 region, ~30 Mbp) and on the short arms of the acrocentric chromosomes (up to ~3.8 Mbp on chromosome 13), which are regions that notoriously break alignments due to their repetitive nature. Briefly, a gap was considered covered if any of the two aligners produced a spanning alignment. Next, we annotated the respective spanning contig with Flagger and NucFreq (see respective sections on Flagger “Flagging regions of putative assembly errors with Flagger” and on NucFreq “Flagging regions of putative assembly collapses with NucFreq” above) in analogy to the curation process applied for the centromeres and the chromosome Y analysis (see main text). For this analysis, we decided not to consider any potential errors identified by Flagger or NucFreq only in the actual gap interval, which can be as small as a single base pair, but to limit the overall error rate of the entire assembled sequence for both tools to <1%.

We emphasize here one special case that resulted from our sample and data selection that included NA24385/HG002, which is the individual that was used to produce the first complete chromosome Y as part of the T2T-CHM13v2 reference²⁴. Consequently, our assembly of

NA24385 is extremely biased towards the reference we use as an alignment target. We assume that this bias in conjunction with the substantial size variation reported for the chromosome Yq12 region²³ explains why only our NA24385 assembly appears to have closed the ~30 Mbp gap in Yq12 despite other complete Y chromosomes in our dataset (see main text).

Results

Inspector

The Inspector analysis using both PacBio HiFi and ONT reads for the Verkko and the hifiasm assemblies resulted in QV estimates ranging from 43.7 to 56.0, with a median of 48.8 for Verkko. Similarly, hifiasm assemblies exhibited QV values ranging from 45.8 to 57.1, with a median of 49.6. Verkko assemblies assessed using ONT raw reads had QV values ranging from 31.2 to 34.2, with a median of 32.8, whereas hifiasm assemblies had QV values ranging from 31.3 to 36.1, with a median of 33.0 (**Supplementary Tables 12, 13, 21, 22**).

Structural errors, defined as the sum of structural assembly errors including collapse, expansion, inversion, and haplotype switch, were also calculated. Verkko assemblies evaluated with PacBio HiFi raw reads had structural error counts ranging from 4 to 51, with a median of 20. In comparison, hifiasm assemblies evaluated with PacBio HiFi raw reads had structural error counts ranging from 13 to 127, with a median of 28.

Genome completeness was assessed using the T2T-CHM13 as the reference. Verkko assemblies evaluated with PacBio HiFi raw reads demonstrated completeness ranging from 85.73% to 94.15%, with a median of 93.40%. Hifiasm assemblies evaluated with PacBio HiFi raw reads exhibited completeness ranging from 88.83% to 96.43%, with a median of 93.77%.

Integrative analysis of QC annotations

The merged QC labels (Methods, section “Assembly production and quality control”) revealed a strong bias for certain tools to flag singleton regions, i.e., regions that were not supported by another tool. This observation was most pronounced for the Inspector analysis using ONT reads, which uniquely flagged >99% of all the flagged regions per Verkko assembly on average (**Supplementary Table 14**). We thus decided to exclude the ONT-based Inspector results, assuming that the inherently higher error-rate in ONT reads makes this data type unsuitable for discovering in particular small-scale errors in the Verkko assemblies (**Supplementary Table 15**).

Next, we computed the overlap and preferential association among all QC labels and regions that are prone to give rise to spurious alignments due to their repetitive nature (telomeres, segmental duplications >98% identity and centromeres), which may affect the performance of all QC tools that rely on read-to-assembly alignments (**Supplementary Table 16**). Limiting to the statistically significant associations (**Supplementary Table 17**; Fisher’s exact test, two-sided, from Python’s `scipy.stats`; p-values multiple-testing corrected with Benjamini-Yekutieli procedure¹⁴⁸), we found a depletion of the Flagger-related labels (FLG*) and of NucFreq in centromeres, which is consistent with the manual curation of those regions before accepting them as correctly assembled (see main text). Other enrichments suggest that high-identity

segmental duplications may either still pose a challenge for hybrid assemblers or are prone to give rise to false positives as many QC flags exhibit a pronounced positive odds ratio in that region type (BUSCO duplicated genes; Flagger false duplications; Strand-seq breakpoints). Almost all tools show an enrichment in assembled telomeres (DeepVariant, Flagger errors and unknowns, Merqury and NucFreq) except for Inspector, which may suggest that Inspector lacks sufficient sensitivity in these areas of the assemblies.

Finally, we created a merged region set for all QC-related annotations and counted only those regions as true positives, i.e., presumably genuine assembly errors, where two distinct tools flagged overlapping intervals (**Supplementary Table 18**). Regions that were not flagged by at least two tools (in the following: “clean” regions) represent 99.6% (median) of the phased sequence in the Verkko assemblies. The minimal (“size min”, **Supplementary Table 18**) region size for both flagged and clean regions (median 1 bp) suggests, though, that a slightly altered merging strategy, i.e., merging based on a distance criterion instead of requiring overlap, would immediately result in a different estimate of the remaining errors in the assemblies. Moreover, we emphasize here that the above estimate refers only to the phased sequences of the assemblies and that sequences such as the rDNA (as identified by Verkko) are excluded from these considerations; the median assembly size (~5.98 Gbp, **Supplementary Table 15**) falls clearly short of a pseudo-diploid T2T-CHM13 reference assembly (~6.02 Gbp). However, restricting the view on the phased sequences is partly a technical necessity because Flagger’s inference of the “haploid” label or NucFreq’s approach of scanning the read-to-assembly alignments for a locally increased abundance of mismatching nucleotides are not applicable if the underlying assembled sequence has an unclear phasing status.

Variant discovery and callset development

Contributing authors: Peter A. Audano, Christine R. Beck, Carolyn A. Paisie

Methods

Genome references

The GRCh38 reference was previously constructed⁸ by obtaining the hg38 reference assembly from the UCSC Genome Browser¹⁴⁹ and removing ALT sequences leaving just the primary assembly (includes chromosome scaffolds as well as unplaced and unlocalized scaffolds). The T2T-CHM13 reference is the hs1 reference assembly obtained from the UCSC Genome Browser and was not modified.

Variant discovery

PAV

We ran PAV⁸, a development version now released under v2.4.1 (**Code Availability**). All default parameters were used specifying only “reference”, “aligner”, and “assembly_table” in the PAV configuration JSON. Both Verkko and hifiasm assemblies were run with PAV. From Verkko, we input haplotypes 1 and 2 and the unassigned contigs as three distinct haplotypes. From hifiasm, we input haplotypes 1 and 2. Each PAV callset was replicated using minimap2⁷⁸ (v2.26) and

LRA⁸¹ (v1.3.7.2) alignments resulting in two independent callsets for each aligner. In total, eight PAV callsets were produced (two assemblies × two aligners × two references).

Genotypes reported in the original PAV VCF files follow the order “h1|h2|un” (haplotype 1, haplotype 2, unassigned).

DipCall

We ran DipCall⁸² (v0.3) on Verkko and hifiasm assemblies using default parameters for both GRCh38 and T2T-CHM13 references. From both Verkko and hifiasm we input haplotypes 1 and 2.

SVIM-asm

We ran SVIM-asm⁸³ (v1.0.3) on Verkko and hifiasm assemblies using svim-asm haploid using the following parameters: “--tandem_duplications_as_insertions --interspersed_duplications_as_insertions” for both GRCh38 and T2T-CHM references. SVIM-asm uses the same alignments as PAV, which was generated with PAV’s CRAM target to create input files for SVIM-asm. From both Verkko and hifiasm we input haplotypes 1 and 2.

Read alignments and simple repeat annotations

PacBio HiFi reads were aligned to both GRCh38 and T2T-CHM13 references using pbmm2 (<https://github.com/PacificBiosciences/pbmm2>; v1.12.0) with “--sort --preset CCS -L 0.1 -c 0”. ONT reads were aligned to both GRCh38 and T2T-CHM references using minimap2 (v2.26) with “-ax map-ont --eqx -R”.

Tandem repeat annotations were generated by retrieving “Simple Repeats” tracks from UCSC for both references. Records were merged to long non-overlapping regions where any records intersecting or within 200 bp were condensed to a single record. These annotations were input into PBSV and other variant callers using tandem repeat annotations on the reference.

PBSV

CCS reads were aligned as described above for both GRCh38 and T2T-CHM13 references, and variants were called using PBSV (<https://github.com/PacificBiosciences/pbsv>; v2.9.0). The PBSV workflow was executed separately for each genome and for each chromosome. First SV signatures were discovered with “pbsv discover --tandem-repeats <TR.bed> --region <CHROM>” where TR.bed is the simple repeat table described above. Variants were called with “pbsv call” with “--ccs -A 2 -O 2 -P 20 -m 20” for CCS. For each genome, the per-chromosome calls were concatenated, sorted, and compressed with BCFtools⁹³ (v1.17).

Sniffles

Variants were called using Sniffles⁸⁴ (v2.0.7) using the following parameters: “--tandem-repeats TR.bed --minsvlen 5” for both GRCh38 and T2T-CHM13 references for both CCS and ONT reads. “TR.bed” is the simple repeats table described above (PBSV).

Delly

Variants were called using Delly⁸⁵ (v1.1.6) using default parameters for both GRCh38 and T2T-CHM13 references for both CCS and ONT reads.

cuteSV

Variants were called using cuteSV⁸⁶ (v2.0.3) using the suggested parameters (<https://github.com/tjiangHIT/cuteSV>) for both GRCh38 and T2T-CHM13 references for both CCS and ONT reads.

DeBreak

Variants were called using DeBreak⁸⁷ (v1.0.2) using the following parameters: “debreak --bam merged.sort.bam --outpath debreak_out/” for both GRCh38 and T2T-CHM13 references for both CCS and ONT reads. Parameter “--rescue_dup” was added for T2T-CHM13.

SVIM

Variants were called using SVIM⁸⁸ (v2.0.0) using default parameters for “svim alignment” for both GRCh38 and T2T-CHM13 references for both CCS and ONT reads.

DeepVariant

For PacBio HiFi reads, variants were called using DeepVariant⁷⁵ (v1.5.0) using default parameters with the model “--model_type=PACBIO” for both references. For ONT reads, variants were called using PEPPER-Margin-DeepVariant⁹⁰ r0.8 using default parameters with the model “--ont_r9_guppy5_sup” for both references.

Clair3

Variants were called using Clair3⁸⁹ (v1.0.4) using default parameters with “--platform=hifi” --model_path=hifi_sequel2” for CCS reads and with “--platform=ont” --model_path=r941_prom_sup_g5014” for ONT reads.

Callset annotation, merging, and QC

Merging and support intersects

Variants were compared by SV-Pop⁸ (v3.4.4) with parameters tuned for each variant type. These comparisons were used to construct a merged callset as well as compare the merged callset to output from other callers to identify support (i.e., determine if another caller makes the same variant call).

Insertions and deletions

A multi-stage merging criterion was used to merge insertions and deletions. All variants are compared by the first criteria and matches are set aside. The remaining variants are compared by the second set of criteria, and so on. Pairs of variants that pass any of the stages are merged together. The staged system allows for the merging process to try different sets of criteria

designed to intersect large and small variants and account for breakpoint differences we have previously observed. For example, 50% reciprocal overlap works well for very large SVs, but small breakpoint differences for smaller SVs have a much larger impact on overlap. The fourth stage was added to better intersect SVs in tandemly repeated loci¹⁵⁰.

SV-Pop Parameters: “nr::exact:ro(0.5):szro(0.8,200):szro(0.8,unlimited,3):match(0.8)”

1. exact: Exact match by size and position
 - a. Best matches first
2. ro(0.5): 50% reciprocal overlap
 - a. For large SVs
3. szro(0.8,200): Within 200 bp and 80% size overlap
 - a. For smaller SVs and indels, less sensitive to small breakpoint differences
4. szro(0.8,unlimited,3): Within 3× SV size and 80% size overlap
 - a. For tandemly repeated variants that may span large regions

All stages require SV sequences to match by 80% identity using a circular alignment in the case of variants shifted through breakpoint repeats. Only variants of the same type are intersected. Some variant callers did not make SV sequences readily available, so we dropped this sequence match requirement when intersecting with those callsets.

To reduce merging artifacts around the SV size cutoff (50 bp), SVs and indels were intersected as a single callset, and we separated SVs and indels after merging. Support intersects were done the same way with SVs and indels together from the PAV and supporting callers.

Inversions

Inversions were intersected by 20% reciprocal overlap to allow for a range of variant sizes that often results from calling inversions from different methods.

SNVs

SNVs were intersected only if they exactly matched by both reference position and alternate base.

Linking variants across references

A complete callset was produced on both T2T-CHM13 and GRCh38, although it is not possible to identify which variants are in both references from these callsets alone. Using assembly coordinates for each variant call reported by PAV, we translated all variants in both references to assembly coordinates and intersected using the same parameters outlined above for intersecting variants in reference space. This was performed independently for each assembled haplotype.

For each variant in the merged callset, we examined the results of linking in all haplotypes it was identified in. If all haplotypes supporting the call pointed to the same merged sample in the other reference, we linked the merged variants. For example, if a variant in T2T-CHM13 is

derived from three haplotypes, then all three haplotype variants should be assigned to the same merged variant in GRCh38 to link the merged variant in T2T-CHM13 to the variant in GRCh38.

Reference annotations

Centromere and satellite annotation for T2T-CHM13

We obtained the “CenSat” track from the UCSC browser for the T2T-CHM13 reference genome on 2024-03-13 (<https://hgdownload.soe.ucsc.edu/gbdb/hs1/censat/censat.bb>). We extracted the type from each record using the “name” column by using the column value up to the first “_” (e.g. “bsat_16_8” becomes “bsat”). Annotation column “REF_CENSAT” was created by intersecting variants with this table and creating a sorted nonredundant list of cen/sat types each variant intersected.

For quality control, filtered regions were obtained by taking all records that were not type “mon” (Monomeric α Sat). We created a non-redundant merge with BEDTools¹²⁸ (v2.31.1), intersecting the remaining (non-mon) records within 50 bp, resulting in 227 Mbp of filtered reference loci. Variants intersecting these regions by 50 bp were excluded from all callsets. We found excessive numbers of variants from these regions, specifically in large α -satellite HORs and Human Satellite III (HSat3) repeats.

Simple repeats

For T2T-CHM13, we obtained the “Simple Repeats” track from the UCSC genome browser for hs1 on 2024-03-13 (<https://hgdownload.soe.ucsc.edu/gbdb/hs1/bbi/simpleRepeat.bb>). For GRCh38, we obtained the “Simple Repeats” track from the UCSC genome browser for hg38 on 2023-11-27 (<http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/simpleRepeat.txt.gz>).

Telomere repeats

For T2T-CHM13, we extracted a set of regions from the simple repeats track (see section “Simple repeats” above) where the consensus sequence was any rotation of “TTAGGG” or its reverse complement and excluded all regions that were not within 10 kbp of either end of the chromosome (142 kbp of reference sequence).

Modeled centromeres

For GRCh38, we obtained the “Centromeres” track from the UCSC Genome Browser for hg38 on 2024-03-21

(<http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/centromeres.txt.gz>).

Merged callset quality filters

Filters applied to all variant types

Any merged variant that was identified in only unphased assemblies (i.e., no h1 or h2 assemblies) was removed from the T2T-CHM13 and GRCh38 callsets.

For T2T-CHM13, variants were dropped that were composed of 50% or more of excluded centromere and satellite sequence (see section “Reference Annotations” above) or that contained 50% or more of telomeric repeats located at chromosome ends (see section “Reference Annotations” above). For GRCh38, variants that were composed of 50% or more of modeled centromeres were dropped (see section “Reference Annotations” above).

Support by callset concordance

To determine if a variant in the merged callset is supported by another caller, we intersect the merge across all individuals (unphased diploid callsets) or haplotypes (phased callsets) with the other callset. We chose a strategy that was designed to mitigate placing too much emphasis on a single individual or haplotype, such as the lead haplotype for a variant in a merge, or that allowed random matches to arbitrarily support a variant call.

For each caller, we intersect the full merged callset with each haplotype (for phased callers) or individual (diploid callset for unphased callers) using the same parameters as described above to perform the intersect. For each variant, we compare the pattern of individuals or haplotypes in the merge with the pattern of individuals or haplotypes from the caller with two metrics: Jaccard similarity (intersection / union) and Fisher’s exact test (FET, two-sided) using SciPy¹⁵¹ (v1.11.4) (`scipy.stats.fisher_exact`). We determine the variant is supported by the caller if the FET p -value is 0.01 or less or if the Jaccard similarity is 0.90 or greater, which mostly rescues FET p -values in the case that the variant is found in all haplotypes or individuals (i.e. $p = 1.0$).

Insertions, deletions, and SNVs

All SNVs, insertions and deletions (including SVs and indels) were accepted into the final callset if they had support from at least three callers. Since SVIM-asm was run using PAV’s alignments, we found that the callsets for all but the largest variants were very similar. For orthogonality, we required that variants be supported by a caller other than PAV and SVIM-asm. Because PAV is able to detect large variants that all other assembly-based callers are unable to detect, we allowed PAV-only calls if they were also supported by PAV with hifiasm assemblies.

Inversions

Inversions were accepted into the final callset if they had support from PAV with hifiasm assemblies or were supported by at least three callers.

Callset growth estimates

Estimates for callset growth by the addition of an individual were performed by averaging the number of new variants each individual would add in a leave-one-out experiment and counting the number of new variants we would obtain from adding the left-out individual. This was simulated using the existing merge all individuals with trio children removed ($n = 62$ individuals) without re-merging. Averages reported in the paper were also computed without trio children and were obtained by running this experiment on all individuals ($n = 63$), all AFR individuals ($n = 29$), and all non-AFR individuals ($n = 33$).

Results

Increases over past callsets

Compared with our previous resource derived from 32 phased human genome assemblies⁸, our current callset yields 1.6× more SV insertions and deletions (177,718 vs. 111,679), which increases to 3.5× for SVs greater than 10 kbp (4,541 vs. 1,299) with improvements in assembly contiguity (**Supplementary Table 27**). Mendelian inheritance error (MIE) has also decreased (**Supplementary Table 28**) with 2.7% MIE for SV insertions and deletions (55% decrease), 5.1% MIE for inversions, 4.9% MIE for indels (5.5% decrease), and 0.6% MIE for SNVs (79% decrease).

Mendelian inheritance error (MIE)

When HG00514 is excluded from MIE analysis, which has a high PacBio HiFi indel error rate compared to other samples (**Supplementary Table 28**), indel MIE falls to 2.1% (-43%), although the indel enrichment was removed during the QC process and is not apparent in per-individual variant calls (**Supplementary Table 29**). When we filter these MIE calls for candidate *de novo* variants and cell line artifacts (Methods), we find 7–40 SV insertions, no inversions, 277–1,690 indels, and 1,027–3,319 SNVs per trio.

Cross-caller support

Using the same merging parameters for callsets, we merged variant calls from multiple sources across one sample, HG00171 (**Supplementary Fig. 9**). We find that ONT-based callers produce results that are often inconsistent with any other call source including other ONT-based callers with Sniffles 2 having the highest false-positive and false-negative on ONT R9 reads. DipCall had the highest false-positive rate of any assembly-based caller. A more complete study would need to be performed to draw solid conclusions for these results across additional samples.

Linking across references

We further linked variants across both references through assembly coordinates (**Supplementary Table 30**) and find SVs identified in both references have lower allele frequencies on average (6.5% insertions, 7.8% deletions) compared to variants present in only T2T-CHM13 (8.3% insertions, 11.1% deletions) or GRCh38 (13.3% insertions, 14.7% deletions).

Mobile elements

Contributing authors: Parithi Balachandran, Christine R. Beck, Jonathan Crabtree, Scott E. Devine, Miriam K. Konkel, Mark Loftus, Ryan Mills, Weichen Zhou

Methods

Mobile element insertions (MEIs) outside of low confidence regions [i.e., MEIs within the Yq12 (see **Main Methods Y chromosome annotation and analysis-for Yq12 MEIs**) and centromeric α -satellite HOR arrays (see **Main Methods Mobile Element Insertions**)] were identified within the 130 haplotype assemblies using two separate pipelines and human references (T2T-CHM13 and GRCh38):

One pipeline utilized a custom calling algorithm, L1ME-AID (v1.0.0-beta) (L1 Mediated Annotation and Insertion Detector) (<https://github.com/Markloftus/L1ME-AID>), that leveraged RepeatMasker⁹¹ (v4.1.6), with the Dfam (v3.8) library⁹², annotation of the PAV merged SV insertion callsets (T2T-CHM13 and GRCh38). PAV SV sequences were designated as MEIs by L1ME-AID if they were annotated by Repeatmasker as containing non-LTR retroposon sequence (*Alu*, L1, SVA) or snRNA, had a poly-A tail, and were semi-low divergence (*Alu*: <6%, L1/SVA: <15%) against the Dfam element consensus sequence. Additionally, elements are selected for younger/active subfamilies (e.g., *AluY*, L1HS, L1PA2, etc.). All candidate sequences were then further selected for minimum mobile element sequence percentage (*Alu*: 70% - to allow for long a-tails, L1/SVA: 20% - to allow for transduced sequence).

The second pipeline called MEIs directly from the alignment of sample contigs to a reference genome with PALMER2 (<https://github.com/WeichenZhou/PALMER>). PALMER2 represents an enhanced iteration of the PALMER algorithm, engineered to directly call non-reference TE families from aligned, assembled contigs (with PALMER --input `{INPUT_hap.bam}` --workdir `{path_to_workingspace}` --output `{output_prefix}` --ref_ver `{reference_version}` --ref_fa `{reference.fa}` --type `{TE_type}` --chr `{chr}` --mode asm). This updated version is crafted to bypass the "masking" phase commonly required in the original workflow^{152,153}, thereby expediting the process by extracting insertion signals straight from the ultra-long contig sequences. By inheriting the comprehensive feature set of its predecessor, PALMER2 retains the capability of meticulously resolving the structure of TE insertions, characterizing signature retrotransposition hallmarks, and identifying customized non-reference insertion sequences. PALMER2 incorporates a merging utility to streamline the consolidation of individual sample callsets into an integrated, unified callset from diverse genomes.

MEIs identified by either caller were then integrated by comparing the MEI coordinate (chromosome and position), element family (*Alu*, L1, SVA), and sequence composition. If PALMER2 called an MEI within 1 kbp of a PAV called MEI site, and the two sequences were identified as belonging to the same mobile element family, a pairwise global alignment of their sequences would be performed using Biopython¹⁵⁴ (v1.82; Bio.pairwise2 module). After alignment, the hamming distance (i.e., the proportion of disagreeing components) between the two aligned sequences would be calculated using SciPy (v1.11.1; `scipy.spatial.distance.hamming`)¹⁵¹. If the sequence alignment hamming distance was <0.1, the two MEI calls would be considered identical and merged.

Following merging, the combined MEI callsets underwent additional refinement/quality control by checking if each MEI was a potential deletion from the human reference genome (T2T-CHM13 or GRCh38 depending on callset), a duplication, or a potential artifact from errors

in genome assembly. For each MEI 500 bases of flanking sequence (500 up-stream and 500 down-stream) of the insertion site within the human reference was retrieved and given to BLAT¹⁵⁵ (Standalone BLAT v36x2, -minScore=850) to query for homologous sites within the human reference, and within an ancestral reference genome (Chimpanzee (NHGRI mPanTro3-v2.0 pri/GCA_028858775.2). All resulting BLAT hits containing a gap +/-20% difference in size compared to the query MEI insertion call sequence length were further investigated. RepeatMasker⁹¹ (v4.1.6), using the Dfam (v3.8) library⁹², was utilized to annotate any elements within the homologous BLAT hits. MEI calls were filtered from the final merged dataset if any secondary sites within the human and/or chimpanzee reference genome were discovered that contain the same mobile element (based on family, orientation, and percent divergence). Additionally, all MEIs within low confidence regions (e.g., centromeres, dense tandem repeat loci; T2T-CHM13: <https://hgdownload.soe.ucsc.edu/gbdb/hs1/censat/censat.bb> (see **Centromere and satellite annotation for T2T-CHM13** above for more information), GRCh38:

https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/technical/filter/20210127_LowConfidenceFilter/LowConfidenceFilter.bed.gz) as well as identical MEIs found across unique locations (e.g., a single MEI called twice due to being found in different copies of a segmental duplication across individuals) were dropped. Finally, any MEIs still remaining and called by a single pipeline were manually curated by double checking elements for classical features of transposon insertions i.e., tails, target site duplications. An F-test was performed (statistic=1.62590, p-value=0.0868) to test for equal variances between superpopulation total MEI counts (AFR/non-AFR total MEIs from T2T-CHM13 callset). Since the F-test p-value was ≥ 0.01 (i.e., variances were equal), and the Shapiro-Wilk test p-values for each population were non-significant (AFR P-value=0.876, statistic=0.982; non-AFR p-value=0.601, statistic=0.9752; i.e., the data was normally distributed), a Student's t-test was performed to test for a significant difference between African and non-African genome MEI counts (mean=1947 (AFR) and 1408 (non-AFR), statistic=49.187, p-value=5.580e-52, degrees of freedom=63.0). All statistical tests were two-tailed.

Callsets were then compared against an orthogonal MEI callset produced by MELT-LRA (see following section, **Supplementary Tables 33, 34**).

Mobile Element Locator Tool-Long Read Assembly (MELT-LRA). MELT-LRA (<https://github.com/Scott-Devine/MELT-LRA>), uses PAV calls derived from whole genome assemblies to identify insertions ranging from 50 bp to 20 kbp that are caused by MEIs. Insertions in the 50 bp to 20 kbp range are first identified from the raw PAV calls, and these insertions are then individually compared to consensus sequences for *Alu*, L1, and SVA elements (the three currently active mobile elements in humans) using Smith Waterman alignments between the PAV calls and MEI consensus sequences. Once candidate MEIs are identified from these alignments, they are further annotated using previously developed MELT tools that have been extensively validated with short-read whole genome sequencing and PCR^{28,156}. Features annotated include flanking target site duplications (TSDs), poly (A) tails, MEI subfamilies, ORFs (for full-length L1 elements), and genotypes. The MELT-LRA package also includes new visualization tools developed to facilitate this analysis. A total of 12362 MEIs were identified from PAV calls generated through GRCh38 alignments and 12434 MEIs from T2T (CHM13) alignments. This includes: 10352 and 10428 *Alu* insertions from the GRCh38 and T2T alignments, respectively; 1381 and 1367 L1 insertions from the GRCh38 and T2T alignments,

respectively; and 629 and 639 SVA insertions from the GRCh38 and T2T alignments, respectively. All but 160 of these MEIs were found in at least one other MEI callset generated with the three independent MEI pipelines (see “Results” below). These 160 were removed from the final callset.

Results

Within the resulting merged and refined T2T-CHM13 and GRCh38 MEI callsets there were an average of 12,709 unique MEI calls (T2T-CHM13: $n=12919$; GRCh38: $n=12642$) with one or more caller support [average of 11,886 unique MEI calls with the support of both callers (T2T-CHM13: $n=11,933$ (92.4%); GRCh38: $n=11,838$ (93.6%))] (**Supplementary Figs. 10, 11, Supplementary Tables 31, 32**). Additionally, our comparison with an orthogonal callset showed a high average concordance of 92.3% (T2T-CHM13: 92.1%, GRCh38: 92.5%; **Supplementary Tables 33, 34, Methods**). The T2T-CHM13 callset (*Alu*: 10,458, L1: 1,664, SVA: 791, HERV-K: 5, snRNA: 1) and GRCh38 dataset (*Alu*: 10,270, L1: 1,604, SVA: 764, HERV-K: 3, snRNA: 1) had similar MEI compositions. The callsets containing only MEIs with the support of two or more callers showed a comparable amount of element family representation (T2T-CHM13: *Alu*: 9,909, L1: 1,409, SVA: 615; GRCh38: *Alu*: 9,855, L1: 1,378, SVA: 605). Next, we screened the PAV deletion callsets and identified 2,450 polymorphic MEIs present in T2T-CHM13 and 2,468 MEIs in GRCh38 (**Supplementary Fig. 12, Supplementary Tables 35, 36**). As expected, we identify a significantly greater number of MEIs within genomes of African descent (mean=1,947 (AFR) and 1,408 (non-AFR), statistic=49.187, p-value=5.580e-52, degrees of freedom=63.0, Student's T-test, two-tailed, T2T-CHM13 reference; **Supplementary Fig. 12**).

Across the 1,246 centromeres analyzed, 383 (30.74%) were found to contain at least one MEI in the active α -satellite HOR array. In total, we identified 806 mobile elements, although most of these elements (88.95%) were the same element/duplication of a single MEI event across the haplotypes (89 unique insertion events; 52 L1, 36 *Alu* element, and 1 SVA insertion). We then selected two unique chromosomes to analyze further (chromosomes 2 and 20). The chromosome 2 centromere was completely and accurately assembled in 70 haplotypes. Within the α -satellite HOR array of these 70 centromeres, 56 contained at least one MEI. There were three unique L1 insertions (all L1HS) as well as three unique *Alu* element insertions (two *AluYb8* and one *AluYa5*) identified. The chromosome 20 centromere was completely and accurately assembled in 42 haplotypes, 22 which contained at least one MEI in the α -satellite HOR array. Across these 22 α -satellite HOR arrays, there were 3 unique insertions identified: two L1 insertions (both L1HS) and 1 *Alu* (*AluY*) element insertion.

Inversions

Contributing authors: Hufsah Ashraf, Jan O. Korb, Tobias Marschall, David Porubsky, Tobias Rausch, Vasiliki Tsapalou

Methods

Genotyping

We performed cell selection prior to genotyping using ASHLEYS, an automated Strand-seq cell quality assessment approach⁹⁶. Strand-seq based regenotyping was performed using ArbiGent, a Bayesian probability framework based regenotyper that determines inversion genotype likelihoods for inversions and copy number changes, using Strand-seq reads as an input¹⁶. ArbiGent is available as a module in MosaiCatcher v2¹⁵⁷ and was executed using default parameters (`--config arbigent=True`).

Results

For T2T-CHM13 based PAV callset, ArbiGent was able to provide an assessment for 246 out of 276 inversion calls, while marking 218 as inversions, 25 as inverted duplications and 2 as potential misorientations (**Supplementary Fig. 14, Supplementary Table 37**). To assess the genotyping performance, we tested the genotypes for Hardy-Weinberg equilibrium. 165 confidently genotyped inversions belonging to autosomal chromosomes with no missing genotype were tested and 156 (~95%) of them were found to be in Hardy-Weinberg equilibrium. For GRCh38 based PAV callset, 262 out of 298 inversions were confidently assessed by ArbiGent, marking 202 as inversions, 32 as inverted duplications, and 18 as potential misorientations (**Supplementary Table 38**). 141 confidently genotyped inversions belonging to autosomal chromosomes with no missing genotype were tested for Hardy-Weinberg equilibrium, 127 (~90%) of which passed. Additionally, we re-genotyped previously developed T2T-CHM13 and GRCh38 based inversion callsets that have been based on a subset of samples^{16,98} (**Supplementary Tables 60, 61**). ArbiGent genotypes showed a Hardy-Weinberg equilibrium statistic of 94% and 91% for balanced, confidently genotyped inversions reported in the T2T-CHM13 and GRCh38 callsets, respectively.

To validate candidate inversions in the PAV inversion callset, we applied NAHRwhals⁹⁷, an R package designed for the visualization of complex, NAHR-driven rearrangements in genome assemblies. NAHRwhals enables pairwise comparisons between an assembly and a reference, facilitating the detection of large structural variants (few kbp to multiple Mbp). For each candidate inversion region, we analyzed one carrier sample, generating dotplots against the T2T-CHM13 reference genome. We visually examined each dotplot, assessing rearrangement patterns to determine the presence of an inversion and classify it into a distinct inversion class based on its structural features (**Supplementary Fig. 13**). To ensure robust inversion characterization, we applied a ≥ 5 kbp size threshold, allowing for both False Detection Rate (FDR) estimation and classification of true positive inversions.

195 inversions were confirmed as true positives, resulting in an FDR of 12.33%. The most prevalent inversion class was homology-mediated inversions, comprising 67.3% of the identified inversions (132 instances). Simple inversions followed, accounting for approximately 14.8% (29 instances). More complex cases, -such as homology-mediated inversions accompanied by deletions-, constituted ~13% of the true positive inversions (26 instances). Inversions with deletions made up 5% (10 instances). The least abundant category was inversions paired with

two deletions, which were observed in only 2 cases (1%). Finally, false positives were detected in 28 instances.

Segmental duplications

Contributing authors: Mark J.P. Chaisson, Evan E. Eichler, Keon Rabbani, DongAhn Yoo

Methods

Annotations

Segmental duplications (SDs) were identified using SEDEF⁹⁹ (v1.1) with the default parameters, on soft-masked genome assemblies, employing TRF¹⁰⁰ v.4.1.0 (*"trf [asm.fa] 2 7 7 80 10 50 2000 -l 30 -h -ngs"*), RepeatMasker¹⁰¹ v.4.1.5 (*"RepeatMasker -s -e ncbi -xsmall -species human [asm.fa]"*), and Windowmasker¹⁰² (v2.2.22) (*windowmasker -mk_counts -mem 16384 -smem 2048 -infmt fasta -sformat obinary -in [asm.fa] -out [asm.count] &&*

windowmasker -infmt fasta -ustat [asm.count] -dust T -outfmt interval -in [asm.fa] -out [asm.interval]); the softmasking was performed using BEDTools-merged regions of the overall repeatmasking results, via seqtk (v1.3) (*"seq -l 50 -M"*). The putative false SDs were filtered out via intersecting with erroneous regions (lower or higher read depth) from NucFreq or Flagger. Additionally, the highly confident segmental duplication calls validated by fastCN previously generated using T2T-CHM13v1.1 (excluding chrY) genome were used for the consequent comparative analyses. Accumulation curve of the SDs was generated based on the number of bases of SDs projected onto T2T-CHM13 reference space; note that the data excludes NA19650, NA19434 and NA21487 with significantly different SD content. Homozygous and heterozygous genotypes among SDs were determined by comparing SD calls in the haplotype assemblies.

Duplicated genes

Gene bodies identified from the transcript alignments (*"minimap2 -x slice -a"*) were remapped back to each assembly (*"minimap2 -x asm20 -p 0.2 -N 100 -m 10 -E2,0 -s 10"*), filtering out alignments containing less than 95% of the coding sequences, or that had greater than 5% divergence. Because there are multiple overlapping alignments, partially mapped reads, and different isoforms, the alignments were grouped by exon overlap using Leiden community detection to avoid marking the same region in the same haplotype as multiple separate duplications (**Supplementary Fig. 66**). Final gene annotations were filtered out if they were marked as erroneous by Flagger¹. To summarize duplications across all genomes, we took the union of the isoform community generated for each haplotype, and considered the isoform with the most alignments as the 'representative' isoform. Duplications were defined as separate alignments of the same community. This is implemented in a pipeline called SegDupAnnotation2 (SDA2) (<https://github.com/ChaissonLab/SegDupAnnotation2>).

Comparison of SDA2 annotations versus HPRC annotations of gene duplications

In the draft human pangenome of 94 haplotypes¹, a combination of liftoff and an earlier version of SDA2 annotations were used to detect multicopy genes. We compared duplicated genes detected by SDA2 versus HPRC to annotate discrepancies. Of the 1,115 genes annotated as duplicated in HPRC, we find that 31 are not listed as protein coding in our database, 104 have one of the copies annotated in flagger regions, 412 have different annotations of gene names from multicopy-gene families. The remaining discrepancies are largely due to community detection in SDA2 and different thresholds of completeness of the gene model.

Comparison of gene duplications using SEDEF versus SDA2 annotations.

The genes in novel segmental duplications annotated by SEDEF should largely be duplicated genes. We compared the gene lists from SDA2 annotations to those in novel SDs. We considered 900 full genes in novel SDs which reflect a superset of the final list of genes because they are not filtered by fastCN^{21,103}. Of these 244 or 27% were excluded by the SDA2 identity filter (<90% identity), 55 or 6% dropped out by length, 85 or 9% dropped out by low gene model alignment (<50% alignment). The remaining 189 genes were not considered by the SDA2 pipeline. Among these, 116 were single-exon genes, 132 were shorter than 5 kbp, and 77 were missing from the original set of genes used to annotate assemblies using SDA2.

Results

Examining genomes by superpopulation, the number of new segmental duplications added per individual is highest in AFR of 3.97 Mbp/individual, compared to other populations, 2.81, 3.37, 2.76, 2.57 Mbp/individual in EUR, AMR, EAS, SAS, respectively. This is consistent with the greater intrachromosomal SD content in AFR than other populations (**Supplementary Fig. 21**). Note that we find no evidence that the intrachromosomal SDs are more divergent in AFR compared to non-AFR ($p=0.32$ and 0.065 for female and male individuals, respectively; **Supplementary Fig. 22**).

In addition to the segmental duplications that are found novel by their location/synteny, we also examined the sequence content of the syntenic ones, by directly aligning the sequences (**Extended Data Fig. 2a**). Among the segmental duplications that are homologous by position, we find that 14.6 Mbp show less than 80% of sequence identity compared to T2T genome, 4.4 Mbp that are at least two-fold expanded length in the HGSVC3 genomes compared with T2T, and 29.0 Mbp with sequence changed (<80% identity) at the same time expanded (>2-fold), suggesting at least 48 Mbp of the segmental duplications that share positional homology, are not necessarily conserved (**Extended Data Fig. 2a,b**). Quantifying the number of genes that are located in the novel or more variable segmental duplications, we find on average 26 complete overlap of protein-coding genes per chromosome (598 genome-wide) and 5 partial overlap (at least 50% coverage) per chromosome (118 genome-wide).

Compared to the previously known human segmental duplications across 170 haplotypes²¹, we observed 23.6 Mbp of new SDs (including 7.3 Mbp that are found in more than one individual, and 16.3 Mbp of that are singleton). This represents complete overlap with 124 protein-coding genes, and partial overlap with 43 protein-coding genes, genome-wide (**Supplementary Fig. 20**).

We separately used an alignment-based pipeline to annotate full-length multi-copy genes in each assembly using multi-mapped alignments of GENCODE genes (release 44). The analysis was limited to multi-exon genes at least 5 kbp to distinguish from processed pseudogenes.

STR and VNTR annotation

Contributing authors: Bida Gu, Mark J.P. Chaisson

Methods

Annotation of tandem repeats

We used *vamos*¹⁵⁸ (v1.3.2) and the corresponding motif catalog¹⁵⁹ (v2.1) to annotate TRs in 94 HPRC and 130 HGSVC3 haplotype assemblies. A unified diploid VCF file was generated from these annotations, encompassing 112 genomes. For each TR locus, distinct alleles were represented in the "ALTANNO" field as strings of annotated motifs. Diploid genotypes were denoted by allele indices as ordered in the "ALTANNO" field. Homopolymers were also annotated, but given their limited capacity to reflect complex TR patterns, we excluded them from most analyses to focus on more informative TR structures.

Results

TR Allele Diversity

To assess the variability of TRs, we calculated the average number of alleles per annotated haplotype for TR loci with more than 20 annotated haplotypes, using two measurement methods: annotation length and motif composition (**Supplementary Fig. 67a**). As anticipated, locus variability decreased when alleles were measured solely by length compared to motif composition. Despite these differences, the majority of TRs were found to be constant across genomes, suggesting that most TRs are evolutionarily highly conserved. Interestingly, although the HGSVC3 genomes were derived from a more balanced population selection and displayed a similar TR compositional variability to the HPRC genomes, they showed reduced length variability. Our analysis identified 114,817 loci with decreased length variability in the HGSVC3 genomes, with 3,659 (3.2%) located in coding regions. Given that coding TRs account for 5.0% of the total TRs, this suggests that the reduced TR length variability observed in the HGSVC3 genomes is predominantly enriched in noncoding regions. Because the population diversity is high in both datasets, this likely reflects an indication of higher quality assembly. Principal Component Analysis (PCA) and genome wide statistical tests were performed as described previously¹⁵⁹, revealing clear clustering of major populations (**Supplementary Fig. 67b**) and 9,297 population informative loci (corrected p-value < 0.05; $p < 10^{-8}$) among the 112 samples.

Intersection of tandem repeats in HGSVC2, HGSVC3, and HPRC genomes

Since the motif database was built using HPRC genomes, unique TRs from the HGSVC3 genomes might not be detected within the original framework. To address this gap, we compared the raw outputs from the Tandem Repeats Finder with only centromere regions masked for the HPRC, HGSVC2, and HGSVC3 genomes. As shown in **Supplementary Fig. 67c**, the majority of loci (732,035, 89.0%) were found to be shared among all three datasets, while unique TRs in each dataset largely reflected population-specific differences. Given the relatively higher quality of the HPRC and HGSVC3 assemblies, we conducted a separate comparison between these two datasets. This comparison revealed 30,030 unique TR loci in the HGSVC3 genomes (as shown in **Supplementary Fig. 68**), whereas 21,516 unique TR loci were identified in the HPRC genomes. Importantly, none of these unique loci were found in regions completely unsequenced in the other dataset, indicating comprehensive coverage of both datasets on the CHM13 reference genome. This finding demonstrates the consistency between the two datasets and underscores the robustness of our TR identification approach. As expected, the unique TR loci were primarily concentrated in ribosomal DNA (rDNA) regions, where the CHM13 reference genome is represented by a perfectly repeated model sequence. This outcome suggests that areas with less robust reference data, such as rDNA, tend to contain unique TR loci, emphasizing the need for continued refinement of reference genomes for more accurate TR analysis.

Comparing tandem repeats calls to SV calls

We annotated tandem repeat motif composition for 1,243,719 TR loci on GRCh38, and compared against the integrated SV callset. Among these, 46,314 showed a discrepancy of at least 50 bases between any allele and the reference, indicating a net structural variant. The majority of structural variant calls are accounted for by tandem repeat loci. Of the 176,438 SV loci in the combined variant callset, 113,412 overlap TRs (TR-SV), defined by either insertion variants that are inside of a tandem repeat ($n=76,094$), or are deletion loci contained entirely within a TR locus ($n=37,318$). The TR-SV are contained within 31,418 TR loci, indicating clustering of SV calls in TR sequences. Nearly all (31327/31418) TR-SV overlap multiallelic TRs, and roughly 50% (15,944/31,418) have at least 10% of genomes with a different length allele.

Y chromosome

Contributing authors: Peter Ebert, Pille Hallast, Miriam K. Konkel, Mark Loftus, Charles Lee

Methods

Construction and dating of Y phylogeny

All sites were called from the Illumina high-coverage data²⁹ using the ~10.4 Mbp of Y-chromosomal sequence previously defined as accessible to short-read sequencing¹⁶⁰. BCFtools^{93,161} (v1.16) was used with minimum base quality 20, mapping quality 20, and ploidy 1. SNVs within 5 bp of an indel call (SnpGap) and indels were removed, followed by filtering all calls for a minimum read depth of 3 and a requirement of ≥85% of reads covering the position to

support the called genotype, followed by removal of sites with $\geq 5\%$ of missing calls (that is, missing in more than 2 out of 32 samples), were removed using VCFtools¹⁶² (v.0.1.16). After filtering, a total of 10,405,284 sites remained, including 11,129 variant sites.

The Y haplogroups of each individual were predicted as previously described²² and correspond to the International Society of Genetic Genealogy nomenclature (ISOGG, <https://isogg.org>, v.15.73). A coalescence-based method implemented in BEAST¹⁶³ (v.1.10.4) was used to estimate the ages of internal nodes. RAXML¹⁶⁴ (v.8.2.10) with the GTRGAMMA substitution model was used to construct a starting maximum-likelihood phylogenetic tree for BEAST. Markov chain Monte Carlo samples were based on 200 million iterations, logging every 1,000 iterations, with the first 10% of iterations discarded as burn-in. A constant-sized coalescent tree prior, the GTR substitution model, accounting for site heterogeneity (gamma) and a strict clock with a substitution rate of 0.76×10^{-9} (95% CI = $0.67 \times 10^{-9} - 0.86 \times 10^{-9}$) single-nucleotide mutations per bp per year was used¹⁶⁵. A prior with a normal distribution based on the 95% CI of the substitution rate was applied. A summary tree was produced using Tree-Annotator (v.1.10.4) and visualized using the FigTree software (v.1.4.4, **Supplementary Fig. 25**). For figures showing the Y phylogeny, subsets of relevant individuals were extracted and visualized.

SVs disrupting genes

Contributing authors: Miriam K. Konkel, Mark Loftus, Gianni V. Martino, Mike E. Talkowski, Xuefang Zhao

Methods

Annotate the genes overlapped by long-read SVs

We used bedtools to annotate the overlap between long-read SVs and coding exons, introns, and UTRs as defined in Gencode V45. SVs were classified as “coding disruptive” if they included deletions or insertions overlapping any coding exons, or inversions with at least one breakpoint within a gene. SVs overlapping introns or UTRs without affecting any exons were categorized as “intronic” or “UTR disruptive,” respectively.

Comparison of short-read and long-read SVs that affect genes

Contributing authors: Xuefang Zhao, Michael E. Talkowski

We benchmarked long-read SVs against SVs generated from matched short-read WGS data. Two SVs are considered to be the same event if they meet the following criteria:

1. Deletions and Duplications that are under 5 kbp should have the same SV types, and share a reciprocal overlap of 10% or more.
2. Deletions and Duplications that are over 5 kbp should have the same SV types, and share a reciprocal overlap of 50% or more.

3. The breakpoints of insertions must be within 100 bp of each other.
4. Insertions can match duplications if the insertion point is within 100 bp of the first breakpoint of the duplication.

Results

We identified 1,535 SVs, including 938 deletions, 80 inversions, 504 insertions, and 13 MEIs, that altered the coding sequences of 985 unique genes. Additionally, 73,489 SVs were found to impact intronic sequences of 9,562 genes, while 1,518 SVs disrupted the untranslated regions (UTRs) of 906 genes in our dataset.

Out of the 65 individuals, 63 had matched short-read SV calls from the 1000 Genomes Project (1kGP). Of the 176,531 SV sites identified from long-read data, 98.9% ($n=174,641$) are present in at least one of these 63 individuals. Consistent with our previous study, long-read sequencing demonstrates superior sensitivity in detecting SVs within highly repetitive regions, such as segmental duplications and simple repeats. Specifically, 82.4% of deletions and 92.1% of insertions in these regions from long-read data do not overlap with short-read SVs. In contrast, a more substantial overlap between the two technologies is observed in the remaining 90% of the genome, which is less repetitive. Here, 11.4% of deletions and 32.1% of deletions and insertions are not overlapped by short-read SVs.

In the 63 individuals, an average of 10,918 deletions and 19,541 insertions were observed per individual genome. Consistent with the site-level results, there is a notable overlap between sequencing technologies for SVs in less repetitive regions. Specifically, 89.4% of deletions and 64.5% of insertions in these less repetitive sequences were detected by short-read SV calls in the matched genomes. In contrast, 19.7% of deletions and 9.0% of insertions in the more repetitive regions were detected by short-read SVs.

Coding exons from 818 unique genes were found to overlap with 1,300 unique SV sites across the 63 individuals. For long-read SVs that overlap coding sequences and are located outside highly repetitive genomic regions, 84% of the deletions and 63% of the insertions are also detected by short-read SVs. In contrast, only 25.1% of these deletions and 7.3% of these insertions in repetitive regions are overlapped by short-read SVs.

By contrast, from short-read sequencing data we find 206 genes altered in these 65 individuals by SV not captured in the long-read assemblies, most of which (83.2%) were disrupted by large CNVs. We leveraged VaPoR to evaluate these SVs by directly comparing raw PacBio alignments against the GRCh38 reference genome. Among the SV sites that are accessible by VaPoR, 27.8% have support from at least one PacBio alignment, indicating potential false negatives in the long-read assembly methods. Interestingly, manual review of these variants in one individual, HG00171 (11 SVs), indicates 1 (8%) was a true deletion between reference assembly gaps that missed by PAV, 2 (17%) are involved in complex SVs that were not fully resolved, 4 (33%) were in contiguously assembled loci with no sign of the deletion, 1 (8%) was present but shifted between the callsets and did not intersect, 2 (17%) were overlapping at the

same locus and likely represent two distinct events that may have been lost to merging in the long-read callset, and 1 (8%) was in an uncallable locus.

Phased transcripts, isoforms, and effects of SVs

Contributing authors: Miriam K. Konkel, Mark Loftus, Gianni V. Martino

Methods

RNA-seq pre-processing

Initial quality control of the Illumina RNA-seq reads for the 12 individuals with corresponding Iso-Seq data was performed using Trimmomatic¹⁶⁶ (v0.39). The trimmed paired-end reads were then aligned to the indexed T2T-CHM13 (NCBI RefSeq GCF_009914755.1-RS_2023_03) using STAR¹⁶⁷ (v2.7.10b). Splice-junction results from all individuals were compiled and provided to STAR for a realignment to optimally identify transcript diversity. Cufflinks¹⁶⁸ (v2.2.1) was then utilized to annotate and quantify the RNA-seq expression data

Phasing Iso-Seq

For preparation of the 12 PacBio Iso-Seq datasets, primers and poly(A) tails were first removed using Lima (v2.1.0; <https://github.com/pacificbiosciences/barcoding/>) and isoseq3 refine (<https://github.com/PacificBiosciences/IsoSeq>), respectively. Iso-Seq reads were then aligned to both haplotype assemblies of the matching individual using pbmm2 (v1.5.0; <https://github.com/PacificBiosciences/pbmm2>). Next, reads were phased through comparison of alignment quality (CIGAR [compact idiosyncratic gapped alignment report] approach) and sequence similarity (*k*-mer approach). Sequence matches and variations (i.e., mismatches and indels) were isometrically rewarded and penalized, respectively, utilizing the CIGAR strings produced by read alignment to each haplotype. The CIGAR-based approach subsequently phased reads to the haplotype assembly that produced the alignment with the greater score. The *k*-mer approach compared the 15-mers present in each read to the 15-mers in the corresponding aligned sequence from both haplotypes. The best alignment was identified as the haplotype that shared the most unique *k*-mers with the respective read. Final read-phasing designations utilized an ensemble approach where reads were phased to a haplotype if both phasing approaches were in agreement. Disagreement between approaches was settled through determination of which alignment produced a greater quantity of matches. In the absence of measurable differences, according to their scoring premises, an “UnableToBePhased” designation was assigned.

Phased Iso-Seq annotation

Only Iso-Seq with a mapping quality $\geq 99\%$ were utilized for downstream analyses. These reads were first divided into haplotype-specific bins based on their phasing designations (i.e., Hap1, Hap2, or Unphasable). Next, reads within each bin were clustered using isoseq3 (v3.8.2; <https://github.com/PacificBiosciences/IsoSeq>) cluster, which removes singleton reads and combines similar reads to form transcripts. These transcripts were aligned to the T2T-CHM13

(NCBI RefSeq GCF_009914755.1-RS_2023_03) reference genome using pbmm2 (v1.5.0; <https://github.com/PacificBiosciences/pbmm2>). Finally, isoforms were generated by running isoseq3 collapse on the mapped transcripts. Isoforms were then annotated and filtered using SQANTI3¹¹¹ (v5.1.2) QC and rule-based filter. For each individual, the isoforms formed from the separate haplotype phasing bins were merged, and unique identifiers for each isoform were generated using individual-agnostic characteristics to facilitate cross-individual tracking of similar isoforms.

Short- and long-read comparison

We selected all isoforms shared across at least two out of twelve individuals present in the PacBio Iso-Seq and/or Illumina RNA-seq, identified all unique protein-coding and noncoding genes represented in either dataset, and compared the gene distribution between both datasets. To achieve this, we filtered the isoforms assembled from Illumina RNA-seq short reads (Cufflinks v2.2.1)¹⁶⁸ retaining only isoforms with non-zero abundance (FPKM). Unique isoforms from the short-read RNA-seq were then identified based upon isoform characteristics (e.g., chromosome, locus, isoform length, gene annotation, and biotype) per individual, as well as across the whole 12 individual set. For Iso-Seq, we used the filtered isoforms generated by the isoseq3 (v3.8.2) pipeline (see **Methods: Phased PacBio Iso-Seq Work-up**), which removed singleton reads. Furthermore, isoforms predicted to undergo nonsense-mediated decay by SQANTI3¹¹¹ (v5.1.2) were also excluded. Isoforms were given unique identifiers based upon individual-agnostic features (gene annotation, splice sites, isoform length, etc.) to compare the total unique genes across all 12 individuals. Lastly, we compared the average number of unique isoforms per unique gene represented for each individual for both sequencing approaches.

Results

Isoform phasing

The availability of long-read Iso-Seq data and haplotype-resolved individual-specific genome assemblies allows the detailed investigation of allele-specific transcription. To ensure haplotype-resolved read alignment quality, we required agreement between a CIGAR and a *k*-mer-based approach (see Methods section “Phasing Iso-Seq” above). Congruence between both approaches was observed 98.3% of the time across the 12 individuals. On average, 58.44% of the reads could be phased (read phasing range 43.98-71.98%) for a given individual (**Supplementary Fig. 69**), permitting the determination of allele-specific expression (**Supplementary Fig. 70**).

Imprinted loci

Our allele-specific Iso-Seq analysis uncovered ten genes whose isoforms were transcribed from a single haplotype. As imprinted genes show a similar pattern, we reviewed the literature, which confirmed that the ten genes (*ZDBF2*, *NAP1L5*, *FAM50B*, *PEG10*, *IPW*, *MKRN3*, *SNRPN*, *PEG3*, *L3MBTL1*, *LPAR6*) are paternally imprinted in EBV-transformed B-lymphocyte cell lines^{169–171} (**Supplementary Fig. 70**). Furthermore in concordance with the literature, isoforms of the *IPW*, *MKRN3*, and *SNRPN* genes, which are located in the Prader-Willi Syndrome imprinting control region¹⁷², displayed monoallelic expression from the same haplotype within a given

individual. The detection of known imprinted genes through phasing allows distinction of paternal from maternal chromosomes. This information can be utilized to determine the origin of a genetic variant in a proband if a parental genome is unavailable. Furthermore, the full agreement of our findings with published imprinted genes provides strong evidence for both our phasing approach and haplotype assembly accuracy.

Comparison of alignments to phased assemblies and reference

While a reference genome serves as a useful baseline for transcriptomic data comparison, the lack of sample specificity can introduce errors into read alignments when the reference genome is discordant from a transcript. Therefore, we quantified the improvement in alignment afforded by mapping long-read Iso-Seq data (12 samples from EBV-transformed B-lymphocytes) to phased individual-specific assemblies versus the T2T-CHM13 reference genome. Out of 26,533,152 total reads, 26,352,868 reads aligned to the personal assemblies and 26,367,272 aligned to the T2T reference. Following comparison of read alignment qualities between the genomes, we observed a set of 105,317 reads that aligned poorly to the assemblies (<50% accuracy) but well to the T2T reference. Interestingly, we found that 93,408 (89.0%) of these poorly aligned reads belonged to just three samples and 99,043 (94.2%) mapped to immunoglobulin genes. These results indicate that most of these poor read alignments can be attributed to somatic rearrangements of immunoglobulin genes in EBV-transformed B-lymphocytes. One possible explanation for the poorly aligned reads being overrepresented in three samples could be that these samples are less clonal than others, leading to increased difficulty for accurate alignments. We further scrutinized the remaining poorly aligned reads (0.02% of total reads) checking for possible misassembly at the read alignment coordinates in the sample-specific assemblies. In total we identified 15 gene loci containing misassembly (4 premature contig ends, 11 incompletely assembled) and 37 loci within 5 kbp of an unassembled region.

To reduce sequencing artifacts, we identified and filtered singleton reads that failed to cluster into transcripts. After removal of singleton reads, we found an equal number of reads (22,432,063) aligned to both the T2T reference and the phased assemblies. However, the phased reads aligned to the assemblies resulted in 21,682,216 reads with at least 99% alignment quality while alignment to T2T produced 21,470,225 reads at the same threshold. This improved alignment of 211,991 reads afforded by phase-specific alignment of Iso-Seq reads to sample-specific haplotype-resolved assemblies marks a 22% reduction in the number of suboptimal alignments (<99% alignment accuracy) resulting from mapping to the T2T-CHM13 reference genome.

Comparison of short- and long-read RNA-seq

Long-read RNA sequencing, which generates reads that span the full length of transcripts, captures a wider breadth of gene isoforms compared to short-read sequencing methods¹⁷³. However, short-read RNA-seq allows for the quantification of transcript abundances. To better gauge the differences in our expression data, we compared isoform profiles produced from Illumina RNA-seq versus PacBio Iso-Seq across twelve individuals. The comparison of total unique genes expressed in at least two individuals revealed that short read RNA seq identified a greater total number of genes (29,858 unique protein coding/noncoding genes) compared to

long-read Iso-Seq (11,134 unique protein coding/noncoding genes; see **Supplementary Fig. 71**). The lower retrieval of unique genes by Iso-Seq is indicative of incomplete saturation of the transcriptome (mean of 2,211,096 Iso-Seq reads per individual). Additionally, RNA-seq identified on average more unique isoforms (51,413 isoforms) on a per individual basis than the Iso-Seq (33,315 isoforms). However, a greater number of unique isoforms across all twelve individuals was identified by Iso-Seq (68,097 unique isoforms) than the short-read RNA-seq (63,028 unique isoforms; **Supplementary Fig. 71**). This disparity can be explained by the Iso-Seq identifying an average of 7,324 unique isoforms per individual while the RNA-seq only identified an average of 523 unique isoforms per individual. Furthermore, the overall average number of unique isoforms per unique gene identified by Iso-Seq across the 12 individuals was 6.72, while for short-read RNA-seq identified an overall average of 1.99 unique isoforms per unique genes (**Supplementary Fig. 72**). Despite capturing fewer genes Iso-Seq was able to resolve in excess of three times more isoforms for those genes than the Illumina short-read approach.

Functional effects of SVs

Contributing authors: Marc Jan Bonder, Mark Gerstein, Matthew Jensen, Yunzhe Jiang, Miriam K. Konkel, Jiaqi Li, Chong Li, Mark Loftus, Gianni V. Martino, Xinghua Shi

Methods

Enrichment of SVs in genomic regions

Genes and transcripts of the human reference genome GRCh38 assembly were derived from the GENCODE¹⁰⁶ annotation (v45). We further assessed four classes of ENCODE v3 cis-regulatory elements (cCREs), including promoters (<200 bp from gene TSS), proximal (between 200-2000bp from gene TSS) and distal enhancers (>2000 bp from gene TSS), and predicted elements bound by CTCF¹¹⁷. Insertions and deletions were intersected with genes and transcripts using BEDTools¹²⁸ (v2.30.0) intersect with a minimum overlap of 1 bp. We also extracted exons and introns for long noncoding RNA (lncRNA) and pseudogene transcripts, as well as coding sequences (CDSs), exons, introns, start codons, stop codons, and untranslated regions (UTRs) for protein-coding transcripts. For a given gene, we counted the insertions and deletions within different genomic elements (such as CDSs, exons, and introns) and normalized the counts by the length of the genomic element. We also calculated the proportion of these genomic elements for different transcript types that were affected by insertions or deletions.

We conducted a permutation test to assess the significance of the depletion or enrichment of SVs for these genomic elements. To achieve this, we generated a catalog of shuffled SVs using BEDTools¹²⁸ (v2.30.0) shuffle. Given the heterogeneity of chromosomes, we ensured that the shuffled SVs remained on the same chromosome (using the -chrom option) and avoided gaps in the GRCh38 assembly (using the -excl option). Similarly, we intersected the shuffled SVs with genomic elements and counted the number of hits to establish the null distribution. Fold changes were determined as the ratio of observed to permuted values for each transcript type, SV type (insertion and deletion), and genomic element (CDSs, exons, and introns, etc). Empirical p-values were calculated and adjusted for multiple comparisons, with significance reported for adjusted p-values less than 0.05.

Differential gene expression and outliers

Expression matrices for the 12 individuals (individuals with Iso-Seq) were obtained from short-read RNA-seq data. Differential expression analysis was performed using DESeq2¹¹⁶ (v1.38.3) for each SV carried by 2-10 individuals out of the 12. SVs were filtered into two categories: 1) direct overlap with an annotated exon (GENCODE v45), or 2) no overlap with an exon but located within 50 kbp of an annotated gene. For genes overlapping or near an SV, the p-values from DESeq2 were corrected using Benjamini-Hochberg adjustment, considering the total number of SV-gene pairs in each category (overlapping with exon or flanking). In these filtered gene-SV pair sets, genes with an adjusted p-value less than 0.05 were reported as differentially expressed.

Outlier expression analysis was performed for SVs carried by one or 11 individuals out of the 12. The gene expression FPKM matrix was used as input, assuming that FPKM follows a log-normal distribution to calculate p-values. The SV-gene pairs were identified and p-values were adjusted using the same methods as in the differential expression analysis. Genes with an adjusted p-value less than 0.05 were considered outliers. In each case, we annotated SVs associated with DE or outlier genes with overlapping cCREs identified using BEDTools¹²⁸ (v2.30.0) intersect.

We performed permutation tests to compare the impact of SVs on gene expression with random effects. SVs were randomly redistributed for 1,000 permutations using BEDTools (v2.30.0) shuffle while maintaining chromosome assignment, preventing overlaps, and excluding GRCh38 assembly gaps¹²⁸. To enable a sufficient number of shuffles while maintaining efficiency in differential expression analysis, we developed a method to directly extract DE results from a pre-calculated set that includes all possible individual contrasts within our dataset. This procedure can be applied consistently to both the real SVs and the shuffled sets. We found the number of SV-proximal DE genes in the real SV set was higher than each of the permuted results, suggesting a nonrandom association between SVs and expression change (empirical $p < 0.001$). In addition to using a 50 kbp window around the TSSs, we also varied the window size to 5 kbp, 25 kbp, and 100 kbp. We observed that narrower windows yielded a higher proportion of expression-altering SVs, underscoring the importance of sequence integrity near TSSs.

Hi-C analysis

We analyzed Hi-C data from 63 individuals to assess the chromatin structures across these genomes. Particularly, we focused on the identification of boundary regions that separate the two adjacent topologically associating domains (TADs) in these individuals. TAD boundaries are recognized for their high conservation across mammalian species and are more evolutionarily constrained than TADs themselves^{174,175}. On average, we have identified 21,021 TAD boundaries per genome in these individuals (**Supplementary Fig. 73, Supplementary Table 44**).

To identify these TAD boundaries, raw Hi-C sequencing reads were aligned to the GRCh38 reference genome for each individual. Specifically, raw sequencing reads were preprocessed using *Juicer* software tools (v1.6) with the default BWA-MEM aligner^{176,177}. Unaligned reads,

including abnormal split reads and duplicate reads, were removed. Additionally, read pairs with a mapping quality (MAPQ) value of less than 30 were filtered out. The resulting high-quality read pairs (Hi-C contacts) were used to construct chromatin contact maps (in *.hic* format, a special format for highly compressed binary files to store contact matrices at various resolutions). These Hi-C contact maps, along with TAD boundary identifications for each individual, were generated using a pipeline following our previous work^{23,178}. The files are primarily supported by *Juicer*, *JuicerTools*, and *Juicebox* command line tools for downstream analysis and visualization¹⁷⁹. Data normalization was done using SCALE matrix balancing¹⁷⁶, which addresses the limitations of KR normalization in unexpectedly not providing coverage for specific regions or chromosomes.

To determine the optimal resolution for calling TAD boundaries, we utilized a script from *Juicer* to calculate the Hi-C map resolution¹⁷⁶, which was directly downloaded from Rao's study¹⁸⁰. The Hi-C map resolution reflects the finest scale at which local features can be reliably detected, and the average resolutions were 9,137 (bp) for GRCh38 reference mapping in our individuals. We employed an algorithm utilizing *Insulation Score (IS)* to detect the coordinates and corresponding boundary scores (BS) for each TAD boundary under 10 kbp resolution, using the FAN-C toolkit¹⁸¹ (v0.9.26b2) with default parameters. TAD boundaries were detected at a 100 kbp window size with a minimum boundary score cut-off value of 0.2, as referenced from the 4DN domain calling protocol^{182–184}. The *IS* algorithm defines a sliding window and sums contacts within this window along the Hi-C matrix diagonal. Regions with low insulation scores represent high boundary scores and are further identified as TAD boundaries, while regions with high insulation scores (low boundary scores) are typically found within domains, referred to as TAD regions (**Supplementary Fig. 73**).

Results

SV effects on GWAS loci

Using the GWAS summary statistics¹¹⁸ (*gwas_catalog_v1.0.2-associations_e111_r2024-04-16.tsv*), which contains 149,081 autosomal SNPs, and we observed 133,695 SNVs present in the GWAS dataset and 10,008 INDELs and 1,547 SVs are at least one bp overlap with any GWAS signal.

Expression impacted by SVs

Structural Variations (SVs) have the propensity to cause drastic alterations to the genomic landscape, with the possibility of impacting transcript diversity through the disruption of coding sequence¹⁸⁵, altering splicing¹⁸⁶, or by rewiring regulatory networks¹⁸⁷. Using the Ensembl¹⁰⁸ (release 111) VEP¹⁰⁷ to intersect 176,531 SVs (PAV freeze 4 callset) with merged Ensembl GRCh38p14 gene annotations, we identified 1,505 unique SVs with evidence for protein-coding exon disruption within 1,070 unique genes across the 130 haplotype assemblies. In the subset of assemblies with corresponding Iso-Seq data (12 individuals; 24 haplotypes) we identified 582 unique SVs that disrupted protein-coding exons within 470 unique genes. After exclusion of 33 immunoglobulin genes, we identified 136 genes with Iso-Seq expression. Of these 136 genes, 85 were found to contain splicing variations absent from the GRCh38 Refseq Annotation¹¹³.

Visual comparison using IGV¹¹² of isoforms phased to variant and wild-type haplotypes for candidate genes revealed SV-driven transcript diversification.

Following scrutinization of the candidate genes, we found *ZNF718* to best exemplify the impact of SV-driven transcriptome diversification. Associated with immune function¹⁸⁸, *ZNF718* was found to contain a 6,142 bp deletion on chr4:127,125-133,267 affecting 10 haplotypes (8 individuals). This deletion resulted in the loss of exons 2 and 3 of the canonical transcript, as well as the majority of an alternate first exon (**Fig. 2e**). We identified a total of 16 unique isoforms in the twelve individuals, nine of which were expressed across the ten variant haplotypes, and seven across the 14 wild-type haplotypes. No isoforms were found in common between the variant and wild-type haplotypes. Of the nine unique isoforms expressed by the variant haplotypes, seven were previously unreported in Refseq¹¹³, GENCODE¹⁰⁶, and CHES¹¹⁴ databases, while the remaining two were known transcripts that lack the deleted exons. Some unreported isoforms displayed previously uncharacterized splice sites located up and downstream of the SV (isoforms 1, 3, 4; **Fig. 2e**), while others (isoform 2) displayed a loss of all splicing. In both cases, the result was the retention of an intronic sequence that had not been observed previously. In the absence of the deleted portion, the repetitive element landscape appeared to influence the aberrant splicing patterns. Apparent activation of novel splice sites were localized within *Alu* elements. Additionally, all previously unrecognized isoforms appeared to terminate in a transposable element. In contrast, the wild-type haplotypes harbored three known and four unreported isoforms, which all contained some portion of the deleted exons. All isoforms harbored by the wild-type haplotypes contained some portion of the deleted exons, indicating the disruptive impact of this deletion.

Enrichment of SVs in genomic elements

Permutation tests revealed significant depletion (empirical $p < 0.05$, Benjamini-Hochberg correction) of both deletions and duplications across most GENCODE-defined genomic elements, including all biotypes within protein-coding genes and long noncoding RNAs, mirroring our previous findings²⁸. Depletions were also observed for SVs within ENCODE-derived cCREs, including promoters and proximal and distal enhancers. However, we found significant enrichments for deletions within unprocessed pseudogenes, in addition to intronic regions of processed pseudogenes (**Extended Data Fig. 3e**), consistent with the SV-mediated mechanisms that lead to pseudogene formation¹⁸⁹.

Effects of SVs on chromatin conformation

Among the 128 SV-gene pairs (122 unique SVs associated with 98 genes) that exhibit significant differential gene expression changes in the 12 individuals with Iso-Seq data, we aimed to identify SVs within these pairs that are also associated with differential insulated regions (DIRs). DIRs represent regions with differential insulation profiles in the presence/absence of SVs across the 12 analyzed individuals. Specifically, we first filtered out SVs with missing genotypes in more than 6/12 individuals. For each remaining SV, we extracted the 50 kbp upstream and downstream of the annotated TSS position for each paired gene with corresponding insulation scores under 10 kbp resolution. For those insulated regions intersecting with more than one SV, we applied a local multi-test correction. An FDR < 0.05 from the two-sided Wilcoxon rank-sum test was considered as being significant. As a result, we

identified 49 SV-gene pairs (29 SVs associated with 24 genes) that were associated with 40 differential insulated regions (DIRs)^{23,178} (**Supplementary Table 44**).

Effects of SVs on GWAS loci

We identified 794,097 SNVs, 138,369 INDELs, and 3,818 SVs in high LD ($R^2 \geq 0.8$) with GWAS SNPs. Specifically, 468,217 SNVs, 77,821 INDELs, and 2,430 SVs were found to be in the highest LD ($R^2 = 1$) with GWAS SNPs (**Supplementary Table 45**, **Supplementary Fig. 74**). Among the SVs in high LD with GWAS SNPs, we observed 3 deletions (chr10-132173822-DEL-112, chr10-132191104-DEL-55, and chr14-67910288-DEL-152) that were identified as SV-differentially expressed gene pairs and linked to DIRs (**Supplementary Table 44**). Notably, chr14-67910288-DEL-152 (associated with the *RAD51B* gene) is in perfect LD with GWAS SNP rs10136790 ($R^2 = 1$), an SNP reported to be associated with the trait of balding measurement¹⁹⁰. This deletion is also in high LD with rs11158717 ($R^2 = 0.8536$), a locus associated with blonde hair, compared to black and brown hair colors¹⁹¹. In addition, chr10-132173822-DEL-112 (associated with the *DPYSL4* gene) is in high LD with GWAS SNPs rs2814162 and rs11146237 ($R^2 = 0.8235$ and 0.9743 , respectively). The SNP rs2814162 is associated with vertex-wise sulcal depth for brain measurement, a robust but underexplored measure of localized brain folding that has been previously associated with multiple neurodevelopmental disorders. Sulcal depth is considered a promising neuroimaging marker that could advance people's understanding of cortical morphology¹⁹². SNP rs11146237, which is also in high LD with the SV chr10-132191104-DEL-55, is associated with body mass index and body weight¹⁹³. Furthermore, we identified four INDELs (chr4-11038-INS-15, chr4-37975-DEL-1, chr4-79676-INS-2, and chr4-82371-DEL-1) that are all in high LD with one GWAS SNP, rs2187874, which is associated with the *ZNF718* gene - a candidate gene that illustrates the impact of SV-driven transcriptome diversification in our study. The SNP rs2187874 has been linked to the response to tumor necrosis factor (TNF) inhibitor in rheumatoid arthritis and joint damage measurement¹⁹⁴.

Genotyping

Contributing authors: Jana Ebler, Glenn Hickey, Tobias Marschall, Benedict Paten, Timofey Prodanov, Tobias Rausch, Michael C. Zody

Methods

Building a pangenome graph with Minigraph-Cactus

We built two pangenome graphs with Minigraph-Cactus¹²⁰ (v2.7.2). First, we constructed a 132-genome graph composed of the 65 HGSVC genomes (130 haplotypes) including GRCh38 and T2T-CHM13 references. Second, we constructed a 216-genome graph including the 65 genome assemblies (130 haplotypes) along with 42 genome assemblies (84 haplotypes) from the HPRC year 1 release¹ (**Supplementary Fig. 31**, Step 1). The input and output data for these two graphs are available at https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=pangenomes/scra/tch/2024_02_26_minigraph_cactus_hgsvc3/ and

<https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=pangenomes/scra>
[tch/2024_02_23_minigraph_cactus_hgsvc3_hprc/](https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=pangenomes/scra), respectively.

The command to build the HGSCV graph was:

```
cactus-pangenome ./js1 ./hgsvc3-2024-02-23-mc-chm13.seqfile.local
--outName hgsvc3-2024-02-23-mc-chm13 --outDir
hgsvc3-2024-02-23-mc-chm13 --reference CHM13 GRCh38 --vcf
--vcfReference CHM13 GRCh38 --odgi --xg --chrom-vg clip --chrom-og
--gbz clip full --xg clip full --haplo clip --giraffe clip --logFile
hgsvc3-2024-02-23-mc-chm13.log --batchSystem slurm --mgCores 64
--mapCores 16 --consCores 64 --indexCores 64 --maxMemory 1.5TB
--batchLogsDir batch-logs --coordinationDir /data/tmp
```

The command to build the HGSCV+HPRC graph was:

```
cactus-pangenome ./js1
./hgsvc3-hprc-2024-02-23-mc-chm13.seqfile.local --outName
hgsvc3-hprc-2024-02-23-mc-chm13 --outDir
hgsvc3-hprc-2024-02-23-mc-chm13 --reference CHM13 GRCh38 --vcf
--vcfReference CHM13 GRCh38 --odgi --xg --chrom-vg clip --chrom-og
--gbz clip full --xg clip full --haplo clip --giraffe clip --logFile
hgsvc3-hprc-2024-02-23-mc-chm13.log --batchSystem slurm --mgCores 64
--mapCores 16 --consCores 64 --indexCores 64 --maxMemory 1.5TB
--batchLogsDir batch-logs --coordinationDir /data/tmp
```

```
cactus-graphmap-join ./js1 --outName hgsvc3-hprc-2024-02-23-mc-chm13
--outDir hgsvc3-hprc-2024-02-23-mc-chm13 --vg
hgsvc3-hprc-2024-02-23-mc-chm13/chrom-alignments/chr*.vg --hal
hgsvc3-hprc-2024-02-23-mc-chm13/chrom-alignments/chr*.hal
--reference CHM13 GRCh38 --vcf --vcfReference CHM13 GRCh38 --odgi
--chrom-vg clip --chrom-og --gbz clip full --logFile
hgsvc3-hprc-2024-02-23-mc-chm13.join.log --indexCores 64 --maxMemory
1.5TB --batchLogsDir batch-logs --coordinationDir /data/tmp
--batchSystem slurm
```

Note that this second command was used to rerun the indexing step due to running out of memory on our cluster. This issue has been fixed in recent versions of Cactus.

We produced normalized versions of the VCF output of Minigraph-Cactus using `vcfwave`¹⁶² for realignment, similar to as was done for the HPRC year 1 data and described here: <https://github.com/ComparativeGenomicsToolkit/cactus/blob/hprc-v1.1/doc/mc-pangenomes/hprc-v1.1-mc.md#vcf-postprocessing>. The commands run for this were:

```
for i in hgsvc3-2024-02-23-mc-chm13.GRCh38
hgsvc3-hprc-2024-02-23-mc-chm13.GRCh38 hgsvc3-2024-02-23-mc-chm13
hgsvc3-hprc-2024-02-23-mc-c
```

```

hml3 ; do docker run -it --rm -v $(pwd)/:/data --user $(id -u):$(id -g) ghcr.io/pangenome/pggb:202402032147026ffe7f bash -c "/data/vcf-b
ubwave.sh /data/${i}.raw.vcf.gz /data/${i}-vcfbub.a100k.wave.vcf.gz"
; done

wget -q
https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemb
lies/analysis_set/chm13v2.0.fa.gz

wget -q
http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/tec
hnical/reference/20200513_hg38_NoALT/hg38.no_alt.fa.gz

for i in hgsvc3-2024-02-23-mc-chm13.GRCh38
hgsvc3-hprc-2024-02-23-mc-chm13.GRCh38; do bcftools norm
${i}-vcfbub.a100k.wave.vcf.gz -f hg3
8.no_alt.fa.gz | bcftools sort | bgzip --threads 2 >
${i}-vcfbub.a100k.wave.norm.vcf.gz ; tabix -fp vcf
${i}-vcfbub.a100k.wave.norm.vcf

.gz ; done

for i in hgsvc3-2024-02-23-mc-chm13 hgsvc3-hprc-2024-02-23-mc-chm13 ;
do bcftools norm ${i}-vcfbub.a100k.wave.vcf.gz -f chm13v2.0.fa.gz

| bcftools sort | bgzip --threads 2 >
${i}-vcfbub.a100k.wave.norm.vcf.gz ; tabix -fp vcf
${i}-vcfbub.a100k.wave.norm.vcf.gz ; done

```

VCF generation and preprocessing

We used the VCF file: https://s3-us-west-2.amazonaws.com/human-pangenomics/pangenomes/scratch/2024_02_23_minigraph_cactus_hgsvc3_hprc/hgsvc3-hprc-2024-02-23-mc-chm13.vcf.gz (Note: VCF prior to the vcfwave-based decomposition) generated by the Minigraph-Cactus pipeline¹²⁰ (see Methods section “Building a pangenome graph with Minigraph-Cactus” above) from the 216-genome containing all 65 genomes as well as 42 genomes previously assembled by the HPRC¹ as a basis for genotyping (**Supplementary Fig. 31**, Step 2). This VCF contains a record for each top-level bubble of the graph, as well as haplotype information for all genomes contained in the graph. Prior to genotyping, we preprocessed this VCF as follows. We converted the genotypes for male genomes on chromosomes X and Y to a homozygous representation by duplicating the haplotype in the VCF that carries the smaller number of “.” alleles. Next, we filtered out all records for which at least 20% of the haplotypes carry a missing allele (“.”). This resulted in a set of 30,490,169 bubbles. These bubbles are often huge and represent not just a single variant event, but rather accumulate many individual variant alleles across the different haplotypes in

the corresponding genomic region. Therefore, we ran a bubble decomposition method that we had developed earlier¹ to find the actual variant alleles nested inside of bubbles. Briefly, the decomposition approach uses the traversals of reference and alternative sequences of the bubbles to determine variation nested inside of the bubble structures¹. As a result, the pipeline adds annotations to the ID tag of the INFO field of each of our VCF records. For each alternative allele in the VCF, the ID tag contains IDs encoding all nested variant alleles it is composed of, separated by a colon (“bubble vcf”, PanGenie input). The decomposition pipeline also outputs a second VCF with a separate, bi-allelic record for each such nested variant ID, defining its reference and alternative sequence¹. This second VCF can be used later in order to convert genotypes computed for the bubbles to genotypes for all individual variant alleles nested in the graph based on the annotations (“decomposed vcf”). We distinguish the following variant types for the decomposed alleles: SNPs, indels (1-49 bp), SV deletions, SV insertions and other SVs (≥ 50 bp). SV deletions include all alleles for which $\text{length(REF)} \geq 50$ bp and $\text{length(ALT)} = 1$, SV insertions include all alleles for which $\text{length(REF)} = 1$ and $\text{length(ALT)} \geq 50$ bp. “SV other” contains all other cases where $\text{length(REF)} \geq 50$ bp or $\text{length(ALT)} \geq 50$ bp. In total, we obtained 39,313,178 variant alleles after decomposing all 30,490,169 bubbles. In the following, all evaluations are based on the decomposed bubbles.

PanGenie input data

PanGenie was run using the VCFs described in the previous section (available here: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/release/Genotyping_1kGP/20240716_PanGenie-genotypes/):

- MC_hgsvc3-hprc_chm13_filtered_bubbles.vcf (bubble vcf)
- MC_hgsvc3-hprc_chm13_filtered_decomposed.vcf (decomposed vcf)

As a reference genome, we used T2T-CHM13 (v2.0) obtained from: https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/analysis_set/chm13v2.0.fa.gz

Illumina reads were obtained from the locations provided below. One FASTA file per sample was generated by concatenating individual FASTA/FASTQ files into a single file.

- 3,202 1kGP samples: <http://ftp.sra.ebi.ac.uk/vol1/fastq/>
- NA24385: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/NIST_Illumina_2x250bps/reads/
- NA21309: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/HPRC_PLUS/NA21309/raw_data/Illumina/child/
- HG01123, HG02486, HG02559: <https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=submissions/30E441F3-6820-4BF6-BCF4-E64D56C8D6A4--TRUSEQ/>

We provide PanGenie command lines in the sections below.

Leave-one-out experiments

We performed “leave-one-out” experiments to evaluate the performance of PanGenie³ (v3.1.0) on all 107 assembly samples contained in our VCF. This was done by iteratively removing a single sample from the PanGenie input VCF (“bubble vcf”) and genotyping it with PanGenie based on the remaining samples using 1000 Genomes high coverage Illumina reads²⁹. PanGenie was run with additional parameter -a 108 using the following commands for each panel sample:

```
bcftools view --samples ^<sample>
MC_hgsvc3-hprc_chm13_filtered_bubbles.vcf | bcftools view --min-ac 1
> MC_hgsvc3-hprc_chm13_filtered_bubbles_<sample>.vcf

PanGenie-index -v MC_hgsvc3-hprc_chm13_filtered_bubbles_<sample>.vcf
-r chm13v2.0.fa -o index -t 24

PanGenie -a 108 -f index -i <(gunzip -c <sample-reads>) -o
pangenie_<sample> -j 24 -t 24 -s <sample>

cat pangenie_<sample>_genotyping.vcf | python3
convert-to-biallelic.py
MC_hgsvc3-hprc_chm13_filtered_decomposed.vcf.gz >
pangenie_<sample>_genotyping_biallelic.vcf
```

The genotypes of the left-out sample were then used as a ground truth for evaluation. Because PanGenie is a re-genotyping method, it can only genotype variants contained in the input panel VCF, that is, it is not able to detect variants unique to the genotyped sample. For this reason we removed all variant alleles (after decomposition) unique to the left-out sample contained in the truth set for evaluation. In order to evaluate the genotype performance, we used the weighted genotype concordance^{1,8}. We evaluated the results within the following regions:

- All regions
- Biallelic bubbles: include all bubbles in the graph with at most two branches
- Multiallelic bubbles include all bubbles in the graph with more than two branches
- SegDup regions: UCSC segmental duplications track (CHM13)¹⁹⁵
- HPRC gap regions: regions with known gaps in HPRC assemblies²
- CMRG regions: challenging medically relevant genes³¹

Results for structural variants ($\geq 50\text{bp}$) are shown in **Supplementary Figs. 32-34**. Overall, we observed high genotype concordances across all regions. Performance is best within biallelic regions of the graph and drops in multiallelic regions. This is expected as multiallelic bubbles often contain a large number of different variant alleles. While genotyping performances within CMRG regions are very similar to the overall concordances, they are slightly worse within segmental duplications, and the worst performance was observed within regions where the HPRC assemblies have gaps. This is expected, since our panel contains 42 HPRC samples, and locally incomplete assemblies negatively affect the genotyping.

Genotyping the 1kGP cohort with PanGenie

We genotyped all 3,202 samples of the 1000 Genomes cohort based on high coverage Illumina reads²⁹ using PanGenie³ (v3.1.0). We first ran PanGenie-index once using the command:

```
PanGenie-index -v MC_hgsvc3-hprc_chm13_filtered_bubbles.vcf -r  
chm13v2.0.fa -o index -t 24
```

Then, we ran genotyping on each sample using the command:

```
PanGenie -a 108 -f index -i <(gunzip -c <sample-reads>) -o  
pangenie_<sample> -j 24 -t 24 -s <sample>
```

```
cat pangenie_<sample>_genotyping.vcf | python3  
convert-to-biallelic.py  
MC_hgsvc3-hprc_chm13_filtered_decomposed.vcf.gz >  
pangenie_<sample>_genotyping_biallelic.vcf
```

We filtered the resulting genotypes using an approach developed previously for the HPRC and HGSCV projects^{1,8} that is based on a regression model. We trained the model based on a positive and a negative set which we defined based on the following filters:

- **ac0_fail**: a variant allele was genotyped with a allele frequency of 0.0 across all samples
- **mendel_fail**: the mendelian consistency across trios is less than 85% for a variant allele. Here, we use a strict definition of mendelian consistency which excludes all trios with only 0/0, only 0/1 and only 1/1 genotypes.
- **gq_fail**: less than 50 high quality genotypes were reported for this variant allele
- **self_fail**: genotyping accuracy of a variant allele across the panel samples is less than 90%
- **nonref_fail**: not a single non-0/0 genotype was genotyped correctly across all panel samples

To select for genotypable variants, we trained a machine-learning classifier to assign a pass or fail filter value to each genotyped variant. To train the model, all variant alleles that pass all five filters are included in the positive set. The negative set contains all variant alleles that passed the *ac0_fail* filter but failed at least three of the other filters. In a similar manner as before^{1,8}, we used Support Vector Regression based on features including the mendelian consistency, allele frequencies and number of alleles transmitted from parent to child samples in order to compute scores between 1 (good) and -1 (bad) for all remaining SV variant alleles. Our final, filtered variant set includes the positive set and all variant alleles with a score ≥ -0.5 . We show the number of variants in the unfiltered, positive and filtered sets in **Supplementary Table 46**. Since our focus is on SVs, we applied our machine learning approach only to SVs. Our filtered set contained 89%, 88% and 81% of all deletions, insertions and other SVs that we genotyped. These percentages are much higher than the ones we previously observed when we analyzed

HPRC data ¹, which is mainly because of several improvements made to the PanGenie software (implemented in versions > 3.0.0, version used here: v3.1.0) in the meantime that led to smaller fractions of variants with allele frequencies of zero across all genotyped samples.

To evaluate the PanGenie genotypes, we compared the allele frequencies of SV alleles that we observed across all assembly samples in our input VCF to the allele frequencies of the same variant alleles after genotyping all 1kGP samples. This was done on all 2,590 unrelated samples. We observed that allele frequencies matched quite well, with pearson correlations of 95.2%, 94.9% and 90.8% for the unfiltered SV deletions, SV insertions and other SV alleles; and 98.1%, 95.7% and 91.9% respectively, after filtering (**Supplementary Fig. 35**). We furthermore compared heterozygosities to allele frequencies and observed a behavior close to what is expected by Hardy-Weinberg equilibrium (**Supplementary Fig. 36**).

We furthermore compared our genotypes to other SV sets, like the PanGenie genotypes we had computed previously for the HPRC data ¹ as well as the Illumina-based NYGC SV callset²⁹. While our genotypes are based on CHM13, the other sets are both based on GRCh38, making a one-to-one comparison more difficult. In order to account for different SV representations between the three sets, we compared them based on the number of SV sites per sample. We ran `truvari collapse`¹⁹⁶ using parameters `-r 500 -p 0.95 -P 0.95 -s 50 -S 100000` to merge similar SV alleles in our filtered genotyped set, as well as in the HPRC genotypes, since no SV merging was performed for both sets. We show the results in **Supplementary Fig. 38**. To ensure that the improvements we see for the our genotypes were not just due to artifacts arising from the different underlying reference genomes of the compared callsets, we additionally counted the number of heterozygous SVs per sample (**Supplementary Fig. 39**), which revealed the same trends.

The pipeline used for the Leave-one-out experiments and the 1kGP genotyping, as well as instructions to replicate the results can be found here: <https://github.com/eblerjana/hgsvc3/tree/main/experiments/genotyping/>

Comparison of PAV and Minigraph-Cactus based calls

We compared CHM13-based SV calls produced by PAV to the ones obtained from the Minigraph-Cactus graph. Four callsets were compared: the unfiltered PAV calls (“PAV-unfiltered”), the final, filtered PAV calls (“PAV”), the Minigraph-Cactus based panel VCF used as input to PanGenie (“MC-panel”) and the filtered PanGenie genotypes (“PanGenie”). Variants were compared based on reciprocal overlap, breakpoint deviations and variant length deviations. More specifically, two SVs were considered to represent the same event if at least one of the following conditions was true:

- reciprocal overlap > 50%
- $abs(s_1 - s_2) < 200$ and $abs(e_1 - e_2) < 200$ and $0.5 < \frac{l_2}{l_1} < 1.5$

Here, (s_1, e_1) and (s_2, e_2) are the start and end coordinates of the MC/PanGenie-filtered and the PAV SVs, respectively (relative to the reference genome), and l_1 and l_2 the variant lengths.

Furthermore, the variant type must be the same, unless one of the SVs has type “OTHER”. The latter ones are allowed to match with any other type.

We performed four pairwise comparisons: MC-panel vs. PAV-unfiltered, MC-panel vs. PAV, PanGenie vs. PAV-unfiltered and PanGenie vs. PAV. Results are shown in **Supplementary Fig. 37**. For overlapping SV calls, counts and variant types were determined relative to the MC / PanGenie-filtered SVs.

Unmatched calls

We found 49,868 SVs that are present in the PAV callset, but could not be matched in the MC-panel. Of those, 14,215 SVs were overlapping with calls previously filtered out when creating the MC-panel VCF from the Minigraph-Cactus VCFs (see section “VCF generation and preprocessing” above), while the remaining 35,653 SVs had no overlap with any of these filtered out calls. Furthermore, we found 85,315 SVs in the PanGenie set for which no match was found in the PAV calls. Since PanGenie was run on the MC-panel containing HPRC assemblies as well (in addition the the HGSVC3 samples), some of these variants might be SVs unique to HPRC samples and thus not included in the PAV set, which was produced based on the HGSVC3 samples only. Therefore, we checked the allele frequencies of these variants across HPRC and HGSVC3 samples and found that for 55,095 of 85,315 SVs the allele frequency across HGSVC3 samples was non-zero.

We further analyzed the 35,653 PAV calls that could not be matched in the MC-panel and the 55,095 SVs in the (MC-based) PanGenie filtered set that could not be matched in the PAV calls. We found that many of these calls (i.e., 24,235 of 55,095 and 15,119 of 35,653 SVs) are located within 5mb around the start and end positions of chromosomes. Therefore, we hypothesize that many of these unmatched are cases in which variant representations between the callsets are too different to be picked up by our matching criteria (see section “Comparison of PAV and Minigraph-Cactus based calls” above). Especially in repetitive contexts, coordinates of variant calls might be shifted, or larger records might be reported as multiple smaller calls and thus are missed by our matching script.

Generating a 1kGP reference panel with SHAPEIT

We used our final filtered PanGenie genotypes for all 3,202 1kGP samples (see section “Genotyping the 1kGP cohort with PanGenie” above) as a basis for phasing all 3,202 short-read 1kGP samples (**Supplementary Fig. 31**, Step 3). Since this genotyped set is based on variant calls obtained from 107 assembly samples, we included additional rare SNPs and indels from an external short-read based callset for the 3,202 1kGP samples obtained from https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=T2T/CHM13/assemblies/variants/1000_Genomes_Project/chm13v2.0/all_samples_3202/. This was done by selecting all variants from this short-read based set that reported genotype 0/0 for all samples overlapping with our 107 assembly samples. In this way, we added 70,174,243 additional variants to our genotyped VCF, resulting in a total of 101,980,762 variants.

In a next step prior to phasing, we set all genotypes with a quality score below 10 to missing (GQ < 10). Then, we ran SHAPEIT5¹²¹ (v5.1.1) using the phase_common command. We provided pedigree information for all trios part of the 3,202 samples, as well as genetic maps for

CHM13 obtained from https://github.com/JosephLalli/phasing_T2T/tree/master/t2t_lifted_chrom_maps. Phasing was run separately on each chromosome. For the sex chromosomes, a list of haploid samples (all males) was provided to SHAPEIT.

Evaluating phased genotypes

In order to evaluate the quality of our phased genotypes, we compared phasing results to other phased sets on a subset of samples. Other sets included the MC-based input panel VCF containing the assembly haplotypes (section “VCF generation and preprocessing”), as well as a phased Illumina-based short variant callset for the 3,202 1kGP samples obtained from https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/variants/1000_Genomes_Project/chm13v2.0/. We used WhatsHap’s compare function¹⁹⁷ to compute switch error rates (**Supplementary Fig. 40**). To compute switch error rates specific to regions provided in a BED file, we treated each BED interval as a separate phased block to avoid penalizing switch errors that happened outside of the regions of interest. In this way, we computed switch error rates for the following regions:

- CHM13-syntenic: all regions shared between GRCh38 and CHM13. Those were derived from the UCSC CHM13-unique track for CHM13v2.0.
- CHM13-easy: regions outside of other difficult regions such as tandem repeats, homopolymers, difficult to map regions, segmental duplications and high/low GC content. Obtained from: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.3/CHM13@all/Union/CHM13_notinalldifficultregions.bed.gz
- CHM13-lowmap-segdup: regions with low mappability and segmental duplication regions. Obtained from: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.3/CHM13@all/Union/CHM13_allowmapandsegdupregions.bed.gz
- CHM13-tandem-repeat: tandem repeat regions and homopolymers, obtained from: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.3/CHM13@all/LowComplexity/CHM13_AllTandemRepeatsandHomopolymers_slop5.bed.gz

In line with our expectations, switch error rates were highest for unrelated samples, i.e., samples not part of any of the 1kGP trios, lower for parent samples and lowest for the children (**Supplementary Fig. 40**). Results for CHM13-syntenic regions were almost identical to the overall error rates (orange and blue line, **Supplementary Fig. 40**). Overall, the best results were observed for the “CHM13-easy” regions.

The overall switch error rates for unrelated samples (all regions) were between 1.16 and 1.50 when comparing to the MC panel VCF, and 1.16-1.55 when comparing to the Illumina phased calls (**Supplementary Fig. 40**). For parent samples, we observed switch error rates between 0.50-0.79 when comparing to the MC panel and 0.39-1.18 when comparing to the Illumina phased calls. For children, switch error rates were below 0.42 and 0.19 for the MC panel and the Illumina phased calls, respectively. We also performed a multiway comparison of all three sets using WhatsHap’s compare function¹⁹⁷. Briefly, it reports the number of times the different

sets agreed or disagreed on the phasing of variant sites. While the phasings of the MC panel and the Illumina calls largely agree, we observed slightly higher rates of disagreements between our and the other two sets (**Supplementary Fig. 41**), especially for the unrelated samples.

Personal genome reconstruction

We generated consensus haplotype sequences for all 1kGP samples based on the phasing results. For each haplotype of each sample, we implanted the variants present on that haplotype into the CHM13 reference genome (chromosomes 1-22 and X). In order to evaluate the quality of the resulting haplotype sequences, we computed QVs and completeness for a subset of 60 haplotypes (30 samples) by comparing k -mers of each haplotype to the k -mers seen in short-read sequencing data of the respective sample (k -mer size used: 21 bp). This was done analogously to how these statistics are computed in Merqury⁷². We furthermore restricted our QV and completeness calculations to certain regions. Given a BED file, we masked all sequences outside of the desired BED intervals with Ns and implanted variants only in the non-masked regions. The Ns were later removed and assembly statistics computed only for the remaining sequence. As a baseline estimate, we additionally computed QV and completeness statistics by flipping sample labels, i.e., by comparing the consensus haplotype for one sample to reads of another sample. In addition to generating consensus haplotypes from our phased genotypes, we also generated consensus haplotypes from the GRCh38-based high-coverage 1kGP set²⁹, which contains phased genotypes for SNPs, indels and SVs across all 3,202 samples, called from Illumina data and phased using SHAPEIT. The QV and completeness values we observed across all 60 evaluated haplotypes are shown in **Supplementary Fig. 42**.

The pipeline used for phasing and the generation of consensus haplotypes, as well as instructions to replicate the results can be found here: <https://github.com/eblerjana/hgsvc3/tree/main/experiments/phasing/>

Targeted genotyping using Locityper

We used Locityper v0.15.1³⁰ to genotype 3,202 samples from the 1kGP cohort²⁹. Target loci were selected based on 268 challenging medically relevant genes³¹ (5 genes not present in the CHM13 reference genome were discarded), extended by genes and pseudogenes from six polymorphic gene families: CFH, CYP2, HLA, KCNK, KIR, MUC and a hyperpolymorphic *LPA* gene. Short genes were extended to both sides to up to 10 kbp, and genes within 1 kbp of each other were combined together. Finally, reference panels were constructed based on the CHM13-based HPRC/HGSVC3 Minigraph-Cactus VCF file, while simultaneously extending locus boundaries if they overlap a pangenomic bubble (variant in the VCF file). Loci that needed to be extended by >300 kbp to one of the sides were discarded, producing a final set of 347 target loci. 61 samples with short read data and HGSVC assemblies were genotyped with two additional reference panels: leave-one-out panel, corresponding to the full panel without the actual sample assemblies; and HPRC-only panel, consisting of the reference genomes and HPRC assemblies. Assembled haplotypes from two individuals (HG00733 and HG02818) were mislabeled as HPRC assemblies, resulting in a bigger HPRC-only panel, which may have diminished the genotyping improvement achieved by switching from the HPRC to the HPRC/HGSVC3 leave-one-out reference panel.

In order to compare actual and predicted haplotypes, we used Locityper align module, which uses largest common k -mer subsequence between two haplotypes, and completes the gaps using wavefront alignment algorithm¹⁹⁸. We then calculated sequence divergence between haplotypes by dividing edit distance by the alignment length, and used Phred-like¹⁴⁵ transformation to convert sequence divergences into QV scores.

Haplotype availability analysis does not require sequencing data and only depends on the assembly sequences. Specifically, for a given sample haplotype we searched for an unrelated haplotype from a limited reference panel that would produce maximal QV score. To evaluate genotyping accuracy, we computed QV scores between predicted and actual haplotypes. All examined loci were diploid, correspondingly, we permuted predicted and actual haplotype pairs to achieve the lowest sequence divergence. Genotypes were marked as low quality and discarded if either weighted distance to other highly probable genotypes was over 30, or there were over 1000 and over 20% unexplained reads³⁰. When evaluating trio concordance, we compared predicted child haplotypes against predicted parent haplotypes, and reported corresponding QVs. Similarly to genotyping, we permuted against possible haplotype configurations to achieve the lowest sequence divergence.

Major histocompatibility complex

Contributing authors: Peter A. Audano, Christine R. Beck, Chen-Shan Chin, Alexander T. Dilthey, Peter Ebert, Lisbeth A. Guethlein, Miriam K. Konkel, Heng Li, Mark Loftus, Tobias Marschall, Paul J. Norman, Nicholas R. Pollock, Timofey Prodanov, Stephan Scholz, Arda Söylev, Ying Zhou

Methods

HLA-DRB and RCCX visualization

Visualization of DR subregion haplotypes and RCCX modules was performed using a custom Python script, the pyGenomeViz library (<https://github.com/moshi4/pyGenomeViz>), and the DNA Features Viewer library (<https://edinburgh-genome-foundry.github.io/DnaFeaturesViewer/>).

Pangenome analysis for the MHC class II region with PGR-TK

The MHC class II region (MHC delta block) exhibits very high diversity. It is unclear whether standard alignment and variant methodologies (small & structural) can be applied to this region. For instance, when attempting to align assembled contigs to the CHM13 reference, most alignments are fragmented, even with parameters that tolerate more differences. For example, using minimap2 for assembly-to-assembly alignment with up to 5% difference allowed (parameters: `-x asm20 -r 10000,100000 -z 100000,100000`), it generated 282 fragmented alignments out of the 111 haplotypes (HGSVC batch1 + batch2+ CHM13), and 167 of them had a soft or hard clip greater than 100 kbp on the 5-prime region (**Supplementary Fig. 75**).

We can evaluate the similarity/difference between the haplotypes using a k -mer-based, alignment-free method. We count all 25-mers from all haplotype sequences. As shown in the

plot below, the peak at 111 indicates conserved k -mers across the haplotype sequences. However, there are many significant non-conserved k -mers, reflecting the high diversity of the MHC class II region in the human population. There are 60.9% (857,179 out of 1,407,809) 25-mers shown in less than 50% of the haplotype sequences (**Supplementary Fig. 76**).

In our PGR-TK haplotype cluster analysis, we only select bundles longer than 6 kbp to generate the pangenome graph such that it is more comprehensible. The haplotype sequences in the bundle share a set of sparse minimizers, which can be aligned to each other through a sparse dynamic programming algorithm. At the sequence level, they are typically highly similar to each other. The plot below compares the k -mer Jaccard (with $k = 25$) distribution of different haplotypes of the entire MHC class II region (~CHM13, chr6:32,168,394-32,812,380) to each other and those haplotype sequences within each bundle (**Supplementary Fig. 77**).

Results

MHC and HLA-DRB

The Major Histocompatibility Complex (MHC), a region of approximately 5 Mbp on the short arm of chromosome 6 (**Fig. 4a**), is the most gene-rich, polymorphic, and disease-associated region of the human genome^{32–34}. Due to a combination of extensive structural variation in the HLA-DR and RCCX regions (**Fig. 4a**), long-range linkage disequilibrium¹⁹⁹, and difficulties in obtaining full sequence resolution from short reads²⁰⁰, however, the majority of MHC disease association signals—for example in the class II region for multiple sclerosis²⁰¹--have not been fine-mapped²⁰². Based on the 128 complete and 2 near-complete MHC haplotypes in our assemblies, representing the largest resource of complete MHC haplotypes so far, we characterized the fine-scale structure of genetic variation in the MHC region, providing the basis for more accurate mapping of MHC association signals and addressing key open questions about the relationship between different MHC haplotypes.

Our comparison of our MHC haplotypes with available MHC reference haplotypes^{39,40} show a high concordance with respect to the number of annotated^{40,203} genes [27-33 Human Leukocyte Antigen (HLA) and 140-146 non-HLA genes and pseudogenes per haplotype] as well as to the overall haplotype repeat content (**Supplementary Table 47**). Our HLA haplotypes were in excellent agreement with classical HLA typing results³⁵, with concordant results observed for 357 out of 360 compared alleles; **Supplementary Table 48**). A total of 826 HLA allele sequences not present yet at full-length resolution in the IPD-IMGT/HLA reference database were submitted³⁶ (**Supplementary Table 50**), including 112 sequences from the HLA-DRB locus, expression variants of which are associated with vaccine response³⁷. The linkage between the allelic states and presence/absence patterns of HLA genes were in agreement with known MHC class I and MHC class II haplotype structures around *HLA-H*, *-K*, *-T*, *-U*, *HLA-Y* and *-OLJ*²⁰⁴, as well as in the HLA-DR region (**Supplemental Fig. 78**). While the HGSVC3 MHC sequences did not contain any major novel HLA gene-defined haplotype structures, we did observe a previously unknown copy number variant, deletion of *HLA-DPA2* on one haplotype (HG03807.1), as well as low-frequency gene-level structural variants, such as deletion of *MICA* on one haplotype²⁰⁵, a gene associated with histocompatibility²⁰⁶. A total of 170 SVs not present in previously available reference haplotypes^{39,40} were detected (**Supplementary Table 51**); the majority of these (64%) had a repeat element content $\geq 70\%$, and 30 of the remaining ones were

larger than 1 kbp. Taken together, while not indicating any major novel haplotype structures, these results further confirm the high accuracy of the HGSVC3 assemblies in a region as complex as the MHC and inform fine-mapping of HLA-related association signals. Additionally, we used Locityper³⁰ to genotype 61 of our individuals using Illumina datasets across 19 protein coding genes and 14 pseudogenes from the MHC locus and compared predicted gene alleles against assembly-based gene annotation, constructed using Immuanot²⁰³. Across all 33 loci, Locityper fully predicted gene alleles in 81.0% cases using a limited HPRC-only reference panel. Adding HGSVC3 assemblies to the reference panel allowed Locityper to accurately predict gene alleles in 86.3% (leave-one-out panel) and 97.1% cases (full panel). When evaluating predicted protein product, Locityper achieved 88.3%, 90.6% and 97.8% accuracy with HPRC-only, leave-one-out and full reference panels, respectively.

We next carried out in-depth analyses at the nucleotide level of the two major structurally variable regions of the MHC, the HLA-DR and RCCX (**Fig. 4a**). We classified the HLA-DR haplotypes of the HGSVC3 assemblies using the established DR group system, which groups HLA-DR haplotypes according to their HLA-DRB1 allele and gene/pseudogene content (**Fig. 4b**, **Supplementary Table 52**, **Supplementary Fig. 79**). Our MHC assemblies contained representatives of all known DR groups, with counts ranging from 9 (DR1) to 25 (for each of DR5 and DR6); no haplotype configurations incompatible with the established DR group system were found. Based on visual inspection of pairwise alignments, with the exceptions of two small (~3.7 and 4.4 kbp) structural variants that contained solitary HLA-DRB exon sequences (see paragraph “Solitary HLA-DRB exon sequences [...]” below), no large SVs distinguishing between DR haplotypes within the same DR group were identified. DR5, DR8 and DR9 were absent from previous analyses of fully resolved MHC haplotypes^{39,40}; in addition, the relationship between the two contracted DR haplotypes DR1 and DR8, which carry only one functional HLA-DRB gene (DRB1; **Fig. 4a**), and other HLA-DR haplotypes have not been investigated at the nucleotide level.

For DR5 and DR9, pairwise sequence alignments (**Supplementary Fig. 48**) showed that these haplotype share their large-scale structure with DR3/DR7 and DR4/9, respectively, the other haplotype groups with the same DRB3- and DRB5-carrying gene configuration (**Fig. 4c**). For DR8, pairwise alignments suggested the highest degree of relatedness to either DR1 or DR3/5/6 (**Supplementary Fig. 48**). Leveraging the fully resolved nature of our MHC assemblies, we constructed high-resolution maps of repeat elements and gene/pseudogene exons within the assembled DR haplotypes (**Supplementary Fig. 46; Methods**). Shared repetitive element architecture encompassing the *HLA-DRB1* gene within DR8 to those found flanking the *HLA-DRB3* and *HLA-DRB1* genes in the DR3/5/6 haplotypes (**Fig. 4c**) indicated that, consistent with earlier analyses of DR8 based on only HLA allele sequences^{41,207}, a sequence contraction event between the *HLA-DRB1* and *HLA-DRB3* genes of a DR3/5/6-class haplotype may have given rise to the DR8 haplotype, which could be fine-mapped to lie within the proximity of exon 4 in HLA-DRB1*08 (**Supplementary Fig. 47**). The putative breakpoint showed a short length (<150 bp) of homology between *HLA-DRB1* and *HLA-DRB3* on DR3/5/6 spanning exon 4 and the surrounding downstream sequence (intron 3) in some DR8 haplotypes suggesting that homology-mediated double strand DNA break repair may be a more likely explanation than nonallelic homologous recombination (NAHR). For DR1, a haplotype structure that had previously been suggested to have arisen from a contraction event within an ancestral *HLA-DRB5* carrying DR2-like haplotype²⁰⁸, high-resolution repeat mapping suggested that DR1

is most likely derived from a recombination event between DR2 and DR4/7/9. DR1 can be split into two segments (**Fig. 4d**), the 5' segment of which exhibits the highest degree of repeat architecture similarity to the 5' region of the *HLA-DRB4*-carrying haplotype groups (DR4/7/9), and the 3' segment of which shows the highest degree of similarity to a region found in DR2. Breakpoint analysis identified homologous LINE/L1 elements at the putative breakpoints in DR1, DR4/7/9 (L1 located between *HLA-DRB9* and *HLA-DRB4* within DR4/7/9), and DR2 (L1 located between *HLA-DRB5* and *HLA-DRB6*) (**Supplementary Fig. 49**).

Solitary HLA-DRB exon sequences have been hypothesized to represent remnants of ancestral nonhomologous recombination events, possibly mediated by the presence of LINE1 and Alu in HLA-DRB intron 1 sequences^{42,209}; in previous analyses of incompletely resolved MHC haplotypes, they were found in the 3' region of *HLA-DRB9*⁴². Based on the set of completely resolved HGSVC haplotypes, we created a comprehensive catalog of the locations of HLA-DRB intron and exon sequences across DR groups (**Fig. 4b**). For the region around *HLA-DRB9*, we refined previously available estimates of the copy number of solitary intron and exon sequences, e.g., observing a copy number of 3 instead of 1 for solitary HLA-DRB exon 1 sequences in the *HLA-DRB9* region of DR1. Solitary exon 1 (DR1 group haplotypes) and intron 1 (for all DR groups but DR8) sequences were also found approximately 10 kbp 3' of *HLA-DRB1*, i.e., at a locus not previously associated with the presence of solitary intron or exons (**Fig. 4b**). Furthermore, the presence of solitary exon 1 sequence in the *HLA-DRB1* region of DR1 and in the DRB9 region of DR3, DR5, DR6, and DR8, was variable even within haplotypes of the same DR group (**Fig. 4b**); the underlying two structural variants (corresponding to the solitary exon-polymorphism in the DRB1 and DRB9 regions, respectively) were identified during the structural analysis of MHC haplotypes (see paragraph "We next carried out in-depth analyses [...]" above) and carried HLA-DRB intron 1 and exon 1 sequences as well as repeat elements (**Supplementary Fig. 80**).

Last, we asked whether the established DR group nomenclature could be recapitulated using unbiased and data-driven analysis. To this end, we analyzed a subset of the HGSVC3 MHC haplotypes ($n=55$) as well as T2T-CHM13⁴ using PGR-TK, a pangenomic multiscale analysis method⁴⁵ (**Fig. 4e**). We examined conserved blocks in 111 haplotypes (110 HGSVC haplotypes + CHM13) greater than 6 kbp, identified by PGR-TK. Multiscale hierarchical clustering of the haplotypes perfectly recovered the traditional DR group system in the region around *HLA-DRB1* (**Fig. 4e**); at the same time, the identified additional diversified subgroups could serve as the basis for a more fine-grained future classification of HLA-DR haplotypes or be used in the context of genome-wide association studies.

RCCX

The RCCX region in the class III region of the MHC is characterized by a modular pattern of copy number variation. Each module encodes one copy of the Complement component 4 (*C4*) gene ("C") as well as one functional or nonfunctional/degenerate copy of Serine/threonine kinase 19 ("R"), Steroid 21-hydroxylase ("C"), and Tenascin XB protein ("X"), referred to as *STK19*, *CYP21A2*, and *TNXB* (functional forms) and *STK19B*, *CYP21A1P*, and *TNXA* (pseudogene / degenerate forms), respectively.

RCCX varies in module copy number, commonly ranging from 1-3 modules per haplotype, although rare four module haplotypes have been observed previously^{43,210}. Duplications and deletions of RCCX modules as discrete genetic units are commonly observed as the region is prone to NAHR^{44,210,211}. Typically, when the RCCX is present as a mono-module, all four genes within the unit are functional, barring any mutation/recombination. However, when the RCCX contains multiple units (bi/tri-module), *C4* is the sole gene that retains its functional capacity across all discrete RCCX modules, while the additional copies of the other three genes (*STK19*, *CYP21A2*, and *TNXB*) often are present as nonfunctional pseudogenes (*STK19B*, *CYP21A1P*, and *TNXA*)

The *C4* gene is composed of 41 exons that together encode for an inactive 1744 amino acid long glycoprotein. This inactive form contains three polypeptide chains (alpha, beta, and gamma) and becomes activated through enzymatic cleavage. The *C4* gene exists in four allelic variants, distinguished by four amino acid residues in exon 26 (positions p1120-1125, *C4* alpha chain), referred to as “A” (PCPVLD) and “B” (LSPVIH) variants of *C4*^{212,213} as well as by the presence or absence of a ~6.36 kbp human endogenous retrovirus (HERV) insertion within intron 9, referred to as “short” (S) and “long” (L) versions of *C4*.

For RCCX, we carried out a comprehensive investigation of the region’s modular structure of genetic variation (**Supplementary Fig. 50**). RCCX haplotypes consist of 1 - 4 RCCX modules^{43,214,215}, each module typically encoding one functional variant of the *C4* gene (*C4AS*, *C4AL*, *C4BS*, *C4BL*) and three additional genes in their functional (*STK19*, *CYP21A2*, and *TNXB*) or pseudogenic (*STK19B*, *CYP21A1P*, and *TNXA*) forms (**Supplementary Fig. 50**); determining the phase and relative order of variants in the region from short-read sequencing or other molecular typing techniques^{43,216} is challenging due to extensive sequence homologies. Consistent with prior reports^{211,217}, RCCX bi-modules were, at 74.6% ($n=97$), observed most frequently in the HGSVC3 MHC haplotypes, followed by 13.1% ($n=17$) for mono-modules and 12.3% ($n=16$) for tri-modules (**Supplemental Fig. 50**). We were able to detect and completely resolve rare RCCX haplotype configurations and characterize likely gene conversion events, which play an important role in RCCX evolution^{44,218,219}. For example, we observed two haplotypes with tri-modular RCCX structures containing two functional *CYP21A2* copies, likely generated by gene conversion turning the *CYP21A1P* pseudogene into its functional counterpart; one mono-modular and one bi-modular haplotype with no functional *CYP21A2* genes; and one tri-modular haplotype exhibiting the unique configuration of *C4B* preceding *C4A* and carrying two copies of *CYP21A2*, one of which was nonfunctional (**Fig. 4d**). The latter haplotype was likely generated by the introduction of one nonsense mutation and two gene conversion events, turning *CYP21A1P* into *CYP21A2* and *C4A* into a *C4B* that unusually encoded the Rodgers (Rg) blood group epitope. An in-depth analysis of *C4* copy numbers (**Supplementary Fig. 51**) and allelic states (**Methods**) revealed seven novel amino acid variants, four residing within the beta-chain (283R->C, 530S->L, 566A->G, 614K->Q), and three within the gamma-chain (1477K->N, 1722R->C, 1724R->W) (**Supplementary Fig. 52**). In addition to providing a more complete catalog of allelic variation, our results thus provide the basis for assessing the differential effects of allelic, copy number and structural variation in RCCX-associated autoimmune phenotypes²²⁰.

C4 copy number analysis

Of a total of 259 *C4* copies, 53.3% ($n=138$) were identified as *C4A* and 46.7% ($n=121$) as *C4B*. Most individuals (56.9%) harbored four *C4* gene copies, ranging from a minimum of two *C4* copies (15.4%) per individual to a maximum of six *C4* copies (3.1%); *C4A* gene copies ranged from 1-3 copies and *C4B* from 0-4 copies per genome. Most genomes contained at least two *C4A* (72.3%) copies, and/or two copies of *C4B* (61.5%).

Most (66.4%, $n=172$) *C4* copies were seen in the long form, *C4L*, and ranged from 1-5 copies within a genome; while the short form *C4S* constituted 33.7% ($n=87$) of all *C4* copies and ranged from 0-3 copies within a genome. Interestingly, an absence of the short *C4* form, *C4S*, was more common in non-African individuals as we identified only one African individual (HG02769) containing zero *C4S*, consistent with ²¹⁶.

Manual curation of HG00514.hap2

STK19 and *TNXB* were found to be present and functional within all assemblies except HG00514.hap2 (version 1). This haplotype contained a nonsense mutation within *TNXB* caused by an indel. Analysis of the parents showed functional *TNXB* gene copies in all four assembled haplotypes. Further inspection of the sequence revealed that the indel occurred within a homopolymeric stretch, a region where long-read PacBio sequencing may have a reduced accuracy^{221–224}. The combination of indel in homopolymeric sequence combined with intact genes in both parents suggests that this variant may be more likely the result of a sequencing artifact; *TNXB* was therefore treated as functional in this haplotype. This indel is not present within the assembly of HG00514.hap2 (version 2).

Complex structural polymorphisms

Contributing authors: Peter A. Audano, Christine R. Beck, Evan E. Eichler, Miriam K. Konkel, Charles Lee, Mark Loftus, David Porubsky, Feyza Yilmaz

Methods

SMN analysis

We start by extracting the FASTA sequence across all phased genome assemblies from this study as well as from previously published studies (HPRC and T2T-primate). For this we aligned each assembly to the reference genome T2T-CHM13 using minimap2⁷⁸ (v2.24). Then we use rustybam (v0.1.33, 10.5281/zenodo.8106233) ‘liftover’ function to subset alignments in PAF format to a desired region (chr5:70300000-72100000). Then we used query coordinates in each subsetting PAF file to extract the query FASTA sequence from each assembly. Next, in order to evaluate the copy number of each multicopy gene (*SMN1/2*, *SERF1A/B* and *GTF2H2/C*), we extract the sequence of each gene from T2T-CHM13 based on the available gene annotation (https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=T2T/CHM13/asssemblies/annotation/chm13v2.0_RefSeq_Liftoff_v5.1.gff3.gz). Then, we align the full gene sequence as well as exonic sequence only to each of the extracted FASTA files. Next we on

each FASTA we run DupMasker²²⁵ (v4.1.2-p1) annotation to define ancestral duplicons in each sequence for visualization purposes.

To assign a specific SMN copy to each haplotype, we extracted FASTA sequence from SMN exon regions for each haplotype ($n=101$) and concatenated them into a single sequence. We constructed multiple sequence alignment using the R package DECIPHER²²⁶ (v2.28.0). We then calculated the distance among all haplotypes and constructed an UPGMA tree using R package ape²²⁷ (v5.7-1) and phangorn²²⁸ (v2.11.1). We set the orangutan sequence as an outgroup. Finally, we split all human haplotypes into two groups representing *SMN1* and *SMN2* gene copies.

Additionally, we constructed a custom RepeatMasker⁹¹ (v4.1.2) library containing the canonical exon sequences of *GTF2H2* (ENSG00000145736.15), *NAIP* (ENSG00000249437.8), *SERF1A* (ENSG00000172058.16), and *SMN1* (ENSG00000172062.17). Exon sequences were sourced from the Ensembl genome browser¹⁰⁸ (release 111). Utilizing RepeatMasker and this custom library, we identified all low divergence (<5%) gene exon hits within our haplotype assemblies. Exon sequences were retrieved using SAMtools⁹³ (v1.15.1), then stitched together to form the full gene exonic sequence. *SMN* gene copies were manually curated as either *SMN1* or *SMN2* based on carriage of either the C or T nucleotide in exon 7.

Amylase haplotype detection using Verkko and optical genome mapping assemblies

We characterized the amylase locus, a ~212.5 kbp region on chromosome 1 (GRCh38; chr1:103,554,220–103,766,732), in 65 genome assemblies (**Supplementary Tables 3, 55**). To resolve haplotype structures at this locus, we compared Verkko and optical genome mapping (OGM) datasets, as described previously^{56,229}. Fragmented OGM assemblies (HG00268, HG01596, HG02011, HG02018, HG03456, and NA19983) were reprocessed using a local de novo assembly approach. The step-by-step process of this approach involves: 1-collecting unaligned molecules and molecules which aligned to amylase locus; 2-running de novo assembly using the molecules from step 1 and aligning the assembly to GRCh38p.12:

```
python2.7 Solve3.5.1_01142020/Pipeline/1.0/pipelineCL.py -T 64 -U -j
64 -jp 64 -N 6 -f 0.25 -i 5 -w -c 3 \

-y \

-b ${bionano_bnx} \

-l ${output_dir} \

-t Solve3.5.1_01142020/RefAligner/1.0/ \

-a
Solve3.5.1_01142020/RefAligner/1.0/optArguments_haplotype_DLE1_saphyr
_human.xml \
```



```
-r ${reference_genome}
```

To compare the Verkko amylase locus haplotype structures to OGM, we first converted fasta files of Verkko assemblies to *in silico* maps using fa2cmap_multi_color.pl (Bionano Solve™ v3.5.1). The resulting *in silico* maps and de novo assemblies of OGMs were aligned to the GRCh38p.12 reference assembly *in silico* map using the refAligner tool (Bionano Solve™ v3.5.1).

```
Solve3.5.1_01142020/RefAligner/1.0/RefAligner
```

```
-ref ${referenceFile} \ -maxthreads ${threads} \
```

```
-i ${inputfile} \
```

```
-o ${outputdir}EXP_REFINEFINAL1
```

Next, OGM contigs that aligned to amylase locus were pairwise aligned to Verkko contigs using the same alignment tool, refAligner [Bionano Solve™ (v3.5.1)]. Pairwise alignment was performed between Verkko hap1 and hap2, and OGM hap1 and hap2 contigs using refAligner alignment results. This process generated two separate alignment files for each sample. The variant files and alignment confidence scores produced by these alignments were used to determine which Verkko haplotype matched with which OGM contig. Matching pairs were validated by visual inspection using the Bionano Access™ software. Amylase haplotypes were characterized from these visualizations using amylase segments and *in silico* map labeling patterns as a guide. The amylase segments were determined as described previously⁵⁷. Segments depicted by colored arrows were determined based on the labeling pattern from the Bionano Genomics optical genome mapping data and the sequence similarity between segments using blastn (BLASTN v2.9.0+)²³⁰. First, using optical genome mapping *in silico* map labeling patterns of GRCh38p.12 reference assembly, amylase segments were identified using pairwise alignments between smaller duplicons within segmental duplications. Then, to obtain accurate coordinates of segments, the GRCh38p.12 fasta sequence of the locus was aligned against itself using blastn (BLASTN 2.9.0+). The alignments with less than 99% sequence similarity and 60% coverage were filtered out and coordinates of all reference segments were finalized. We identified one copy of AMY2B (purple; 45.7 kbp), two copies of Intergenic1 (green; 5.3 kbp), one copy of AMY2A.1 (red; 12.4 kbp), two copies of AMY2A.2 (orange; 14.1 kbp), three copies of Intergenic2 (maroon; 4.3 kbp), and three copies of AMY1 (pink; 26.3 kbp) segments within the GRCh38p.12 amylase locus. Next, the GRCh38p.12 segment sequences were used as a guide to detect amylase segments from each haplotype included in this study using blastn. Same sequence similarity (> 99%) and coverage thresholds (> 60%) were used to obtain the amylase segment coordinates from each haplotype. Finally, all amylase haplotypes were visualized using pygenomeviz (<https://github.com/moshi4/pyGenomeViz>) to confirm haplotype structures, and haplotypes were clustered using pgr-tk⁴⁵ (v0.5.1).

Amylase haplotype rarefaction analysis

Rarefaction analysis was performed using the R package iNEXT²³¹ to evaluate haplotype diversity and determine whether the number of samples included sufficiently captures majority of

diversity at the amylase locus. The following command was used for all amylase haplotypes ($n=130$ alleles):

```
Data <- read.table("rarefaction_allPopsCombined.txt", header =TRUE,
sep = "\t")[6,]

Obj <- iNEXT(as.numeric(Data), datatype="abundance",
q=0,knots=sum(Data)*2)

Temp_data <- Obj$iNextEst$size_based

Rate <- c()

for (i in 1:nrow(Temp_data)){Rate[i+1] <-
Temp_data[i+1,5]-Temp_data[i,5]}

Plot_data <- data.frame(NumChr=Temp_data$m, NumHap=Temp_data$qD,
Rate=Rate[-length(Rate)])

pdf("Sample-size-based_REcurve_with_rate_of_change.pdf",
width=6,height=4)

twoord.plot(Plot_data$NumChr,Plot_data$NumHap,Plot_data$NumChr,Plot_d
ata$Rate, type="l", xlab="Number of alleles",ylab="Haplotype count",
rylab="Rate of change", lytickpos=seq(0,70,10),
lcol="blue3",rcol="red2")

abline(v=sum(Data), lty="dashed")

dev.off()
```

Results

There are ~60 genomic disorders associated with neurodevelopmental delay, autism, epilepsy and intellectual disability where recurrent CNVs are associated with large complex regions of segmental duplications. The organization and orientation of the SDs is a critical risk factor in predisposing to NAHR that predispose and/or protect from CNV risk^{16,232–234}. These SD rich regions have also been a hotspot for non-allelic gene conversion that plays an important role in evolution of many multi-copy genes²³⁵. Here, we focus on a biomedically relevant and structurally complex region containing *SMN1* and *SMN2* gene copies. *SMN1* is the primary gene associated with spinal muscular atrophy (SMA) and its duplicated copy, *SMN2*, which modifies SMA severity and is a target of one of the most successful ASO-mediated gene therapies. The genes are embedded in a very large SD region (~1.5 Mbp) that has been almost impossible to fully sequence and assemble over the last two decades^{1,2,8,236} (**Supplementary Fig. 54**). In addition to recently released HPRC assemblies, we were able to fully assemble the *SMN1/2*

region in approximately two-thirds of the haplotypes in this study, providing us a unique opportunity to investigate the structural complexity of this region (**Fig. 5b**). We characterized this combined set of 101 complete human assemblies of this region that allows us to map copy number (full or partial gene alignments) of *SMN1/2* genes and detailed structure of this region. Focusing on multi-copy genes *SMN1/2*, *SERF1A/B* and *GTF2H2/C* we find that about half ($n=48$) of all haplotypes carry exactly two copies (aligned to two distinct regions in a single haplotype) of each of these genes. On the other hand *NAIP* gene is present mostly in a single copy (gene alignment per haplotype). (**Fig. 5c**). We highlight 11 human haplotypes including orangutan that carry distinct gene copy numbers and are organized into structurally diverse duplication structures composed of multiple palindromic regions (**Fig. 5d**). We further defined functional *SMN1* copy and duplicated *SMN2* copy in each haplotype using orangutan ancestral *SMN1* copy as an outgroup (**Supplementary Fig. 55**). For individuals where both haplotypes are fully assembled ($n=31$) we report 100% concordance with the short-read based genotyping methods (**Supplementary Fig. 56**). We have assigned 98 haplotypes to carry ancestral *SMN1* copy and 3 haplotypes that do not (**Fig. 5e**, **Supplementary Fig. 57**). We speculated that such haplotypes are likely to occur via gene conversion. Evaluation of an extent of directly oriented and highly identical repeats we define 2 haplotypes that are likely predisposed to CNVs via NAHR.

The amylase locus in the human genome spans approximately 212.5 kilobase pairs (kbp) and is located on chromosome 1 (GRCh38; chr1:103,554,220–103,766,732). This locus contains the *AMY2B*, *AMY2A*, *AMY1A*, *AMY1B*, and *AMY1C* genes (**Supplementary Fig. 58**). We constructed haplotype-resolved diploid assemblies for 65 individuals, resulting in a total of 130 sampled alleles (haploid sample size; **Supplementary Table 55**). By analyzing copy number and orientation of the amylase segments, we identified 35 distinct amylase haplotypes (**Supplementary Fig. 58**, **Supplementary Table 55**). These haplotypes were supported by both Verkko and optical genome mapping de novo assemblies. To classify these haplotypes, we adhered to the established nomenclature⁵⁷. The notation HXAYBZ represents the copy number of *AMY1*, *AMY2A*, and *AMY2B* genes, respectively, with superscript “a” indicating ancestral and superscript “r” indicating reference haplotypes (**Supplementary Fig. 58**). Notably, this study described the largest number of base-pair resolution amylase haplotypes described to date. The length of these amylase haplotypes varies, ranging from 111 kbp (H1^a.1 and H1^a.2) to 582 kbp (H11.1) (**Supplementary Fig. 58**), capturing those that are structurally identical to the GRCh38 (H3^r.1) and the T2T-chm13 (H7.3) assemblies. Among the identified haplotypes, four common ones, H1^a.1 ($n = 14$), H3^r.1 ($n = 13$), H3^r.2 ($n = 19$), and H3^r.4 ($n = 22$), constitute approximately 57% of all amylase haplotypes. Notably, 23 out of 35 distinct amylase haplotypes are singletons. Furthermore, this study provides base-pair level resolution for nine haplotypes that were previously supported only by optical genome mapping data. Finally, we introduce a novel haplotype, H11. 1, which also has base-pair level resolution.

To assess how effectively the 65 genomes capture the amylase locus within the population, we performed rarefaction and extrapolation analyses based on 39 unique amylase haplotypes, including both sequence-resolved and OGM-based haplotypes, derived from the 130 alleles examined (**Supplementary Fig. 59**). This analysis indicates that the number of distinct haplotypes saturates as more alleles are sampled, with the rate of discovery being relatively steady (0.17), meaning new haplotypes are being identified at a predictable rate per additional allele analyzed. Based on this, we estimate that the current sample of 130 alleles captures

~83% of the haplotypes in the population, which is a substantial representation, though further sampling could help refine this estimate and potentially capture additional rare haplotypes. Therefore, 65 genomes provide a robust representation of the amylase haplotypes.

Centromeres

Contributing authors: Evan E. Eichler, Miriam K. Konkel, Mark Loftus, Glennis A. Logsdon, Keisuke K. Oshima

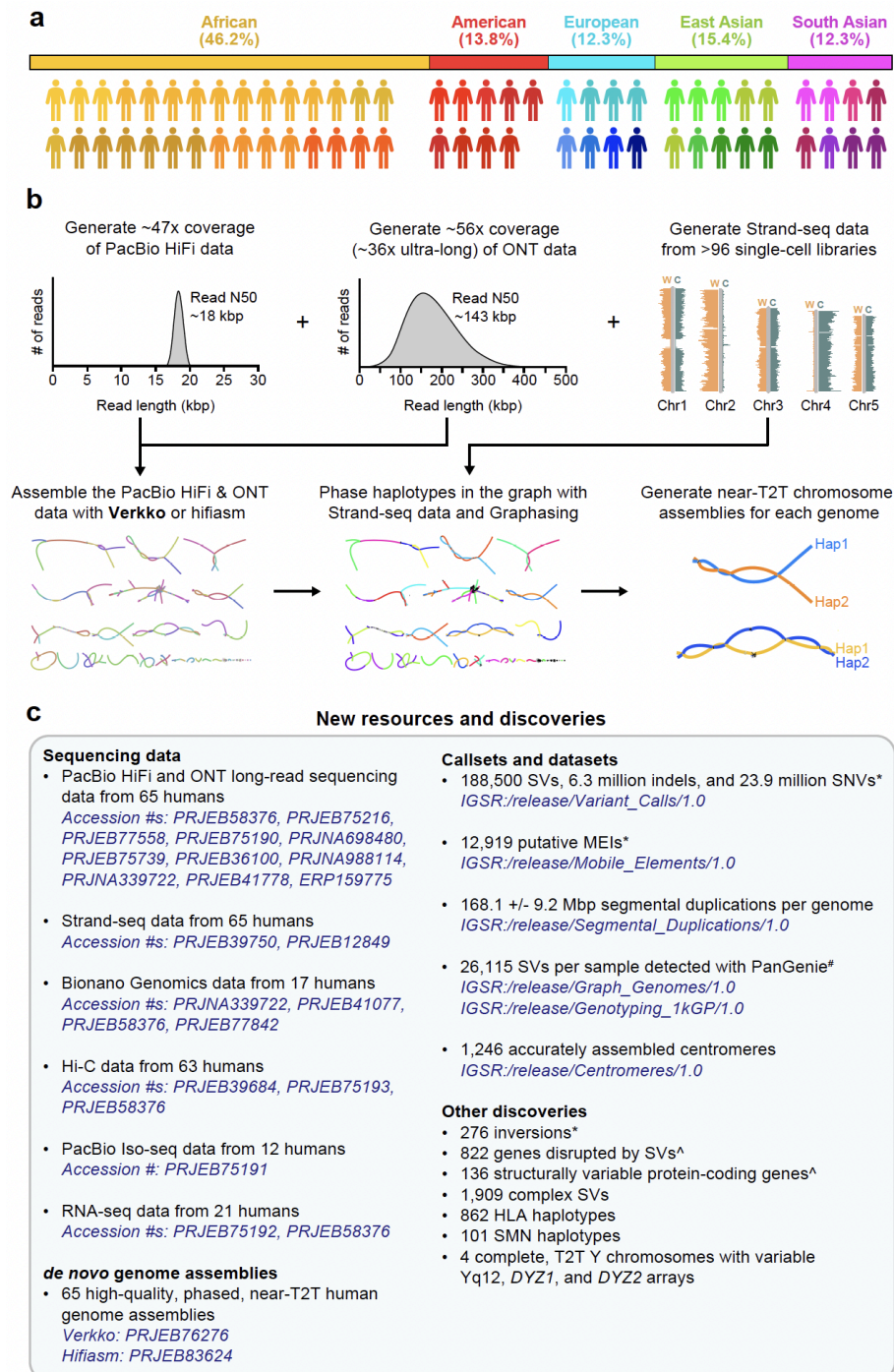
Methods

Chromosome 2 α -satellite HOR array MEI network analysis

The most prevalent LINE/L1 insertion (L1HS variant 1) and its duplications within the chromosome 2 centromeric α -satellite HOR array (**Extended Data Fig. 5e**) were selected to show how MEIs could be utilized to tag α -satellite HOR expansion events within the arrays. The sequence of this MEI and its duplications were pulled from the genome assemblies using SAMtools⁹³ (v1.15.1) and then a multiple sequence alignment was performed using MUSCLE²³⁷ (v3.38.31). Next, the similarity between all sequences (single-nucleotide variants found in only one sequence, or an indel within a homopolymeric region (e.g., a-tail), were dropped) was then calculated using the hamming distance (i.e., the proportion of disagreeing components) between the two aligned sequences using SciPy (v1.11.1; `scipy.spatial.distance.hamming`)¹⁵¹. Identical sequences were grouped and represented as a single SNV profile node within an undirected network using NetworkX²³⁸ (v3.1). A weighted edge was placed between nodes which were the most similar to one another (best match and second best match only; ties are allowed). Edges were weighted based on sequence similarity (i.e., a higher weight between nodes with more similar SNV profiles). An asynchronous label propagation algorithm (`asyn_lpa_communities` as implemented in NetworkX) was then utilized for community detection within the network.

Supplementary Figures

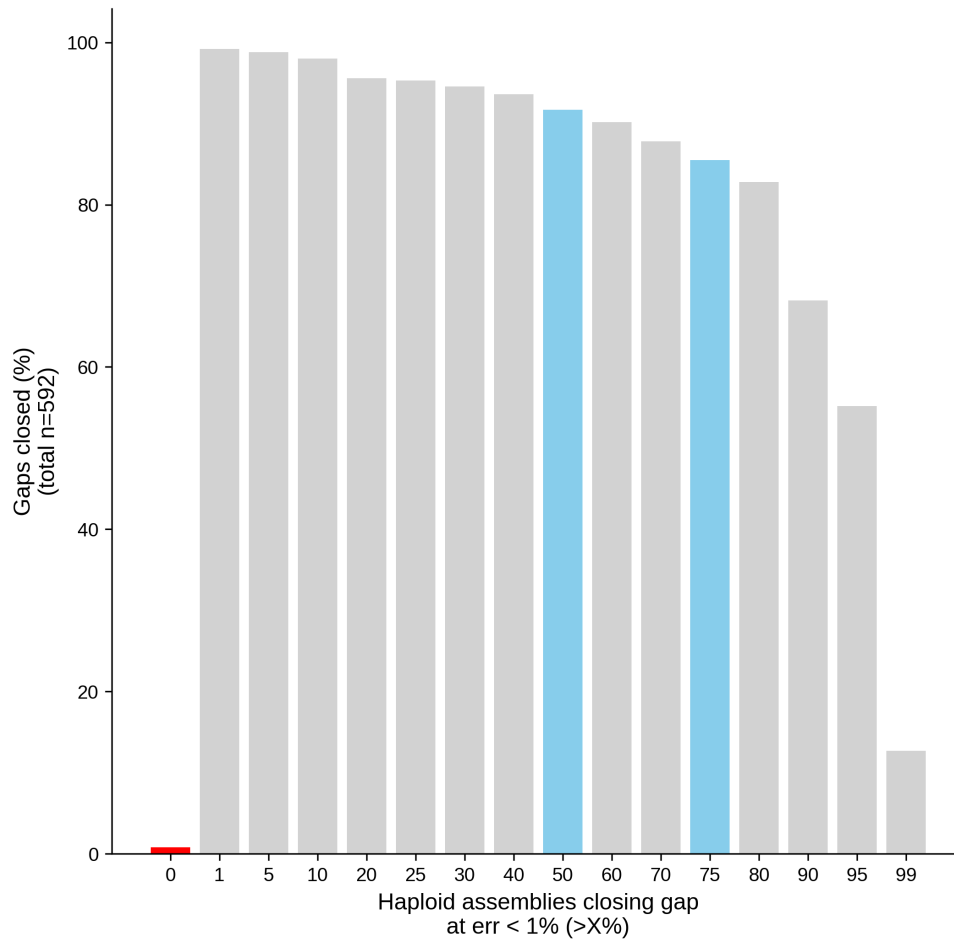
Fig. 1



Supplementary Figure 1. Overview of the individuals sequenced, genome assembly workflow, resources generated, and discoveries made in this study. a) Continental and population groups of the 65 individuals sequenced with long- and short-read sequencing technologies. The color designations are identical to those in Figure 1a. **b)** Workflow for

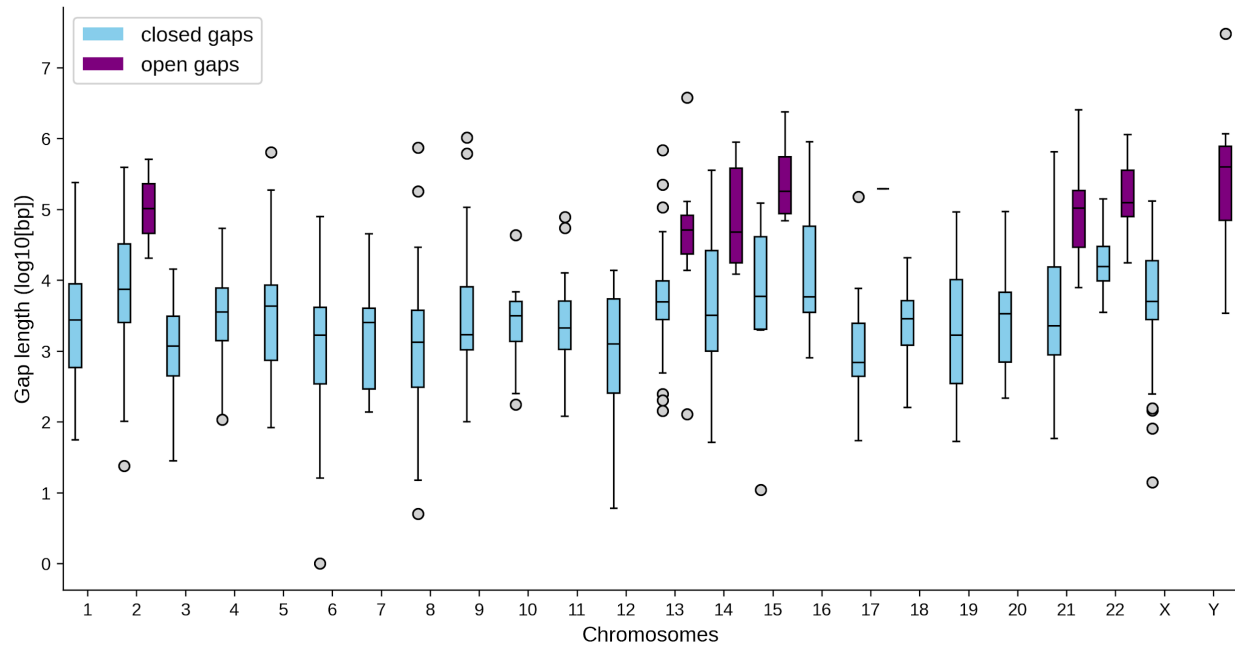
generating near-telomere-to-telomere (T2T) genome assemblies from 65 diverse humans. **c)** Resources generated and discoveries made in this study. *, relative to the T2T-CHM13 v2.0 reference genome; ^, relative to the GRCh38 reference genome; #, relative to a pangenome graph with HGVC3 and HPRC genome assemblies.

Fig. 2



Supplementary Figure 2. Closed HPRC gaps. Bar chart depicting the percentage of closed HPRC gaps (vertical axis) ² at increasingly stringent thresholds for the percentage of haploid assemblies closing the gap (horizontal axis). Highlights are shown for gaps that are never closed (red bar, $n=5$, ~0.8%), for gaps that are closed in more than 50% of all haploid assemblies (left blue bar, $n=543$, ~91.7%) and for gaps that are closed in at least 75% of all haploid assemblies (right blue bar, $n=506$, ~85.5%). For this summary, the error rate of the contigs closing the gap was set to less than 1% (“err < 1%”).

Fig. 3



Supplementary Figure 3. Size distribution of HPRC gaps. Boxplots of gap lengths (vertical axis, log10) stratified by chromosome (horizontal axis) and status [closed (blue) and, if applicable, open (purple)] of previously reported HPRC gaps ($n=592$)². In this summary, gaps were counted as “closed” at an error rate of <1% (**Methods**) if they were labeled as closed in more than 50% of all haploid assemblies and “open” otherwise. Boxplot whiskers depict 1.5*(interquartile range); outliers beyond that range appear as gray points; the horizontal line indicates the median.

Fig. 4



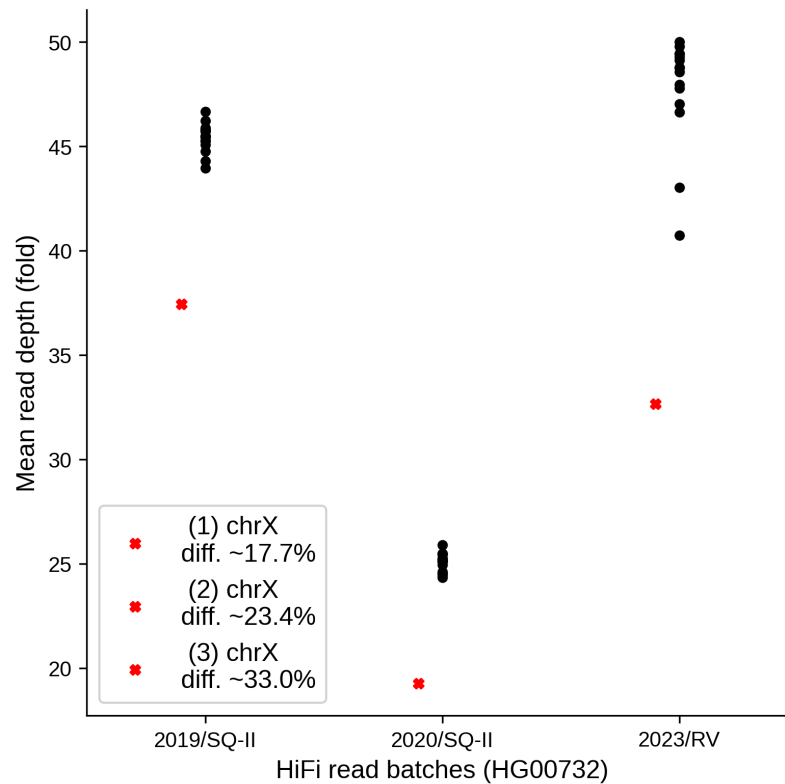
Supplementary Figure 4. Verkko genome assembly quality evaluation with PacBio HiFi raw reads. Boxplot representation of each evaluated parameter for 65 genome assemblies with two haplotypes from the Verkko assembler. Evaluation was performed using Inspector v1.2⁷⁴ with PacBio HiFi raw reads. Results were grouped in superpopulation. The boxplot whiskers depict 1.5* interquartile range, outliers beyond that range appear as grey circles; the horizontal line indicates the median.

Fig. 5



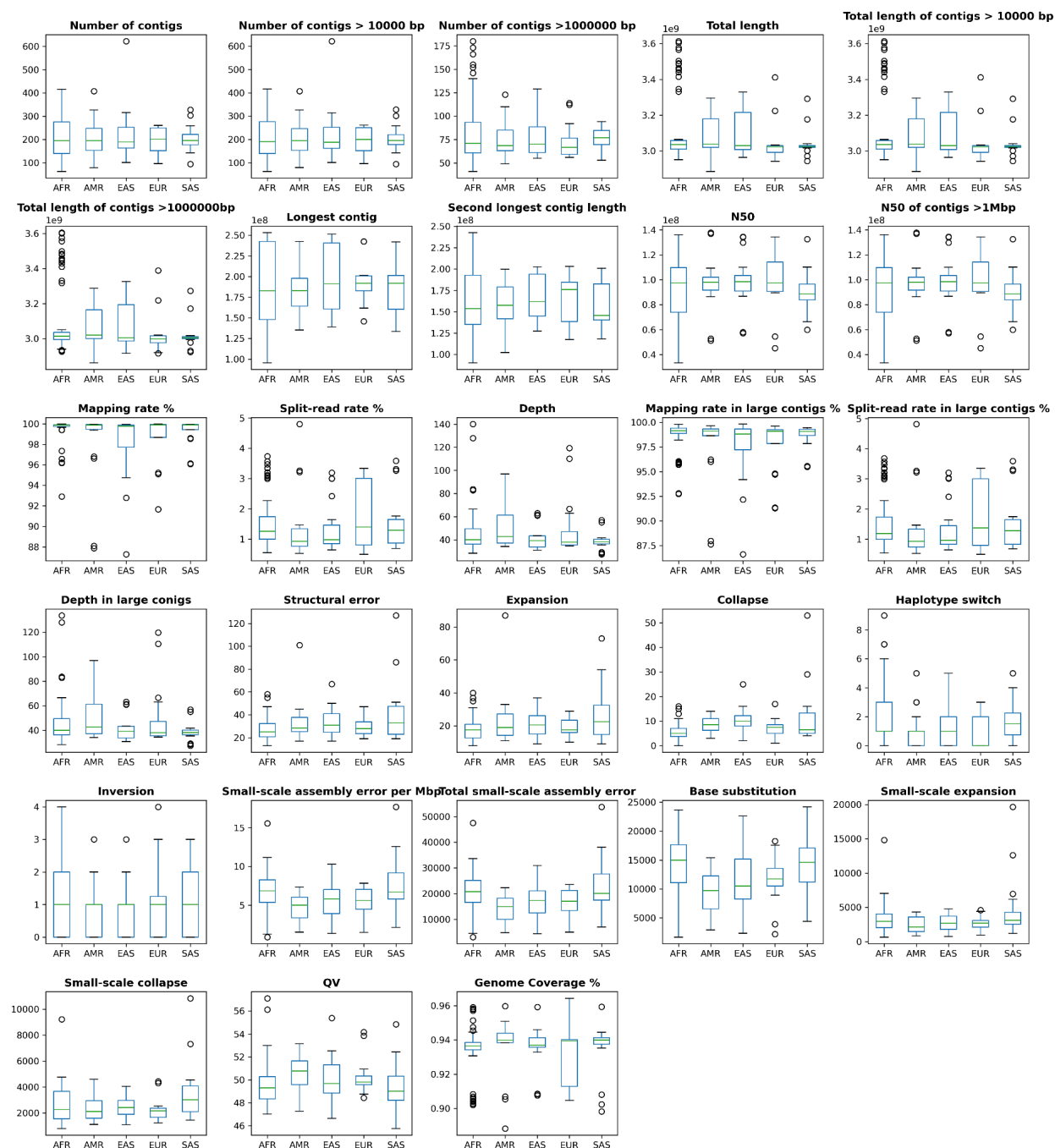
Supplementary Figure 5. Verkko genome assembly quality evaluation with ONT raw reads. Boxplot representation of each evaluated parameter for 65 genome assemblies with two haplotypes from the Verkko assembler. Evaluation was performed using Inspector v1.2⁷⁴ with ONT raw reads. Results were grouped in superpopulation. The boxplot whiskers depict 1.5* interquartile range, outliers beyond that range appear as grey circles; the horizontal line indicates the median.

Fig. 6



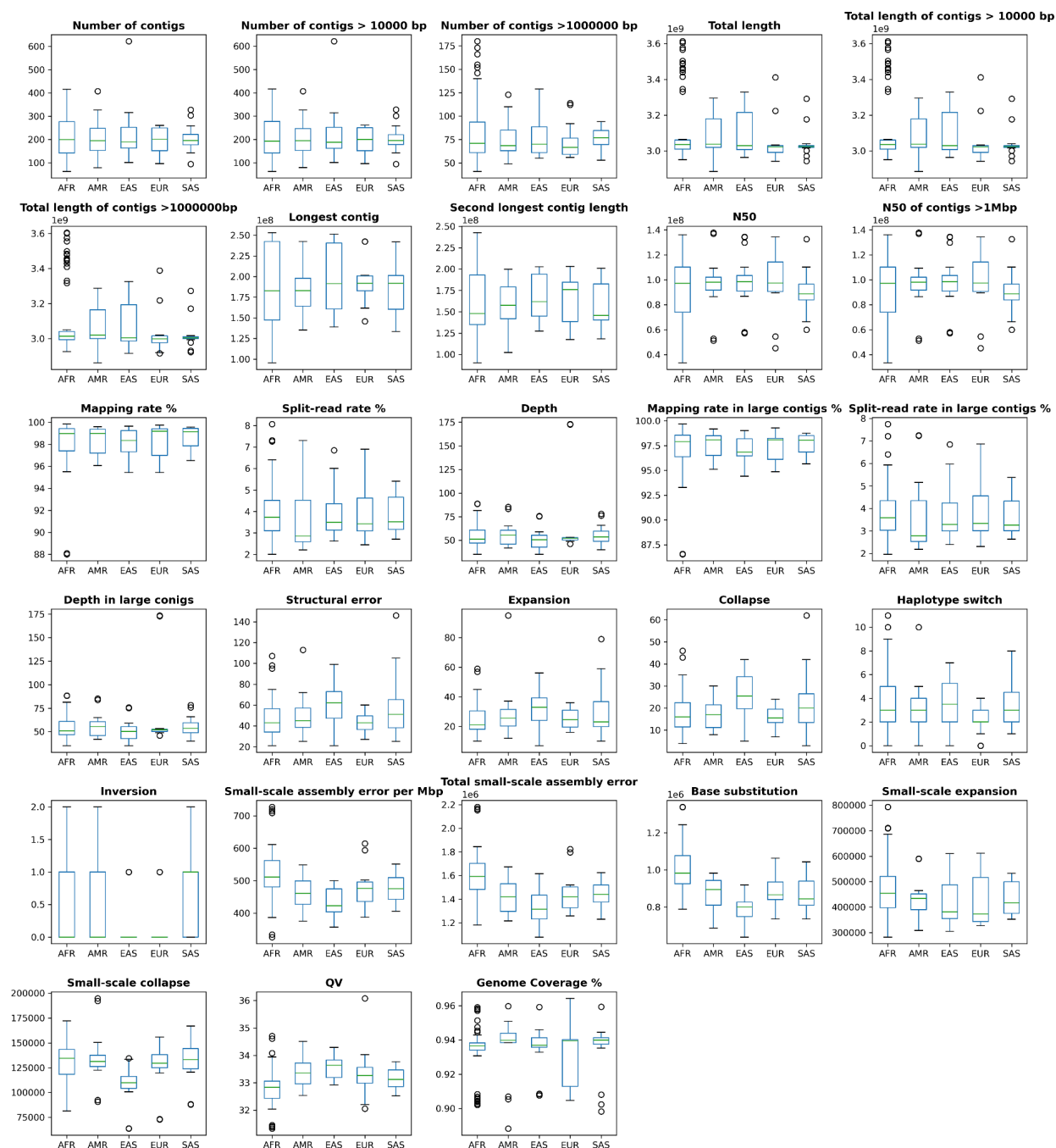
Supplementary Figure 6. PacBio HiFi read coverage dropout over chromosome X in HG00732. Summary of mean PacBio HiFi read fold coverage stratified by sequencing batch/year (horizontal axis) over main reference chromosomes (acrocentric chromosomes and chromosome Y excluded). The mean read depth for chromosome X is highlighted as a single red “x” per batch/year. The relative difference of the X-chromosomal read depth compared to all other chromosomes is stated in the legend. SQ-II: PacBio Sequel-II platform; RV: PacBio Revio platform.

Fig. 7



Supplementary Figure 7. Hifiasm assemblies quality evaluation with PacBio HiFi raw reads. Boxplot representation of each evaluated parameter for 65 samples with two haplotypes from the hifiasm assembler. Evaluation was performed using Inspector v1.2⁷⁴ with HiFi raw reads. Results were grouped in superpopulation. The boxplot whiskers depict 1.5* interquartile range, outliers beyond that range appear as grey circles; the horizontal line indicates the median.

Fig. 8



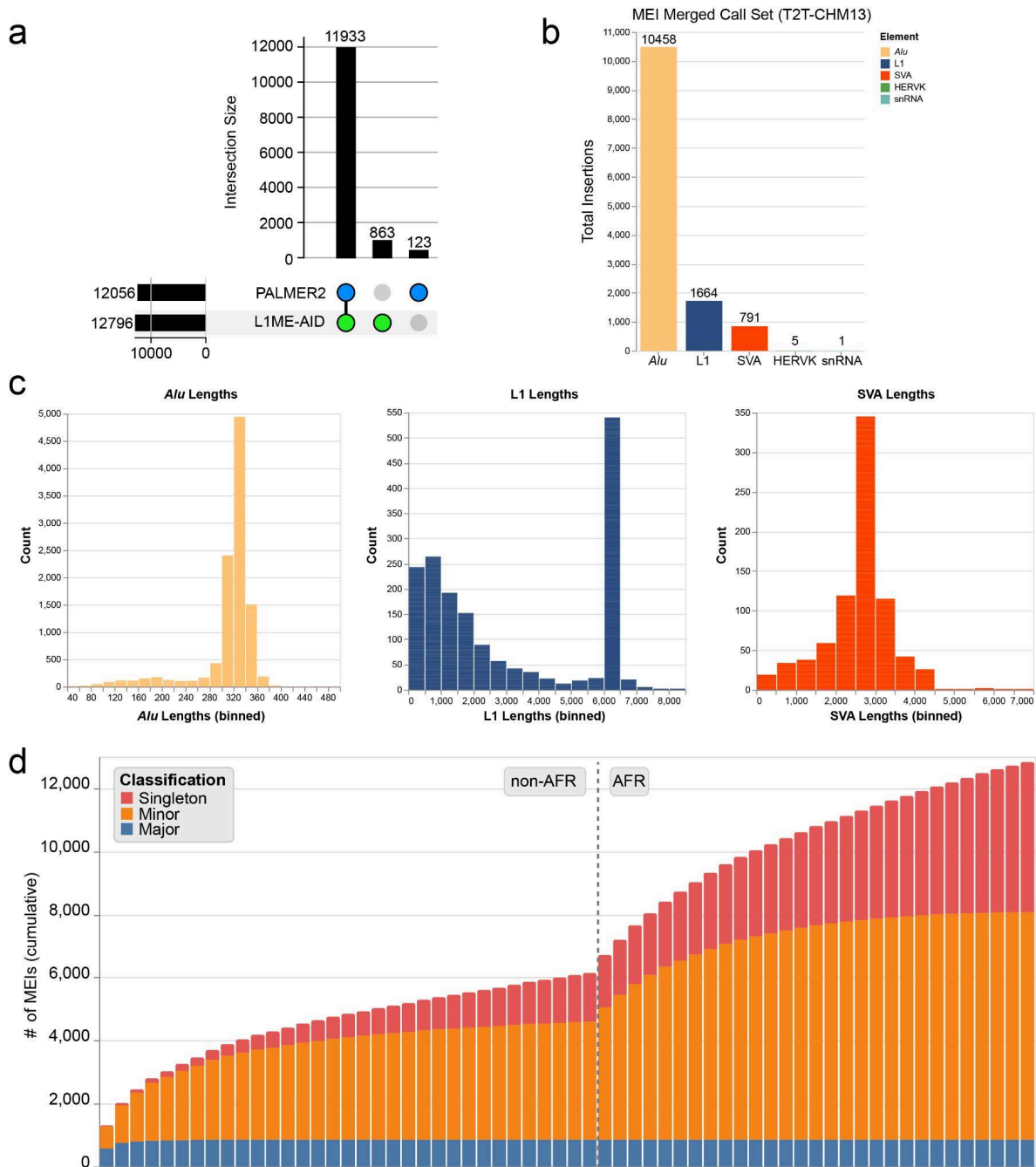
Supplementary Figure 8. Hifiasm genome assembly quality evaluation with ONT raw reads. Boxplot representation of each evaluated parameter for 65 genome assemblies with two haplotypes from the hifiasm assembler. Evaluation was performed using Inspector v1.2⁷⁴ with ONT raw reads. The boxplot whiskers depict 1.5* interquartile range, outliers beyond that range appear as grey circles; the horizontal line indicates the median.

Bar chart showing the number of variants (n) for various methods across four categories: PAV, PAV Trimmed, Covered, and No Assembly. The methods are listed on the x-axis, and the y-axis represents the number of variants. The chart shows that PAV has the highest number of variants, followed by PAV Trimmed, Covered, and No Assembly. The number of variants decreases as the methods move from left to right.

Method	PAV	PAV Trimmed	Covered	No Assembly
PAV (Vk)	5,723	0	0	0
SVIM-asm (Vk)	3,107	0	0	0
DipCall (Vk)	0	0	2,506	0
PBSV (HiFi)	0	0	1,736	0
Snif2 (HiFi)	0	0	1,329	0
Snif2 (ONT)	0	0	1,311	0
Delly (HiFi)	1,108	0	0	0
Delly (ONT)	1,108	0	0	0
DeBrk. (HiFi)	0	0	1,063	0
DeBrk. (ONT)	0	0	986	0
SVIM (HiFi)	0	0	689	0
SVIM (ONT)	0	0	626	0
CuteSV (HiFi)	0	0	555	0
CuteSV (ONT)	0	0	534	0
PAV (Vk)	461	0	0	0
SVIM-asm (Vk)	451	0	0	0
DipCall (Vk)	0	0	439	0
PBSV (HiFi)	0	0	339	0
Snif2 (HiFi)	0	0	277	0
Snif2 (ONT)	0	0	256	0
Delly (HiFi)	0	0	243	0
Delly (ONT)	0	0	213	0
DeBrk. (HiFi)	0	0	205	0
DeBrk. (ONT)	0	0	203	0
SVIM (HiFi)	0	0	199	0
SVIM (ONT)	0	0	191	0
CuteSV (HiFi)	0	0	175	0
CuteSV (ONT)	0	0	165	0
PAV (Vk)	153	0	0	0

73

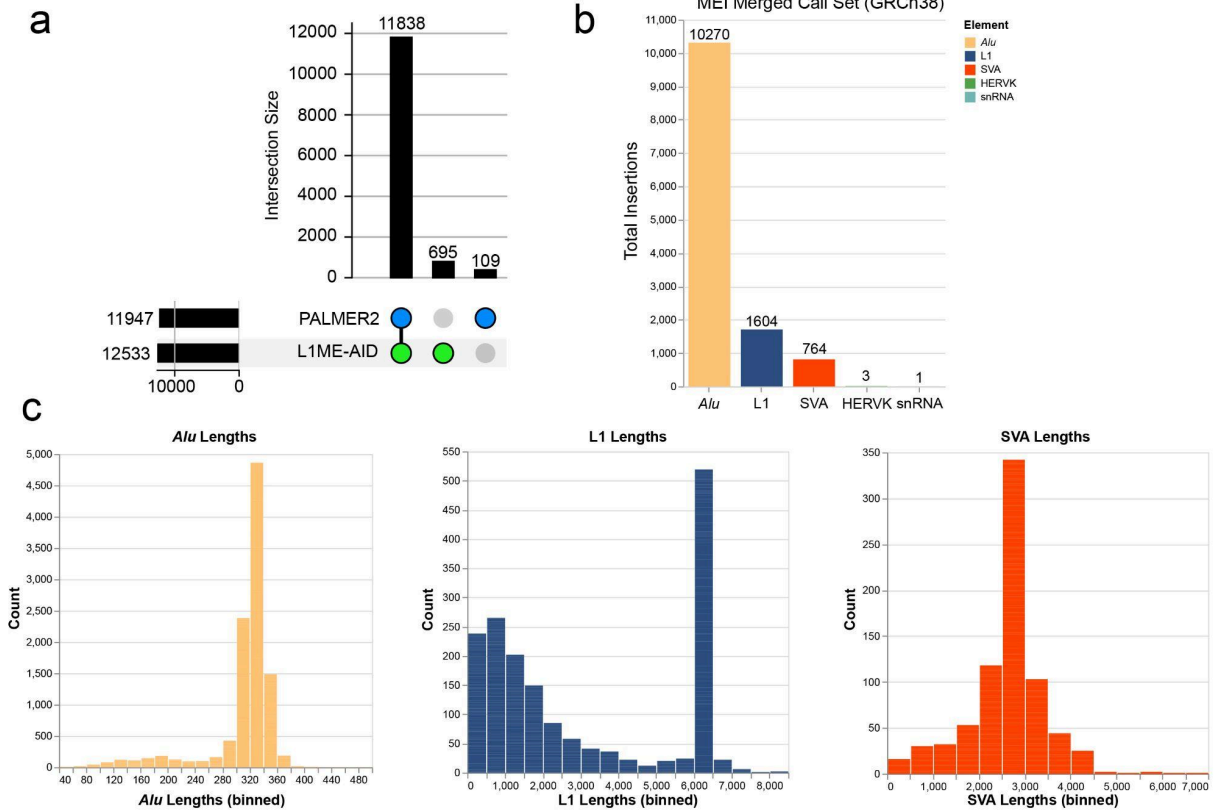
Fig. 10



Supplementary Figure 10. T2T-CHM13 mobile element insertion (MEI) merged callset. a) An upset plot showing the intersection of MEI calls across the two individual MEI callsets built against the T2T-CHM13 reference genome. Most (92.4%) of the 12,919 MEIs were identified by both callers (PALMER2 and L1ME-AID). **b)** A bar graph displaying the total *Alu*, L1, SVA, HERV-K, and snRNA insertion calls within the merged callset. **c)** Three histograms displaying the length distribution of *Alu* elements (yellow), L1s (blue), and SVAs (red) within the merged callset. **d)** Cumulative non-redundant MEIs (y-axis) across 62 diverse individuals (x-axis) in this study (children from the three trios were excluded), called with respect to the T2T-CHM13

reference genome. MEIs are classified as minor if they are found in fewer than half of the individuals and major if present in more than half. MEI detection begins to plateau until the addition of individuals from African descent.

Fig. 11

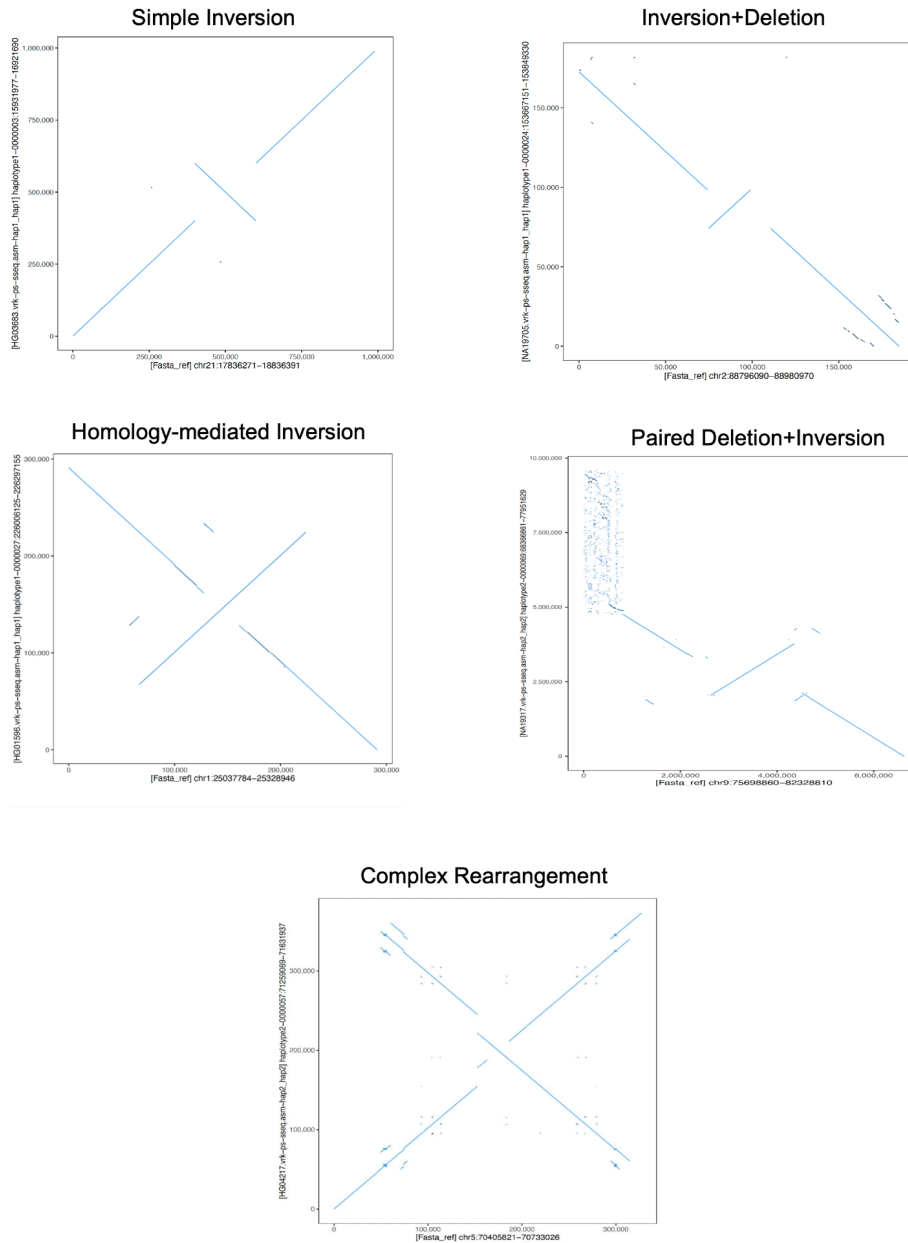


Supplementary Figure 11. GRCh38 MEI merged callset. **a)** An upset plot showing the intersection of MEI calls across the two individual MEI callsets built against the GRCh38 reference genome. Most (93.6%) of the 12,642 MEIs were identified by both callers (PALMER2 and L1ME-AID). **b)** A bar graph displaying the total *Alu*, L1, SVA, HERV-K, and snRNA insertion calls within the merged callset. **c)** Three histograms displaying the length distribution of *Alu* elements (yellow), L1s (blue), and SVAs (red) within the merged callset.

Supplementary Figure 12. MEI counts. Two heatmaps [GRCh38 (left) and T2T-CHM13 (right)] showing the total unique sample-specific and reference-specific mobile element insertions per sample. Both heatmaps show samples from African descent containing a greater total number of sample-specific MEIs.

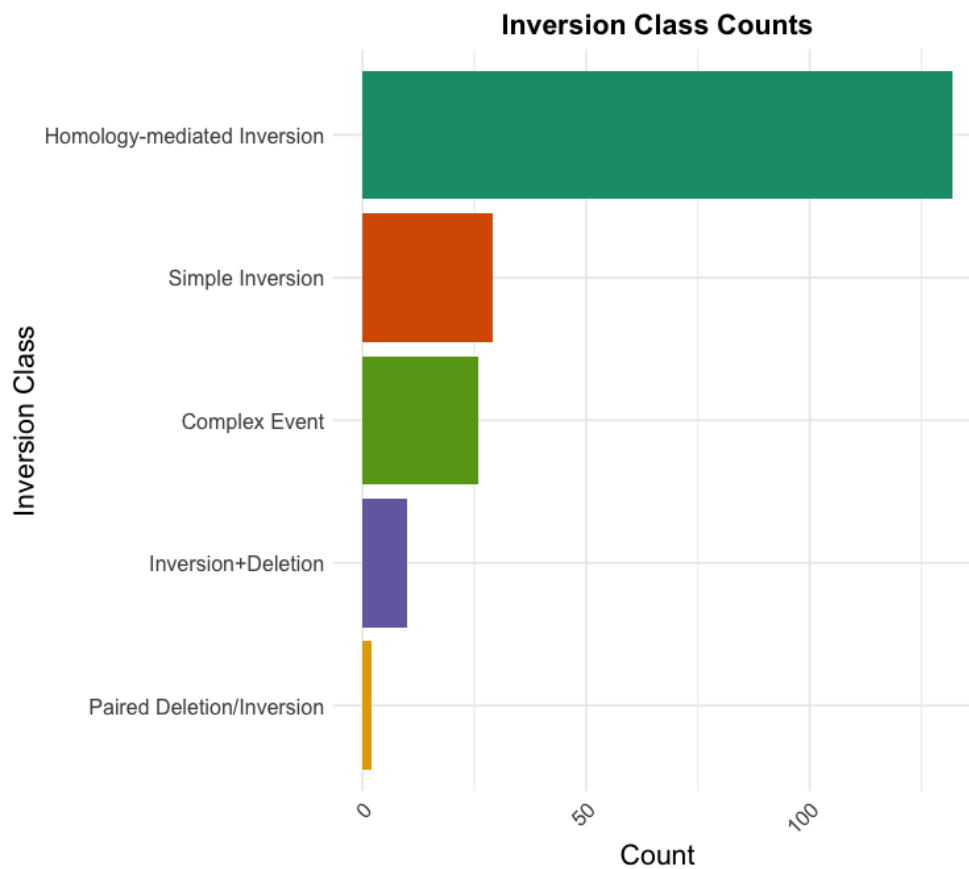


Fig. 13



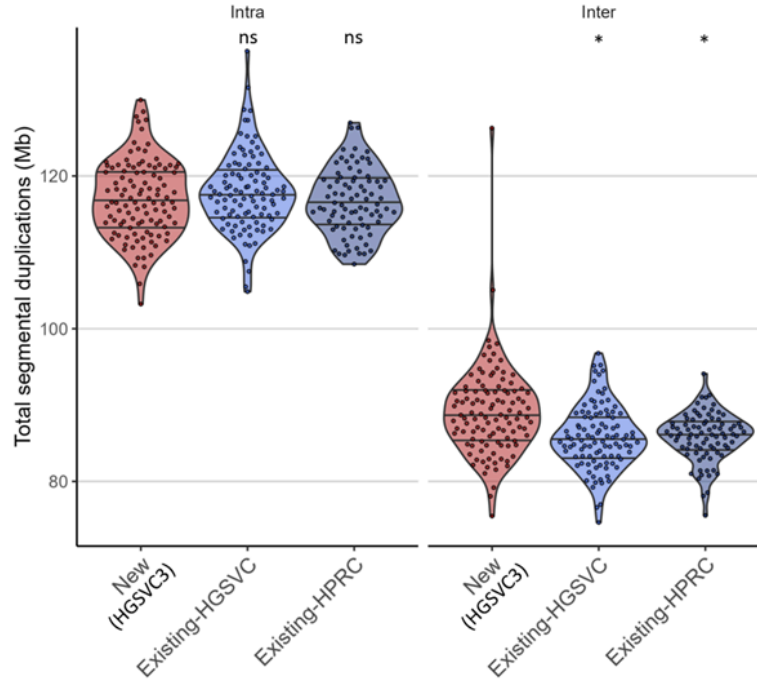
Supplementary Figure 13. Inversion dotplot analysis. Representative dot plots illustrating the categorization of different inversion classes in our dataset. The y-axis represents the specified genomic location in one of the inversion carrier samples, while the x-axis represents the corresponding region in the CHM13 reference genome. Each candidate inversion locus was manually examined through dot plot analysis, facilitating the validation of candidate inversion locations and their classification into one of the five distinct categories showcased in the graph.

Fig. 14



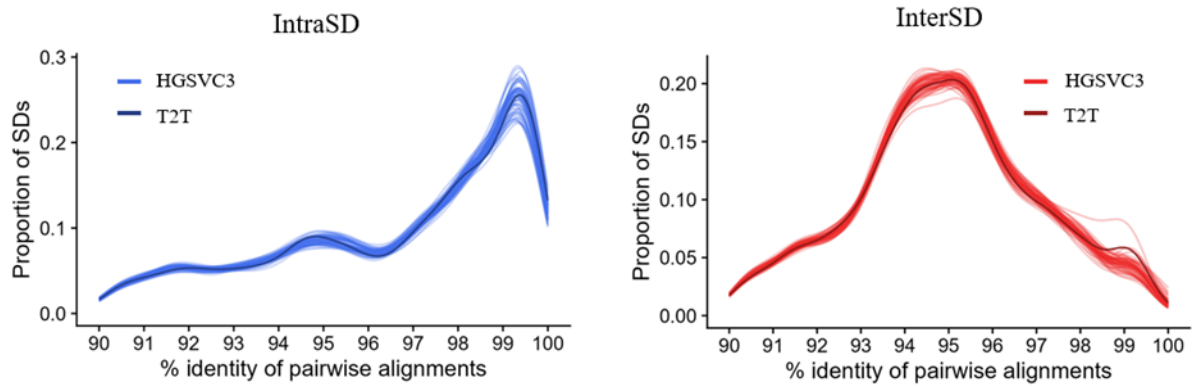
Supplementary Figure 14. Inversion classes. The bar plot represents the count of instances for each inversion class identified in the manual dotplot analysis.

Fig. 15



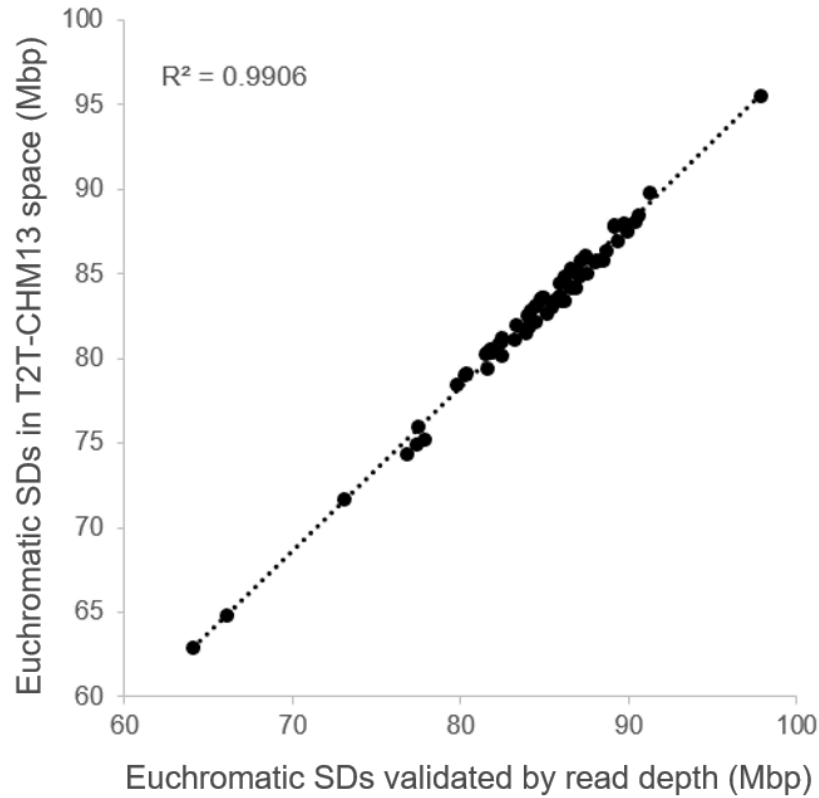
Supplementary Figure 15. Comparison between segmentally duplicated bases in this study and those in PacBio HiFi genome assemblies. Total segmental duplication (SD) content is quantified and compared between this study (HGSVC3; $n=130$), HGSVC2/3 PacBio HiFi assemblies; $n=102^{21}$, and HPRC draft genome assemblies (HPRC; $n=84$)¹. The three horizontal lines in the violin plots indicate the first quartile (top), median (middle), and third quartile (bottom). Significance (*) is shown on the top and is measured with a two-tailed Wilcoxon ranked sum test ($p=3.1e-6$ and $2.8e-6$ for Existing-HGSVC and -HPRC, respectively).

Fig. 16



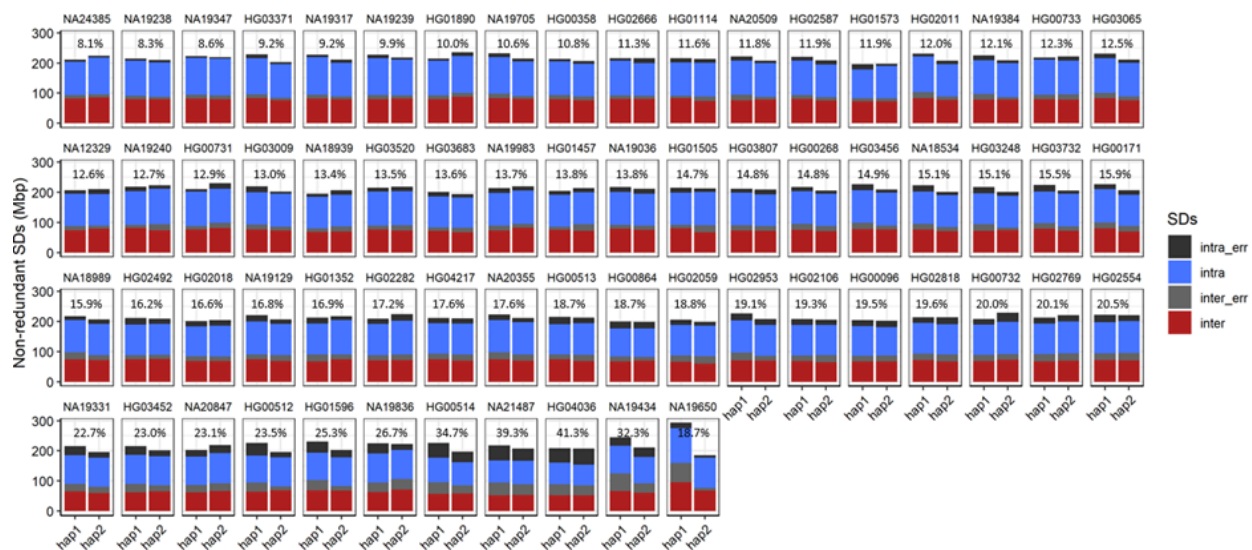
Supplementary Figure 16. Inter- and intra-SD content. Distribution pairwise identity among inter/intra-chromosomal segmental duplication pairs in the HGSVC genomes ($n=126$ haplotypes).

Fig. 17



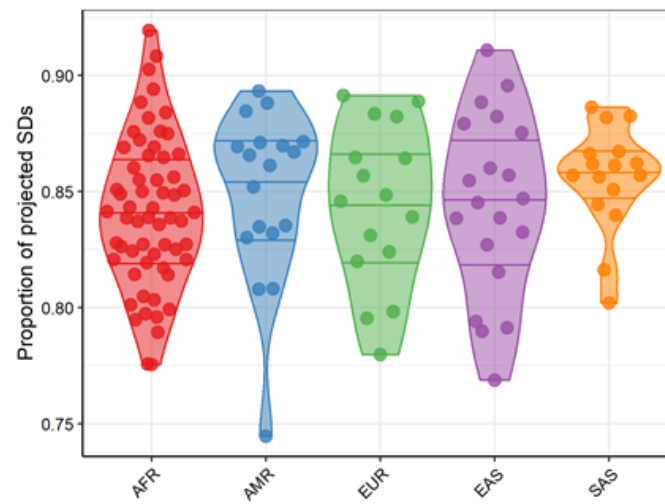
Supplementary Figure 17. Correlation between the euchromatic SDs detected with an assembly-based approach vs those validated by read depth. Plot showing the comparison between the amount of non-redundant euchromatic SDs predicted from the genome assembly versus the amount of SDs supported by read-depth analysis using Illumina WGS data from the same sample mapped to the reference, T2T-CHM13 (v1.1). Each data point represents a diploid individual ($n=64$), and an R-squared statistic is shown.

Fig. 18



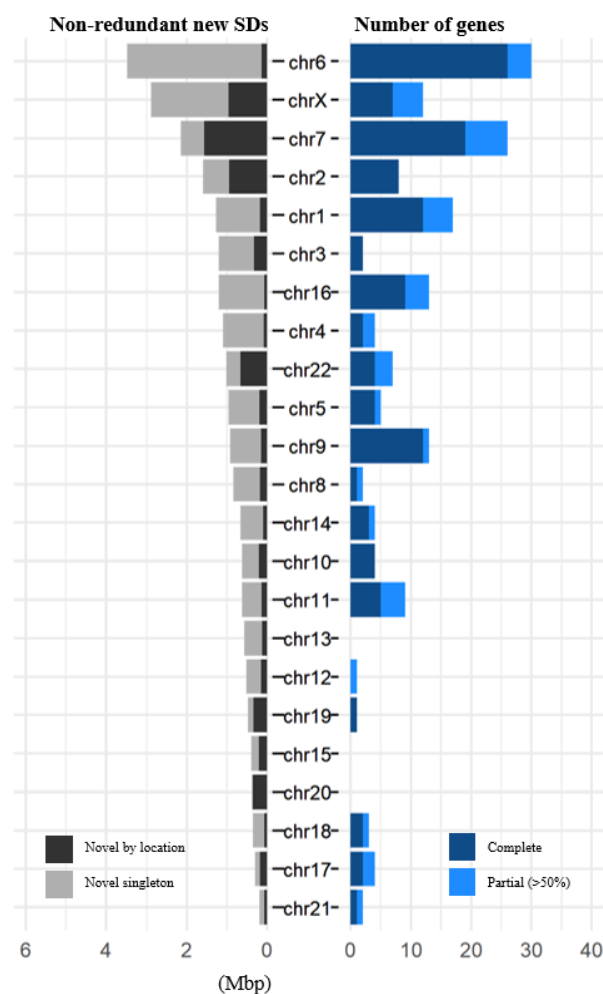
Supplementary Figure 18. Non-redundant SD content. Nonredundant bases of inter/intra-chromosomal segmental duplications and those that overlap with flagged regions via Flagger/NucFreq.

Fig. 19



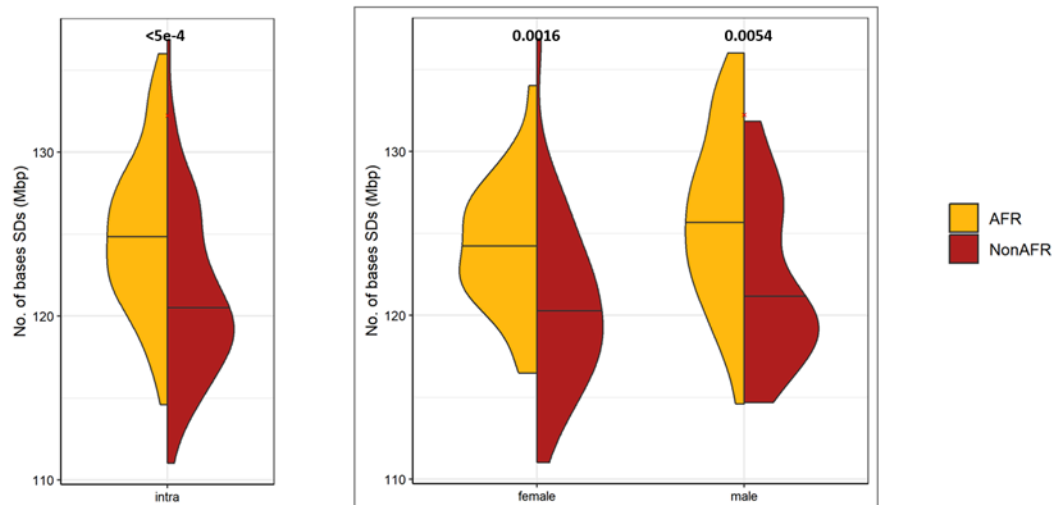
Supplementary Figure 19. Proportion of segmental duplications successfully projected onto T2T-CHM13 space.

Fig. 20



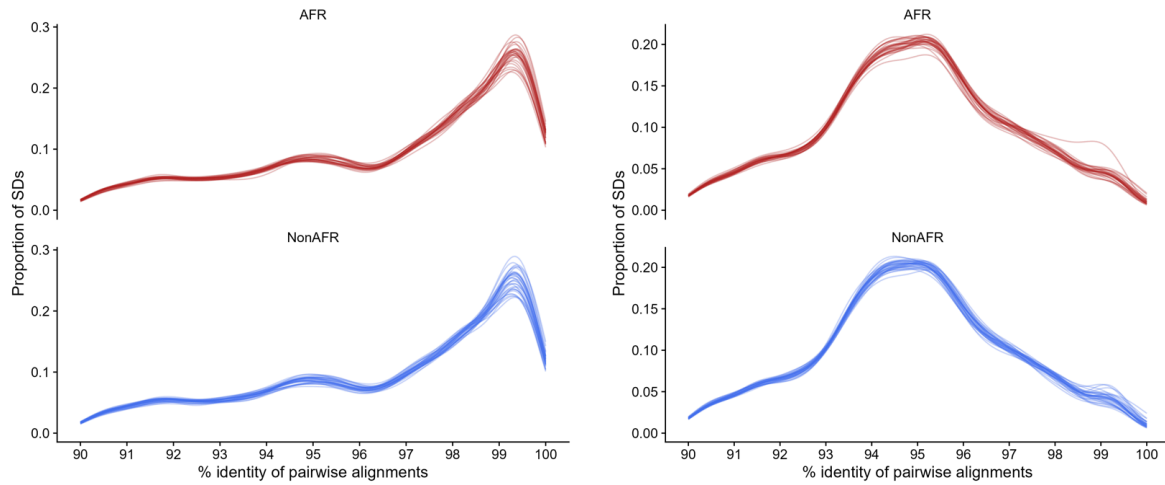
Supplementary Figure 20. New SD content. New segmental duplication content of HGSVC3 genomes compared to the segmental duplications of 170 haplotypes²¹. In the respective plots, the left shows SDs with content changes, expansion, and whether they are new (based on location/synteny), while flagging those that are singleton. On the right quantifies the number of complete (100% coverage), and partial overlaps (>50% coverage) with protein-coding genes.

Fig. 21



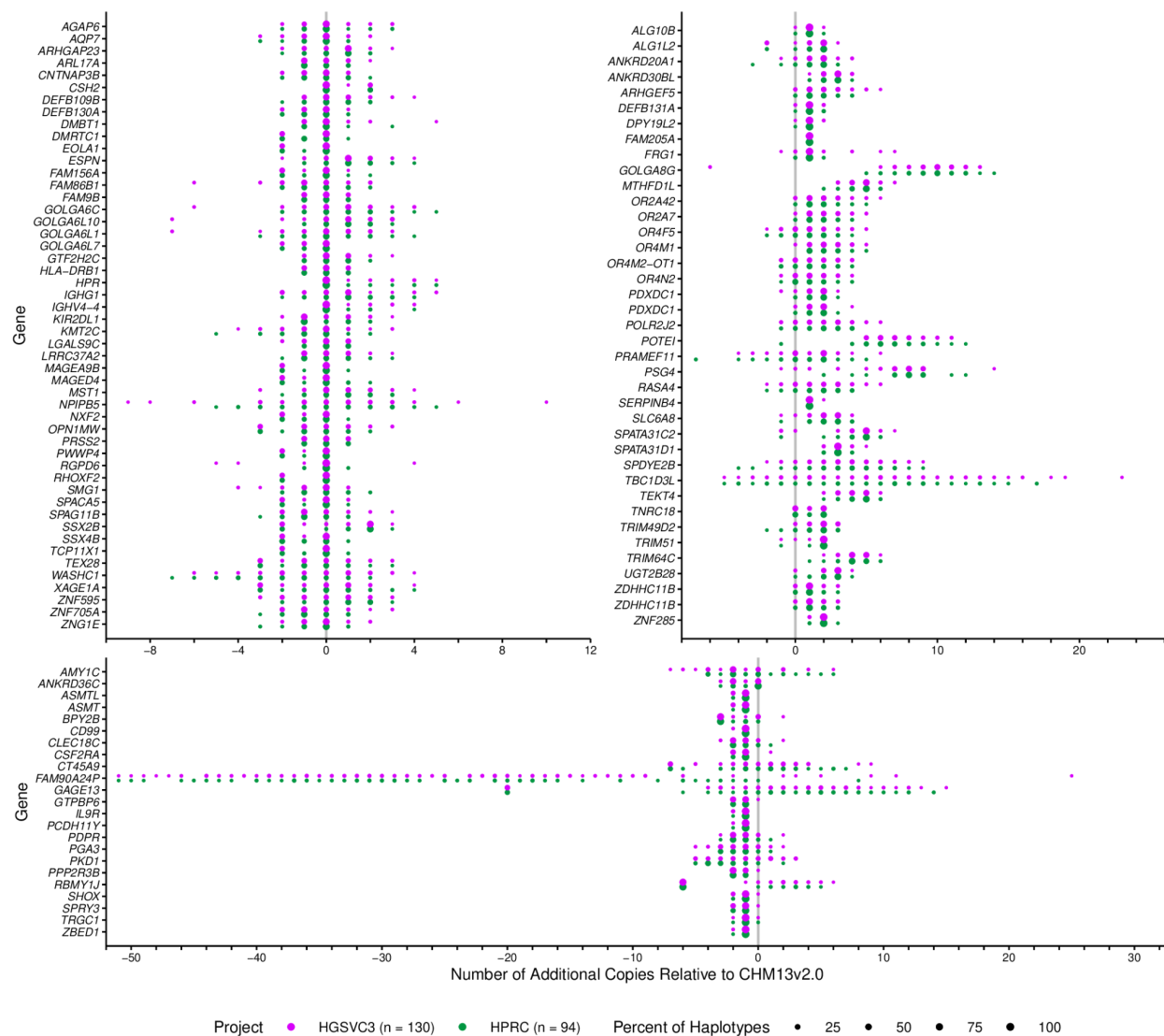
Supplementary Figure 21. Comparison between the number of intrachromosomal SDs in African (AFR) and non-African (non-AFR) genomes. We observe a significant shift in the distribution between African ($n=60$; yellow) and non-African ($n=70$; red) individuals (two-tailed Wilcoxon rank sum test); the p-value is indicated above each violin plot.

Fig. 22



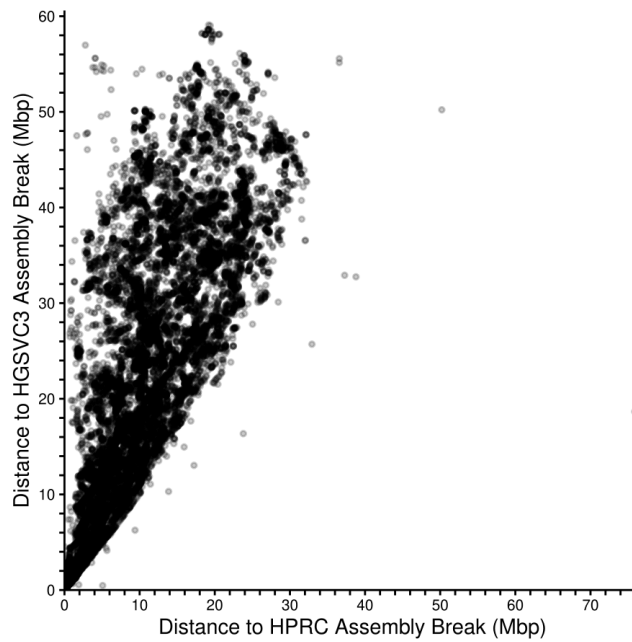
Supplementary Figure 22. Comparison of the distributions of pairwise sequence identity between African and non-African individuals among intra- and inter-chromosomal SDs. The comparison for intrachromosomal SDs is shown on the left, and those for interchromosomal SDs is shown on the right.

Fig. 23



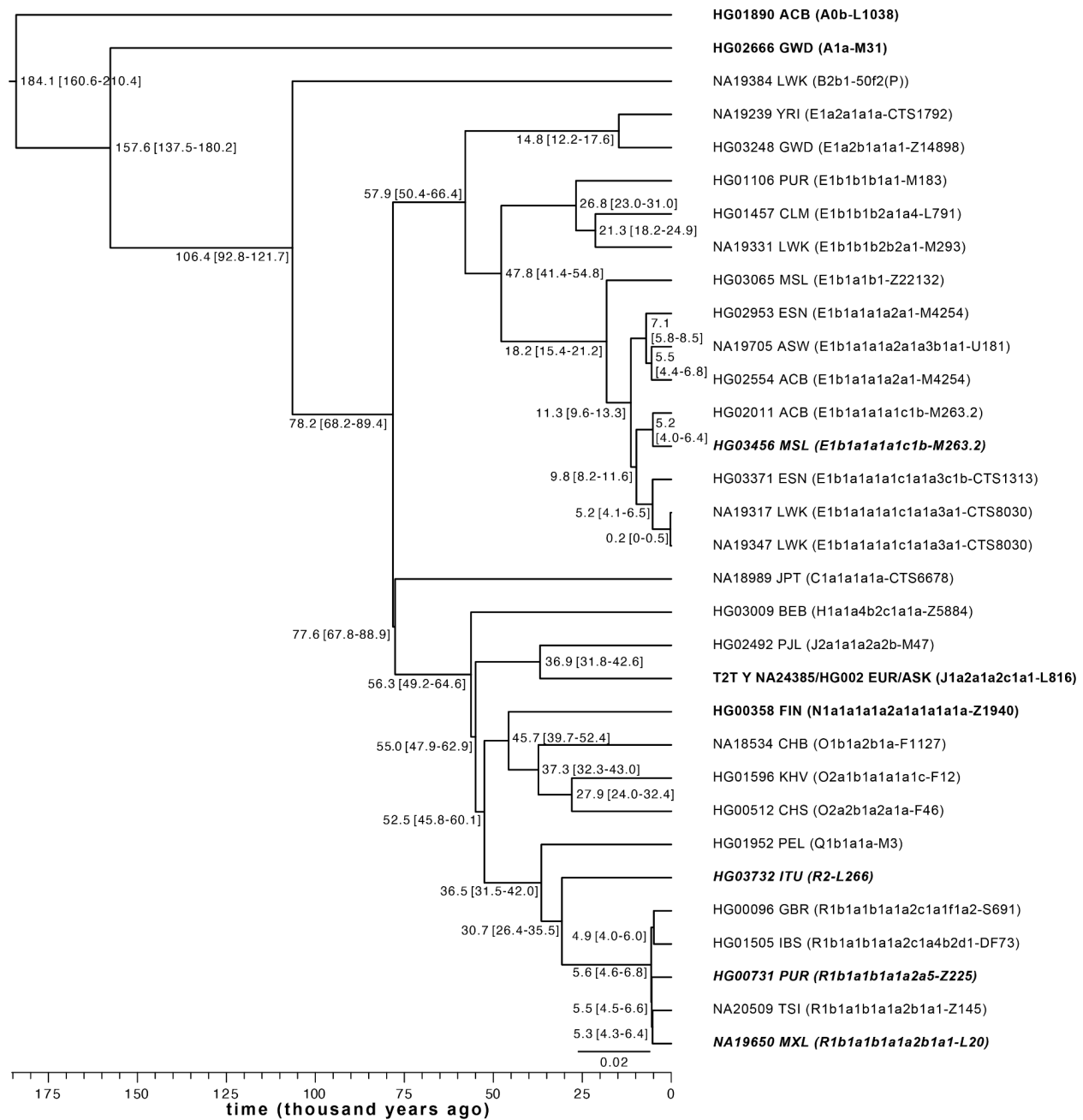
Supplemental Figure 23. Top 50 CNV genes. A comparison of the top-50 CNV genes in the HGSVC3 (violet) and HPRC (green) assemblies. (top-left) CNV genes with equal average copy number in HGSVC3 as CHM13. (top-right) CNV genes with greater average copy number in HGSVC3 versus CHM31. (bottom) Genes with lower average copy number than CHM13.

Fig. 24



Supplemental Figure 24. CNV gene contiguity. Contiguity analysis of CNV genes in HGSVC3 and HPRC assemblies. Each point reflects a gene that is duplicated in both assemblies. The x and y coordinates reflect the closest distance to an assembly break defined by the end of a contig or an unassigned base ("N"). Points above the diagonal reflect greater contiguity in HGSVC3 assemblies.

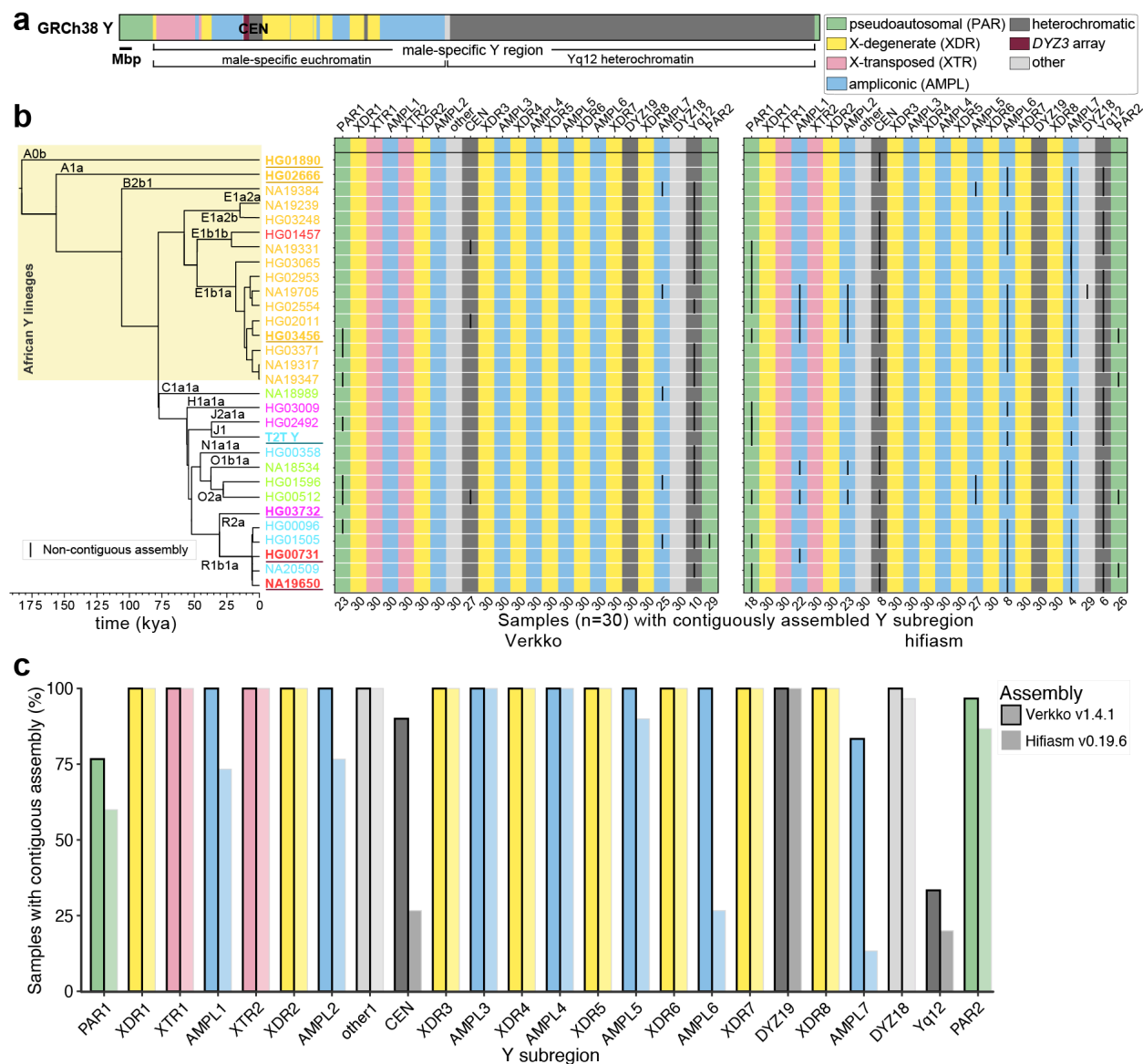
Fig. 25



Supplementary Figure 25. Phylogenetic relationships of 32 Y chromosomes. Split times as estimated according to the BEAST analysis are shown with 95% HPD interval in brackets. Sample ID is followed by population designation, full Y haplogroup label according to ISOGG v15.73 and terminal marker ID. Population abbreviations: ACB - African Caribbean in Barbados; ASW - African Ancestry in SW USA; BEB - Bengali in Bangladesh; CHB - Han Chinese in Beijing, China; CHS - Han Chinese South; CLM - Colombian in Medellín, Colombia; ESN - Esan in Nigeria; FIN - Finnish in Finland; GBR - British From England and Scotland; GWD - Gambian in Western Division – Mandinka; IBS - Iberian Populations in Spain; ITU - Indian Telugu in the U.K.; JPT - Japanese in Tokyo, Japan; KHV - Kinh in Ho Chi Minh City, Vietnam; LWK - Luhya in

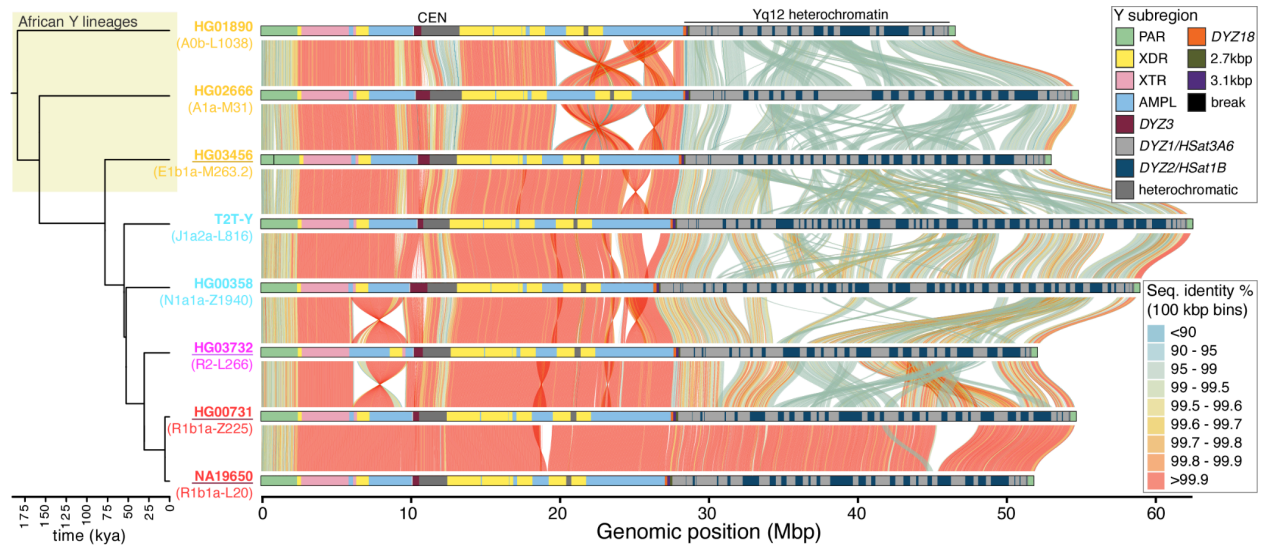
Webuye, Kenya; MSL - Mende in Sierra Leone; MXL - Mexican Ancestry in Los Angeles CA USA; PEL - Peruvian in Lima Peru; PJI - Punjabi in Lahore, Pakistan; PUR - Puerto Rican in Puerto Rico; TSI - Toscani in Italia; YRI - Yoruba in Ibadan, Nigeria. Four novel contiguously assembled Y chromosomes (note - HG03456 has an assembly break in pseudoautosomal region 1) are shown in bold italics and previously reported contiguous Y chromosomes^{23,24} in bold.

Fig. 26



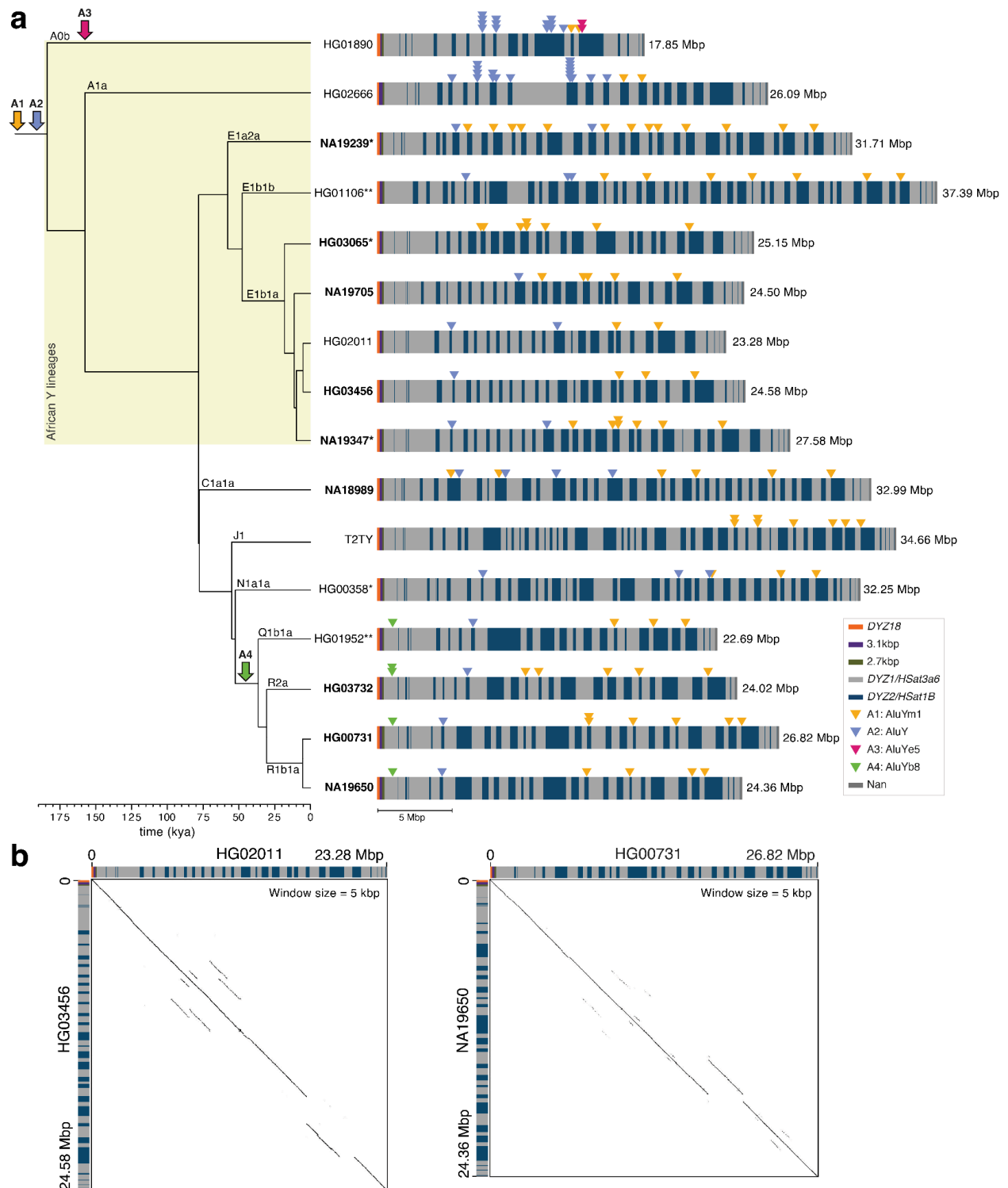
Supplementary Figure 26. Assembly contiguity of the Y chromosomes. **a.** Human Y chromosome structure based on the GRCh38 reference sequence. **b.** Dated Y phylogeny (left) and assembly completeness of Y-chromosomal subregions comparing Verkko and hifiasm (right). The color of sample ID corresponds to superpopulation designation (**Fig. 1a**). A Y-chromosomal subregion was considered to be contiguous if it was assembled contiguously from the subclass on the left to the subclass on the right. PAR1 and PAR2 regions were considered contiguously assembled if they contained the telomeric repeat sequence. kya - thousand years ago. **c.** The proportion of contiguously assembled Y subregions across 30 males comparing Verkko and hifiasm. See panel **a** for subregion labels.

Fig. 27



Supplementary Figure 27. Pairwise comparison of the eight contiguously assembled Y chromosomes. Four novel contiguous Y assemblies, representing E1b1a, R2a and R1b1a Y lineages (three assembled telomere-to-telomere and one, HG03456, with a break in PAR1 region) are underlined. Six chrY assemblies were generated here using Verkko (v1.4.1) and two (T2T Y and HG00358) were included from ^{23,24}.

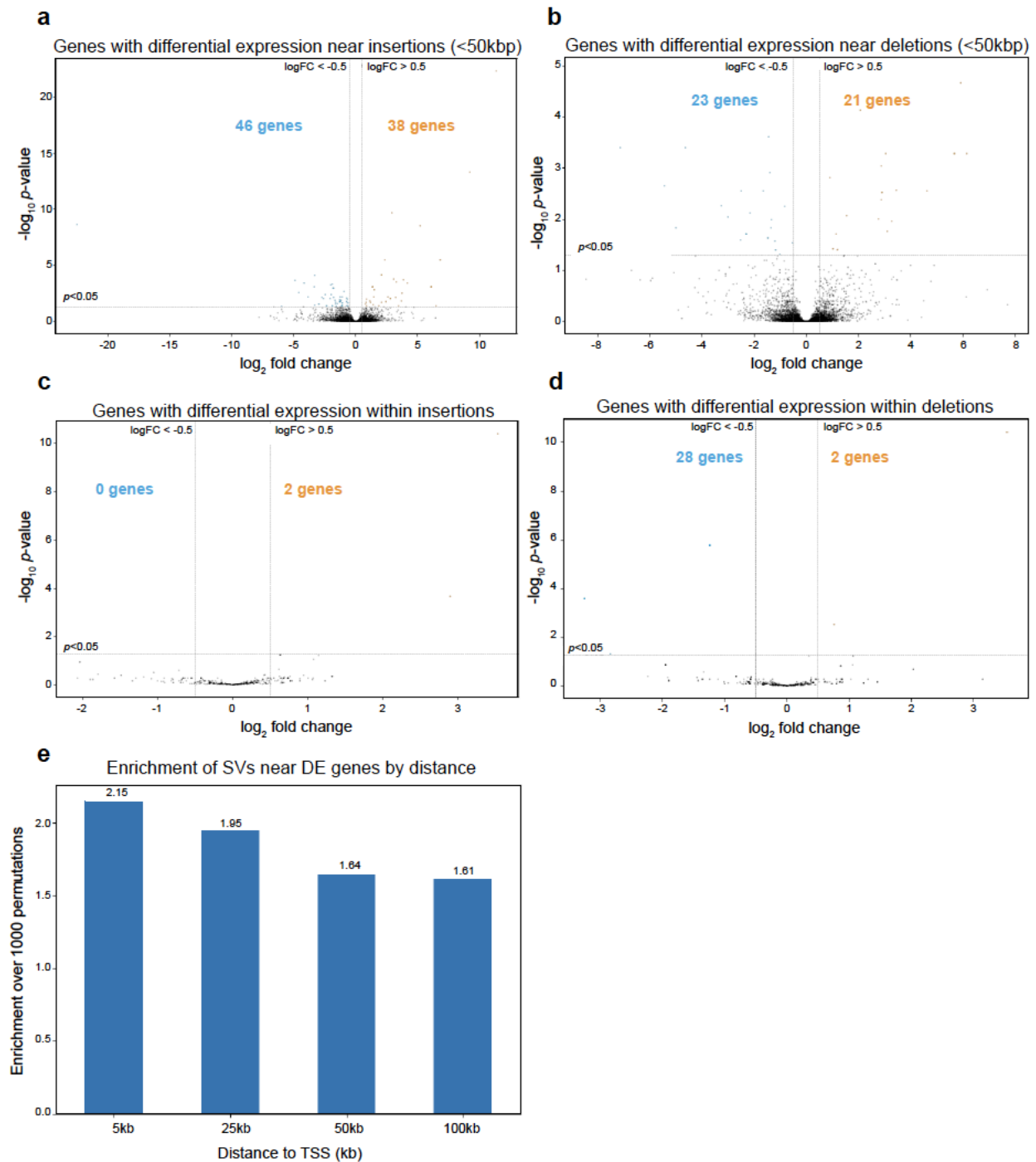
Fig. 28



Supplementary Figure 28. Yq12 heterochromatic region. a) Repeat composition of the contiguously assembled Yq12 heterochromatic regions across 16 individuals. Distribution of the four reported²³ *Alu* insertions across the Yq12 (labeled A1-A4) are shown as filled triangles.

Based on their phylogenetic distribution, the insertion of *Alu* elements A1 and A2 occurred prior to the split of studied Y chromosomes, while A3 insertion is present only in the A0b lineage and A4 shared occurred in the common ancestor of Q and R lineages. Individuals included were assembled using Verkko (this study, $n=9$), hifiasm (this study, $n=4$, indicated by an asterisk). Yq12 sequences for two individuals (indicated by double asterisks) were included from Ref. ²³ and the T2T Y chromosome sequence from Ref. ²⁴. Novel, previously unpublished Yq12 sequences ($n=9$) are shown in bold text. The total length of the Yq12 region is indicated on the right. ka, thousand years ago. **b)** Dotplots of the Yq12 regions of two closely related chrY pairs (haplogroup E1b1a: HG02011 and HG03456 and R1b1a: HG00731 and NA19650) using window size = 5,000 bp and showing evidence of two (left) and one (right) large insertion/deletion events.

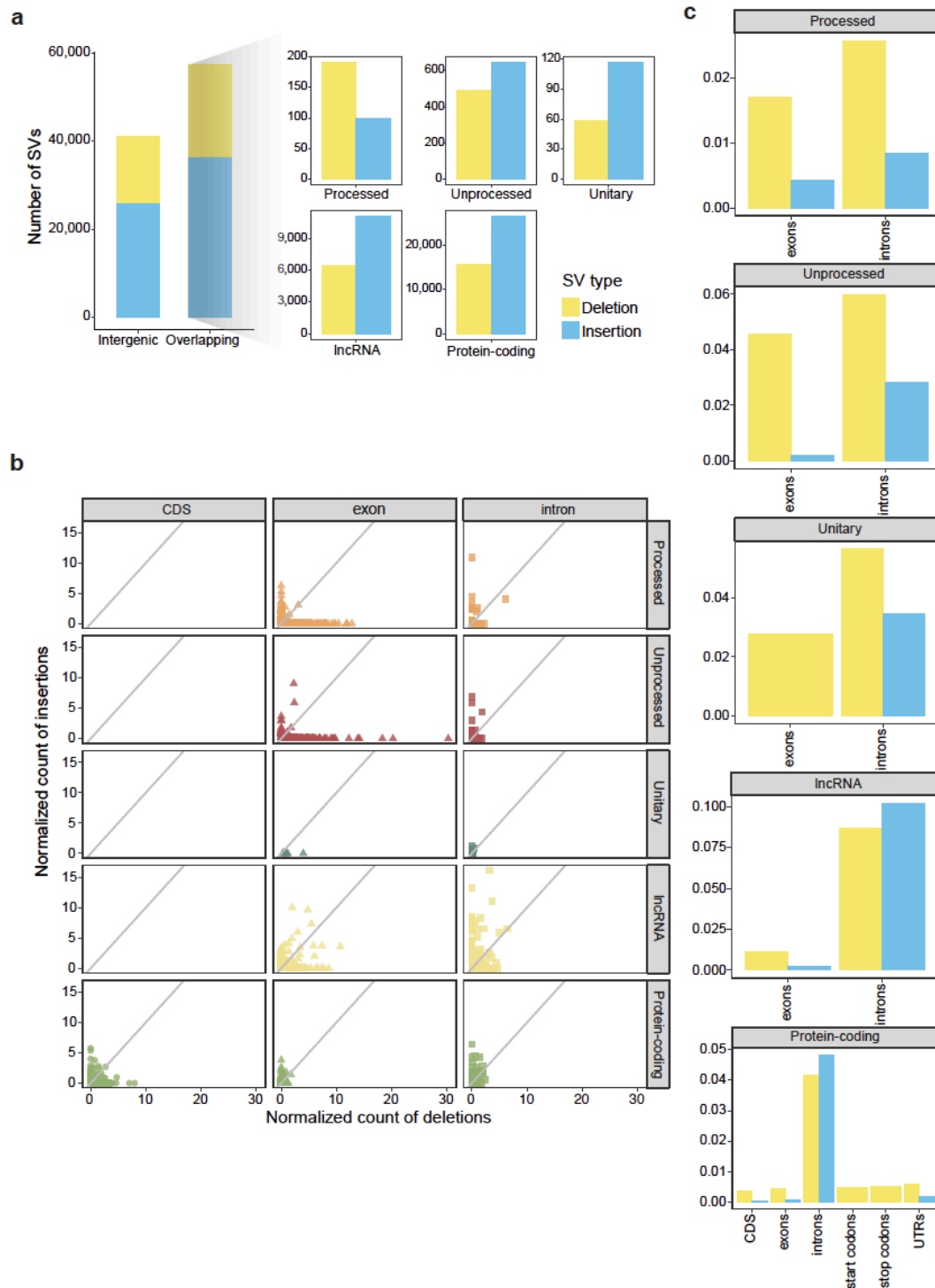
Fig. 29



Supplementary Figure 29. Differential expression analysis. a,b,c,d) Volcano plots of differentially expressed genes with (a) insertions and (b) deletions located <50 kbp from the gene TSS without overlapping the gene, or those that directly overlap (c) insertions and (d) deletions. Differential expression was calculated by comparing groups of 2-10 individuals who carried the SV vs. the remainder of individuals without the SV. The x-axis is the $\log_2$ -fold-change of read counts, and the y-axis is $-\log_{10}$ p-values (adjusted with Benjamini-Hochberg correction). The significant expression increases and decreases are defined as points with adjusted p-value

< 0.05 and $\log_2$ -fold-change > 0.5 or < -0.5 . The numbers indicate counts of SV-gene pairs considered significant. A full list of p-values and test statistics for the summary genes are provided in Supplementary **Table 43**. **e)** Enrichment of proportions of genes with differential expression occurring near SVs per genomic distance, compared with the proportion of genes near SVs averaged across 1000 permutation tests. The decreasing trend indicates that SVs closer to TSSs are more likely to affect expression.

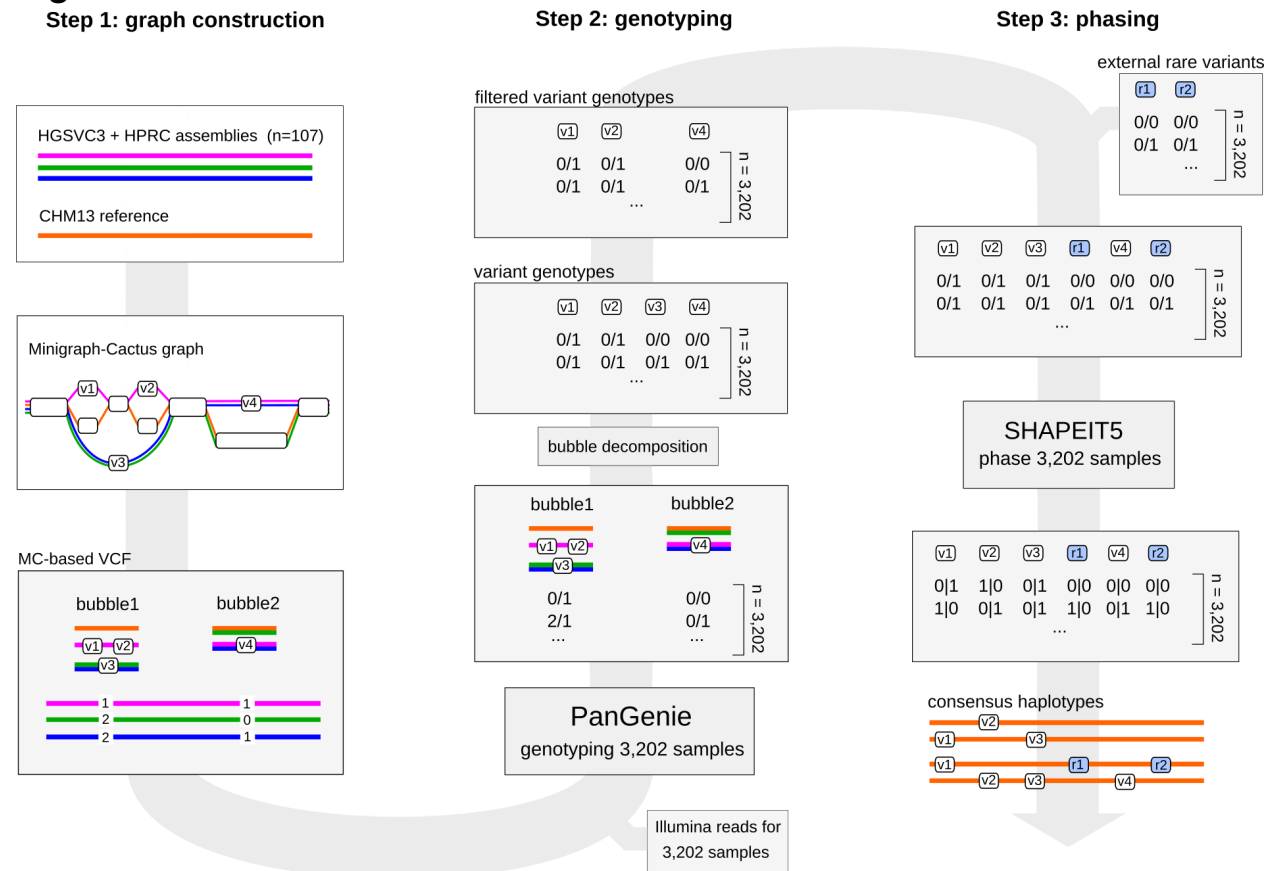
Fig. 30



Supplementary Figure 30. SV intersects with functional elements. **a)** Number of intergenic and overlapping SVs. Over 40% of SVs in the human genome were intergenic, indicating that these SVs did not overlap with annotated protein-coding genes, long noncoding RNAs (lncRNAs), or pseudogenes (including processed, unprocessed, and unitary types). For the SVs that did overlap with genes, we categorized the genes into different biotypes based on

GENCODE v45 annotations. Insertions outnumbered deletions across all gene biotypes, except for processed pseudogenes. **b)** Normalized counts of deletions and insertions. The X-axis represents the number of deletions, while the Y-axis represents the number of insertions. Number of SV overlaps were counted per gene and per genomic element, and we normalized the counts by dividing them by the length of the genomic elements for a given gene and multiplying by a scaling factor of 1,000. Overall, pseudogenes tended to have more deletions within both exons and introns, while protein-coding genes exhibited a balance between insertions and deletions within coding sequences (CDSs), exons, and introns. **c)** Proportion of different genomic elements affected by SVs. At the transcript level, we calculated the overlapping length of SVs per genomic or transcript element (i.e., intron or exon). The proportion was defined by dividing the overlapping length by the total length in the genome. As expected, a higher proportion was consistently observed for both insertions and deletions within introns compared to exons across different transcript biotypes. Notably, protein-coding transcripts exhibited a significantly reduced proportion for insertions and deletions within CDSs, exons, start codons, stop codons, and untranslated regions (UTRs) compared to introns.

Fig. 31



Supplementary Figure 31. Overview of the graph constructing, genotyping and phasing pipelines. **Step 1:** graph construction. A pangenome graph is constructed from haplotype-resolved assemblies of 107 samples (65 HGSVC3, 42 HPRC). A VCF representation of the bubbles with phased genotypes for all assembly samples is generated. **Step 2:** genotyping. The bubbles are genotyped across all 3,202 1kGP samples with PanGenie using the panel VCF generated in step 1 as input, as well as Illumina reads for all samples. Bubbles are decomposed into variant alleles and bubble genotypes are translated to variant genotypes. Finally, genotypes are filtered using a machine-learning approach. **Step 3:** phasing. The filtered genotypes are combined with additional rare variant calls produced from short reads across all 3,202 samples and the combined set is phased with SHAPEIT5. Based on the resulting phased genotypes, consensus haplotypes are generated for all 3,202 samples by implanting variants located on each haplotype into the reference genome.

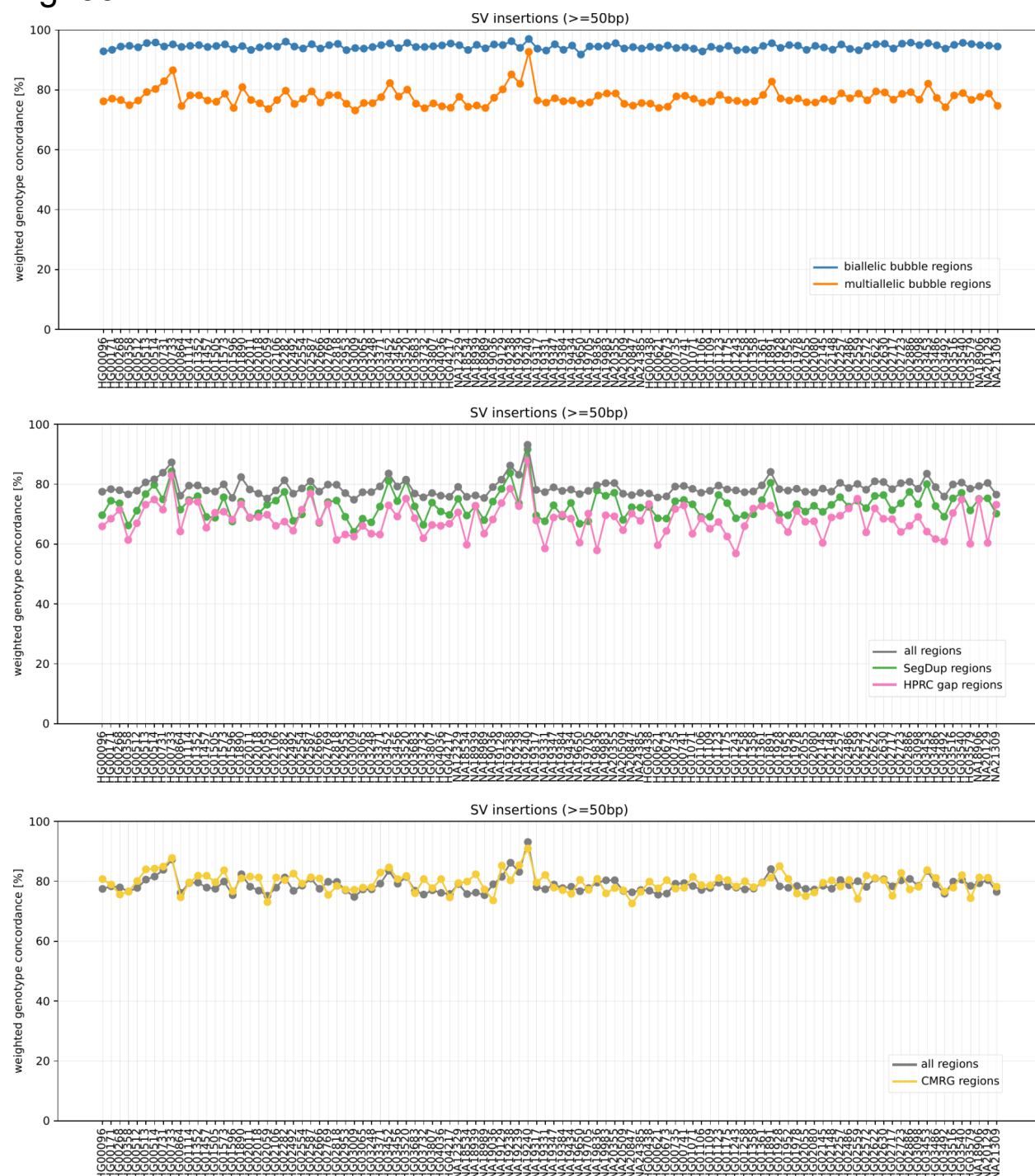
Fig. 32



Supplementary Figure 32. PanGenie Leave-one-out validation for deletions. We conducted a leave-one-out experiment by repeatedly removing one sample from the MC-based panel VCF and genotyping it with PanGenie. The left out sample was used as a ground truth for evaluation based on the weighted genotype concordance³. The plots show concordances of variants falling within certain regions: all biallelic bubble regions of the graph (blue), multiallelic bubble regions (orange), segmental duplications (green), regions of gaps in the HPRC assemblies¹ (pink) and

within challenging medically relevant genes³¹ (yellow).

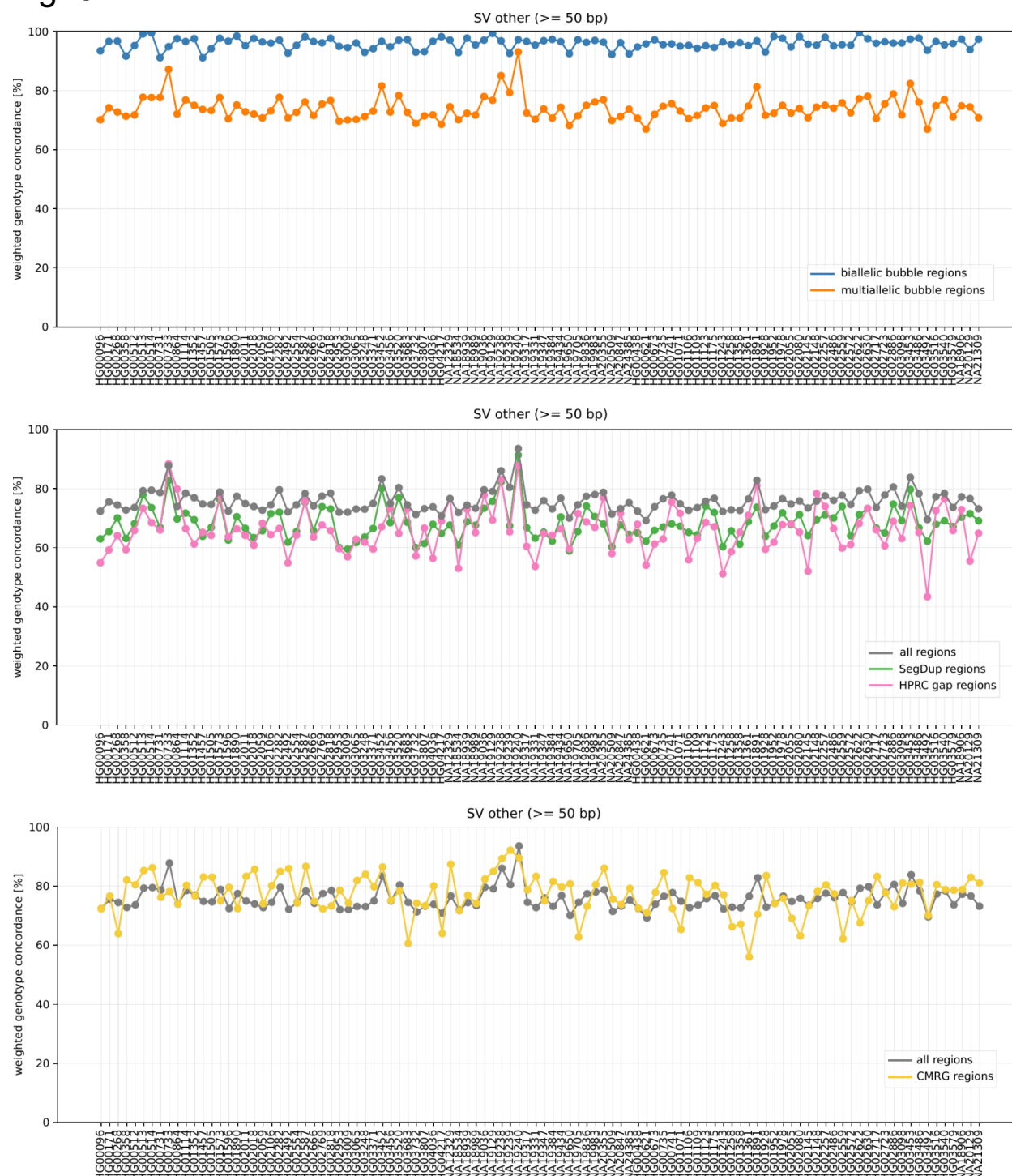
Fig. 33



Supplementary Figure 33. PanGenie leave-one-out validation for insertions. We conducted a leave-one-out experiment by repeatedly removing one sample from the MC-based panel VCF and genotyping it with PanGenie. The left out sample was used as a ground truth for evaluation based on the weighted genotype concordance³. The plots show concordances of variants falling within certain regions: all biallelic bubble regions of the graph (blue), multiallelic bubble regions

(orange), segmental duplications (green), regions of gaps in the HPRC assemblies¹ (pink) and within challenging medically relevant genes³¹ (yellow).

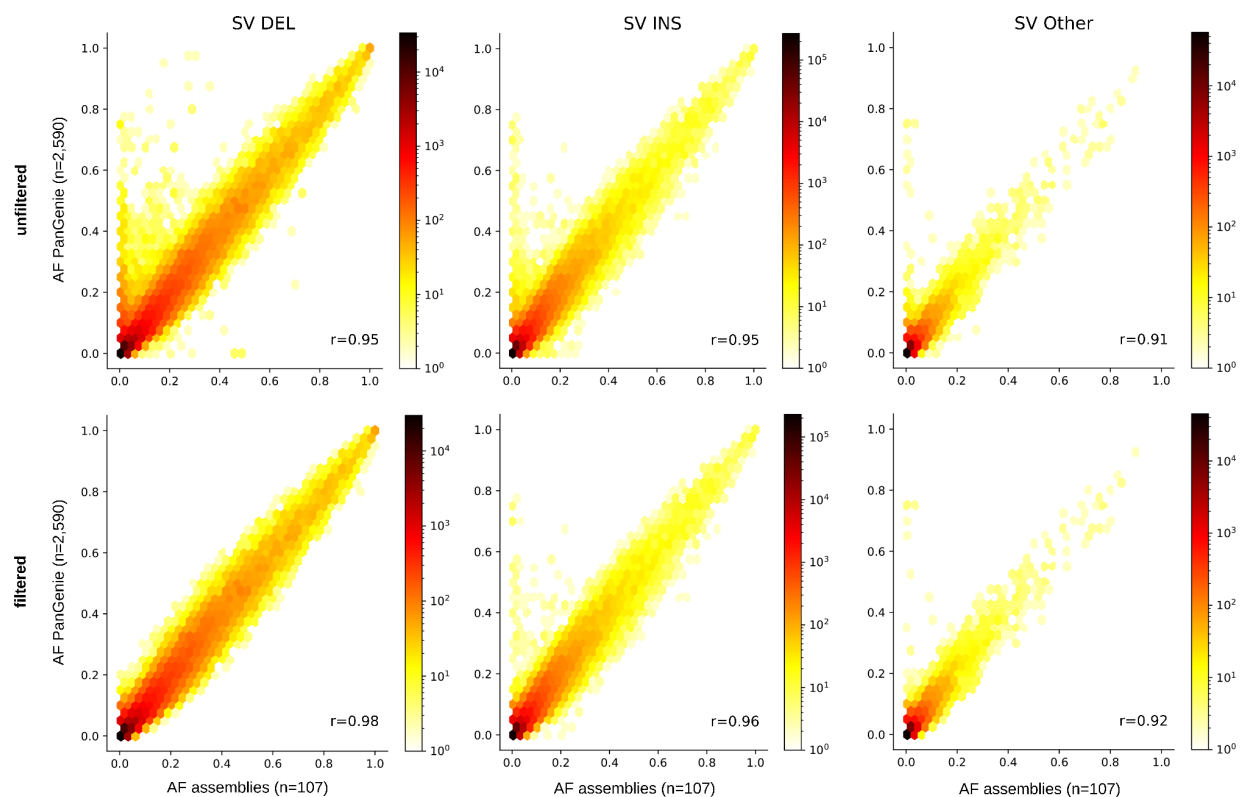
Fig. 34



Supplementary Figure 34. PanGenie Leave-one-out validation for other variants. We conducted a leave-one-out experiment by repeatedly removing one sample from the MC-based panel VCF and genotyping it with PanGenie. The left out sample was used as a ground truth for evaluation based on the weighted genotype concordance³. The plots show concordances of variants falling within certain regions: all biallelic bubble regions of the graph (blue), multiallelic

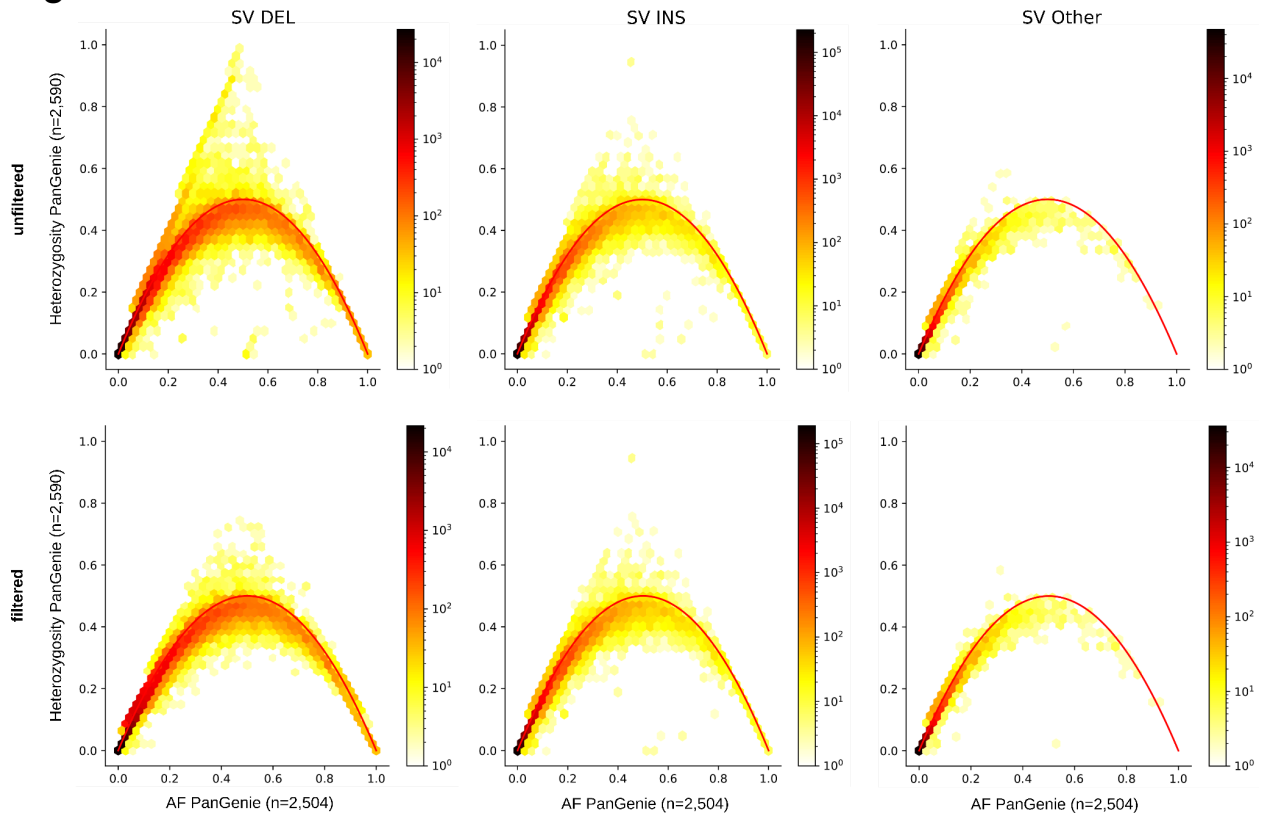
bubble regions (orange), segmental duplications (green), regions of gaps in the HPRC assemblies¹ (pink) and within challenging medically relevant genes³¹ (yellow).

Fig. 35



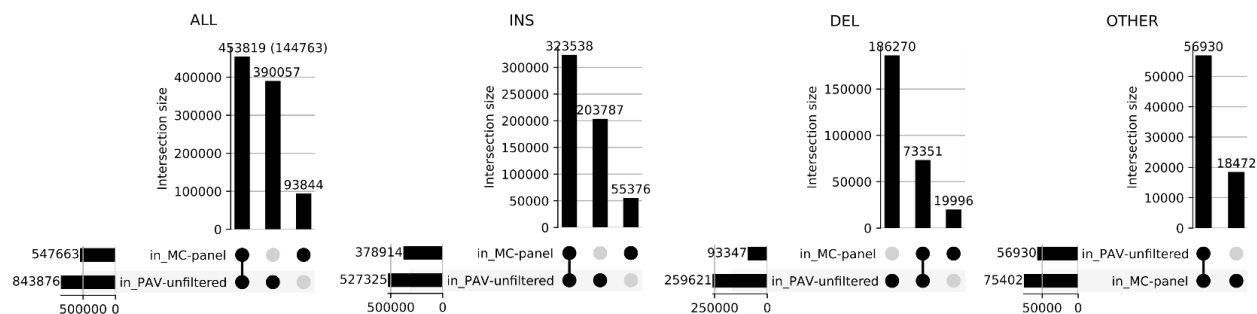
Supplementary Figure 35. Comparison of allele frequencies by discovery and genotyping. Shown is a comparison of allele frequencies for SV alleles across the assembly samples in the panel VCF ($n=107$) and corresponding allele frequencies after genotyping these variants across $n=2,590$ unrelated 1kGP samples.

Fig. 36

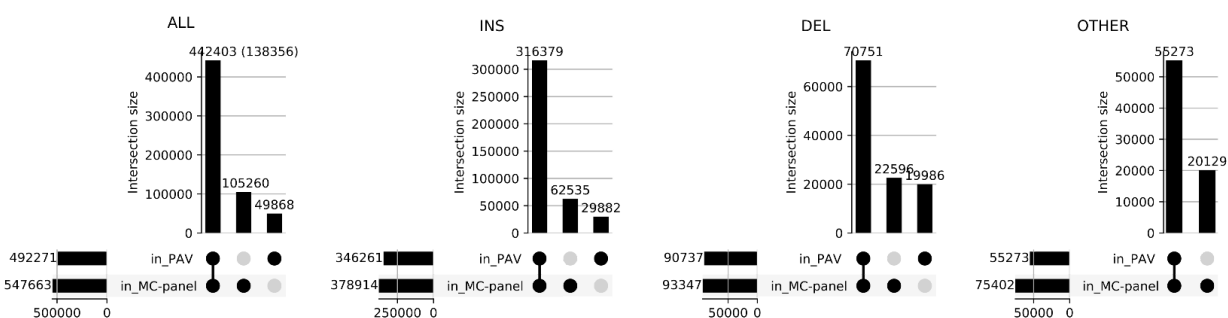


Supplementary Figure 36. Hardy-Weinberg Equilibrium for PanGenie genotypes. Shown is a comparison of heterozygosities and allele frequencies after genotyping $n=2,590$ unrelated 1kGP samples.

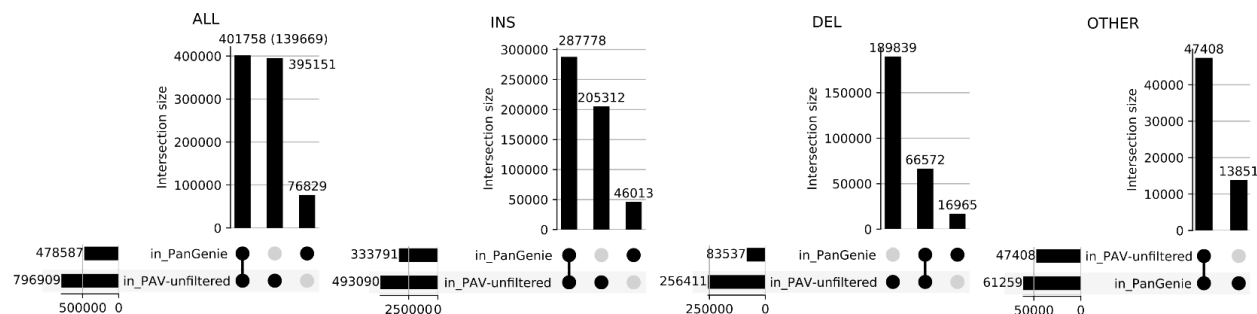
Fig. 37
MC panel vs. PAV unfiltered



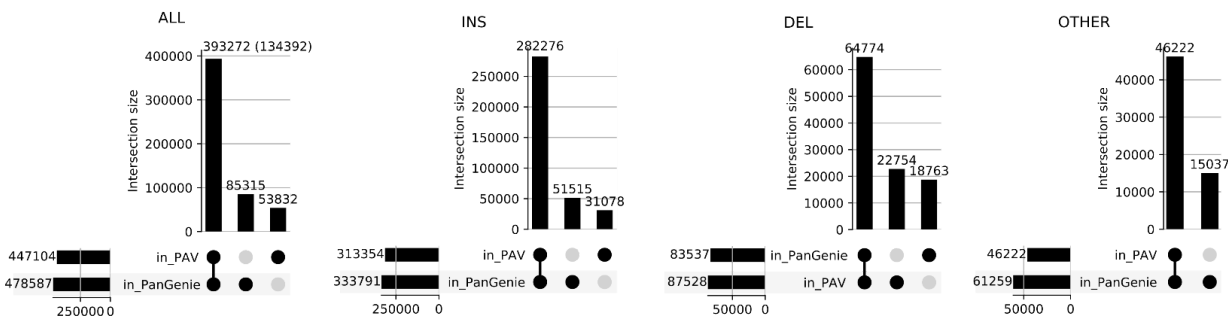
MC panel vs. PAV



PanGenie vs. PAV unfiltered



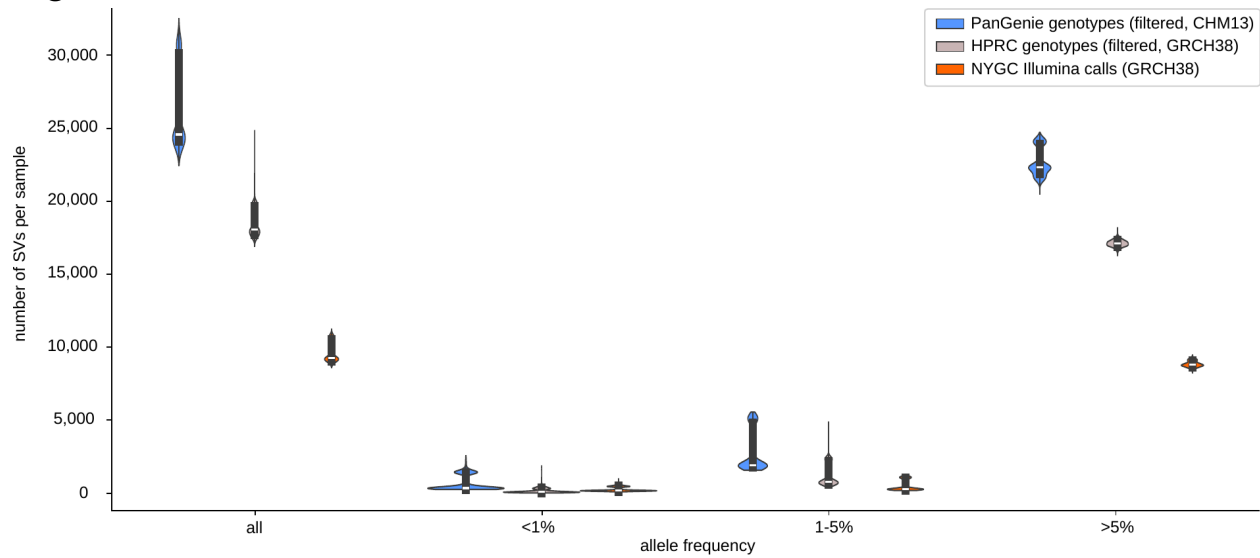
PanGenie vs. PAV



Supplementary Figure 37. PAV vs graph genotypes. Comparison of SVs in Minigraph-Cactus panel VCF (“MC-panel”), PanGenie filtered genotypes (“PanGenie”), PAV unfiltered (“PAV-unfiltered”) and PAV (“PAV”) callsets. Variant types and counts in intersections correspond

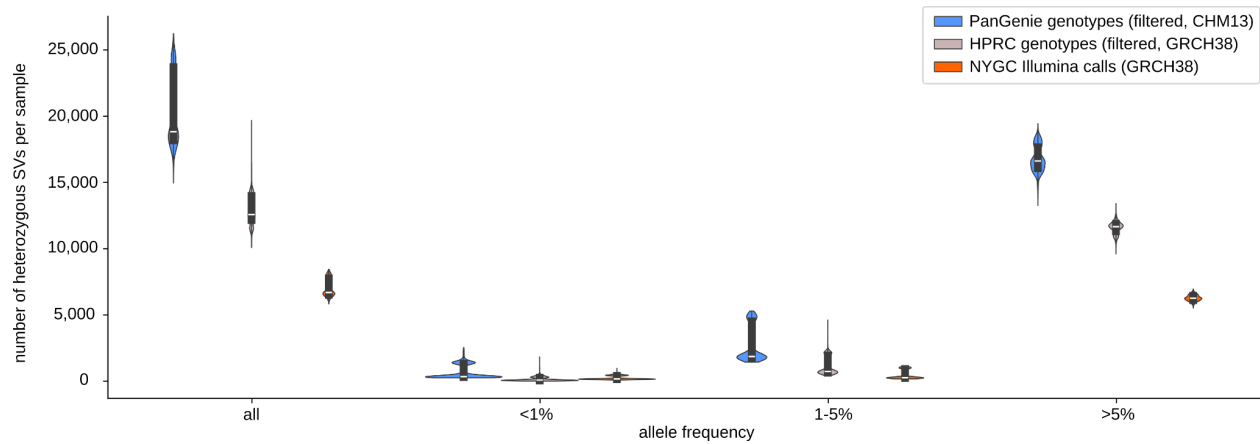
to the MC / PanGenie sets; the numbers in brackets give the number of unique PAV calls matched (only for “ALL”).

Fig. 38



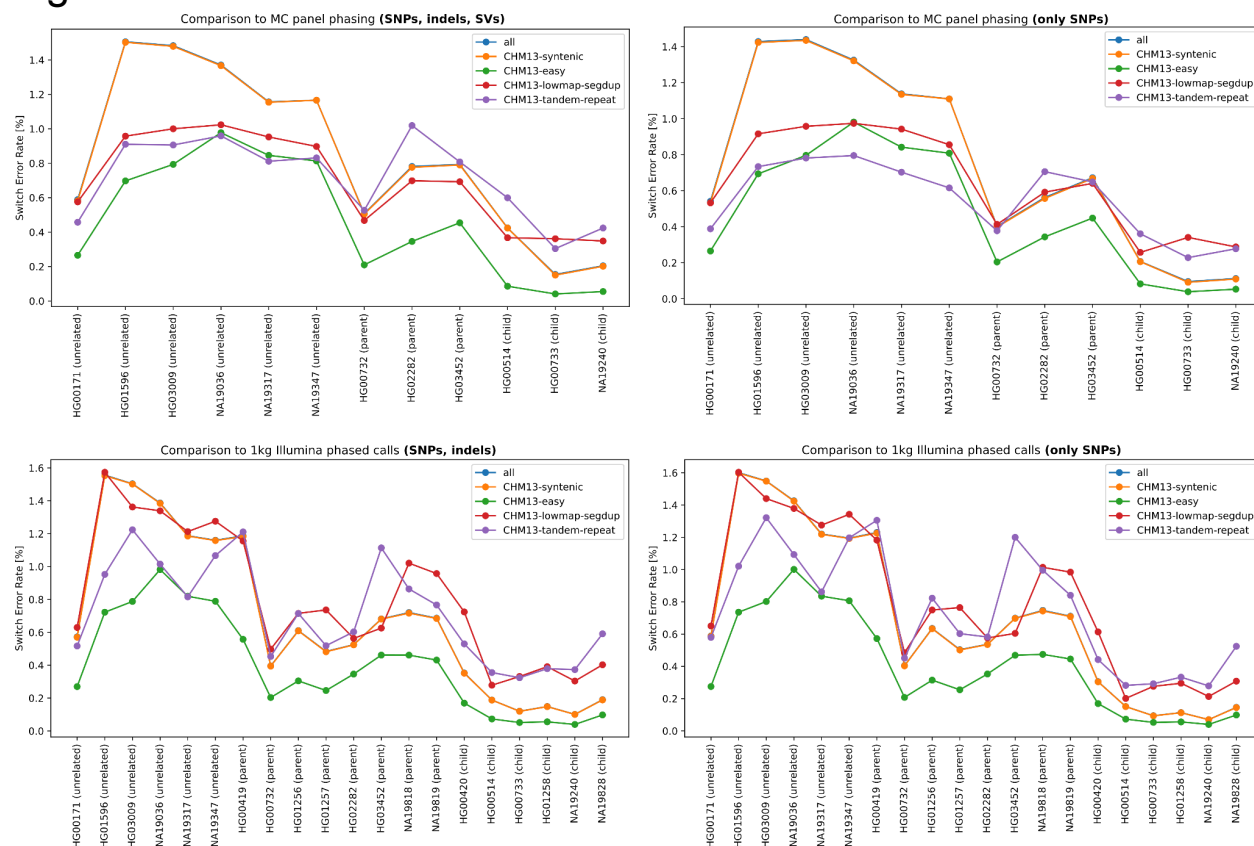
Supplementary Figure 38. Comparison of PanGenie genotype frequencies. We show the number of SV alleles present (0/1 or 1/1) in each sample ($n=3,202$) of each callset. Compared were our filtered genotyped set (blue), the HPRC genotyped set (grey) and the Illumina-based NYGC SV callset (orange). The boxes inside the violins represent the first and third quartiles of the data, white dots represent the medians, and black lines mark minima and maxima of the data.

Fig. 39



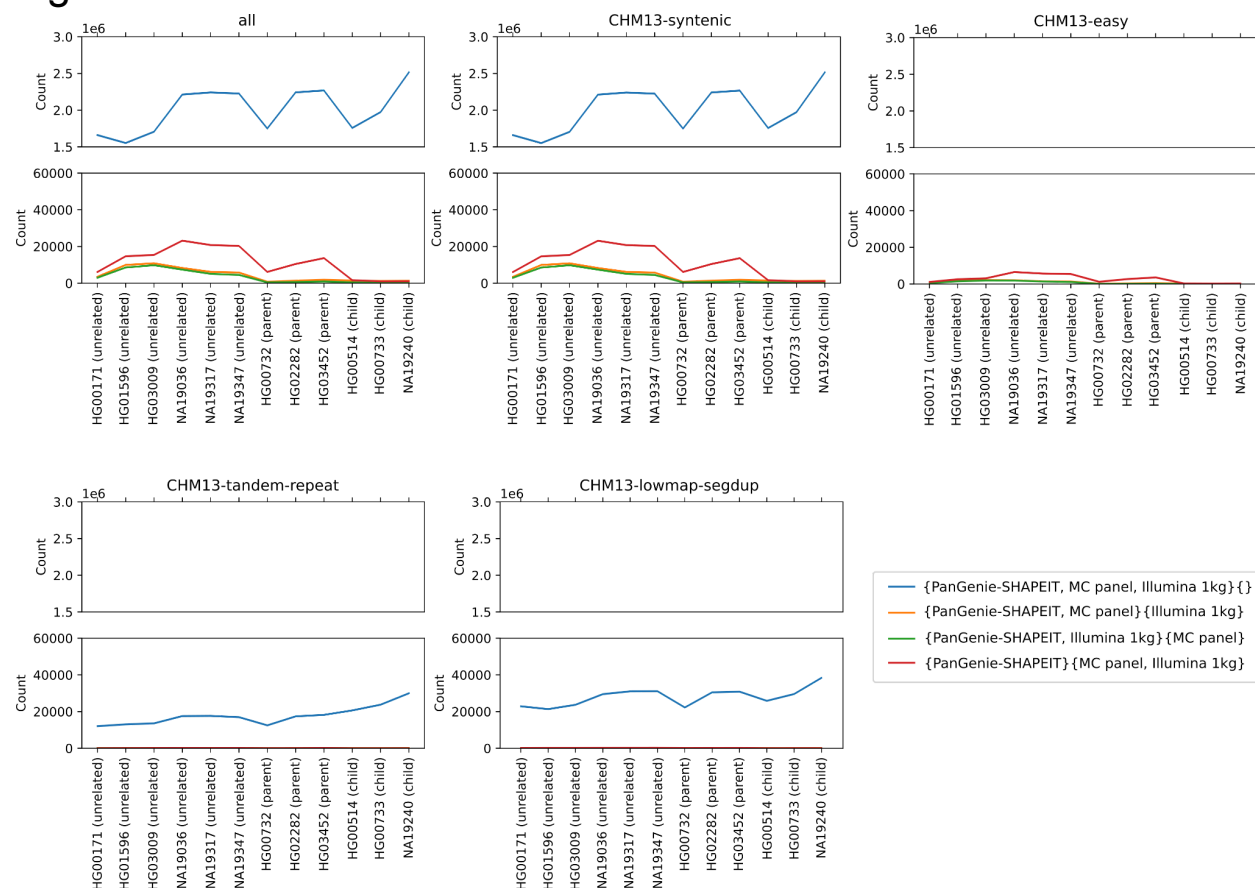
Supplementary Figure 39. Comparison of PanGenie frequencies for heterozygous genotypes. We show the number of heterozygous SV alleles present (genotype 0/1) in each sample ($n=3,202$) of each callset. Compared were our filtered genotyped set (blue), the HPRC genotyped set (grey) and the Illumina-based NYGC SV callset (orange). The boxes inside the violins represent the first and third quartiles of the data, white dots represent the medians, and black lines mark minima and maxima of the data.

Fig. 40



Supplementary Figure 40. Genotype switch error rate. Comparison of our phased genotypes to the phasing of the assembly samples in the MC VCF used as input for genotyping (top), as well as to a CHM13-based phased set of Illumina-based SNP and indel calls (bottom). The plots show switch error rates computed for all regions, as well as inside of CHM13-syntenic regions (regions shared between GRCh38 and CHM13), inside of low mappability and segmental duplication regions (“CHM13-lowmap-segdup”), tandem repeats (“CHM13-tandem-repeat”) and “easy” regions (“CHM13-easy”), i.e., regions outside of other difficult regions such as tandem repeats, homopolymers, difficult to map regions, segmental duplications and high/low GC content.

Fig. 41



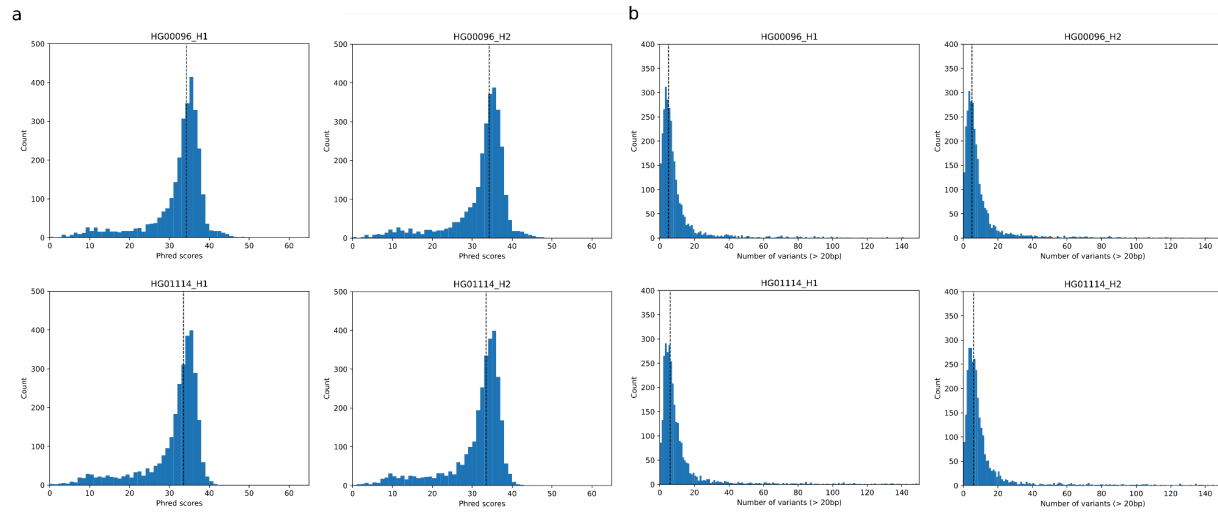
Supplementary Figure 41. Three-way comparison of phased callsets. Multiway comparison of three phased sets (all CHM13-based): our phased genotypes (“PanGenie-SHAPEIT”), the 1kGP Illumina-based phased calls and the MC panel VCF (input used for genotyping). The plots report the number of times different sets agreed / disagreed on the phasing of variant sites, i.e., the yellow line shows the number of cases in which our phasing and the MC panel phasing agreed, but the 1kGP phasing was different.

Fig. 42



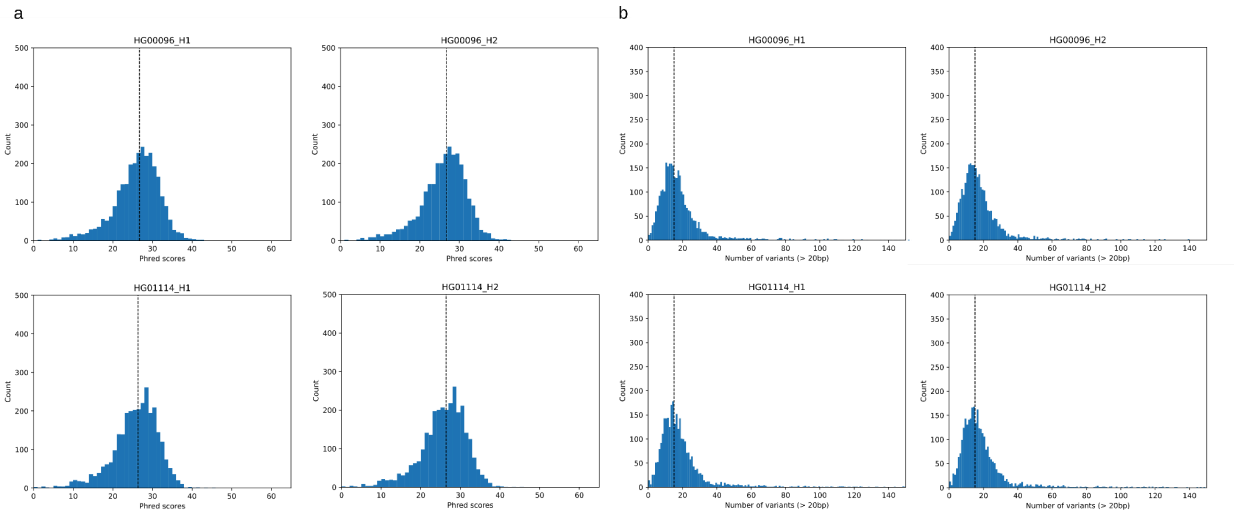
Supplementary Figure 42. Phasing QV statistics. a) QV statistics computed for the assemblies (pink), consensus haplotypes produced from the CHM13-based PanGenie-SHAPEIT phased genotypes for all regions (orange), for PanGenie-SHAPEIT within syntenic regions (regions shared between GRCh38 and CHM13) shown in green and consensus haplotypes produced from the GRCh38-based NYGC phased Illumina-calls (blue). In addition, we computed QVs for the two phased sets by switching sample labels (“randomized”), i.e., by using reads and phasings for different samples (in purple, brown, red). **b)** Completeness statistics computed for the consensus haplotypes produced from the CHM13-based PanGenie-SHAPEIT phased genotypes for all regions (yellow), for PanGenie-SHAPEIT within syntenic regions (regions shared between GRCh38 and CHM13) shown in green and consensus haplotypes produced from the GRCh38-based NYGC phased Illumina-calls (blue). In addition, we computed completeness for the two phased sets by switching sample labels (“randomized”), i.e., by using reads and phasings for different samples (in purple, brown, red).

Fig. 43



Supplementary Figure 43. Performance of genome inference with PanGenie across samples. PAV was run using one of the consensus haplotypes produced from PanGenie-SHAPEIT calls as a reference (H1 or H2) and the respective other haplotype as a query, together with the two assemblies of the same individual. Variants with conflicting genotypes between consensus and assembly haplotype were determined. **a)** Distribution of QVs estimated in windows of 1 Mbp along the reference haplotype computed based on the base pair changes caused by erroneous genotypes for all samples using the first (H1) or second (H2) consensus haplotype as a reference. **b)** Distribution of the number of erroneous variants (>20 bp) in windows of 1 Mbp along the reference haplotype.

Fig. 44



Supplementary Figure 44. Performance of 1kGP-HC call set across samples across samples. PAV was run using one of the consensus haplotypes produced from 1kGP calls as a reference (H1 or H2) and the respective other haplotype as a query, together with the two assemblies of the same sample. Variants with conflicting genotypes between consensus and assembly haplotype were determined. **a)** Distribution of QVs estimated in windows of 1 Mbp along the reference haplotype computed based on the base pair changes caused by erroneous genotypes for all samples using the first (H1) or second (H2) consensus haplotype as a reference. **b)** Distribution of the number of erroneous variants (>20 bp) in windows of 1 Mbp along the reference haplotype.

Figure 2 displays histograms of variant-based QV estimates and the number of variants for four different comparisons. The left column shows histograms of variant-based QV estimates, and the right column shows histograms of the number of variants (> 20bp).

Row a: GRCh38 vs CHM13

- Variant-based QV estimate: The histogram shows a distribution peaking around 24.55 (indicated by a vertical dashed line).
- Number of variants (> 20bp): The histogram shows a distribution peaking around 20 (indicated by a vertical dashed line).

Row b: 1kGP-HC vs assembly (HG00096)

- Variant-based QV estimate: The histogram shows a distribution peaking around 26.73 (indicated by a vertical dashed line).
- Number of variants (> 20bp): The histogram shows a distribution peaking around 15 (indicated by a vertical dashed line).

Row c: Pangenie vs assembly (HG00096)

- Variant-based QV estimate: The histogram shows a distribution peaking around 34.23 (indicated by a vertical dashed line).
- Number of variants (> 20bp): The histogram shows a distribution peaking around 5 (indicated by a vertical dashed line).

Row d: Assembly vs Q100 (HG002)

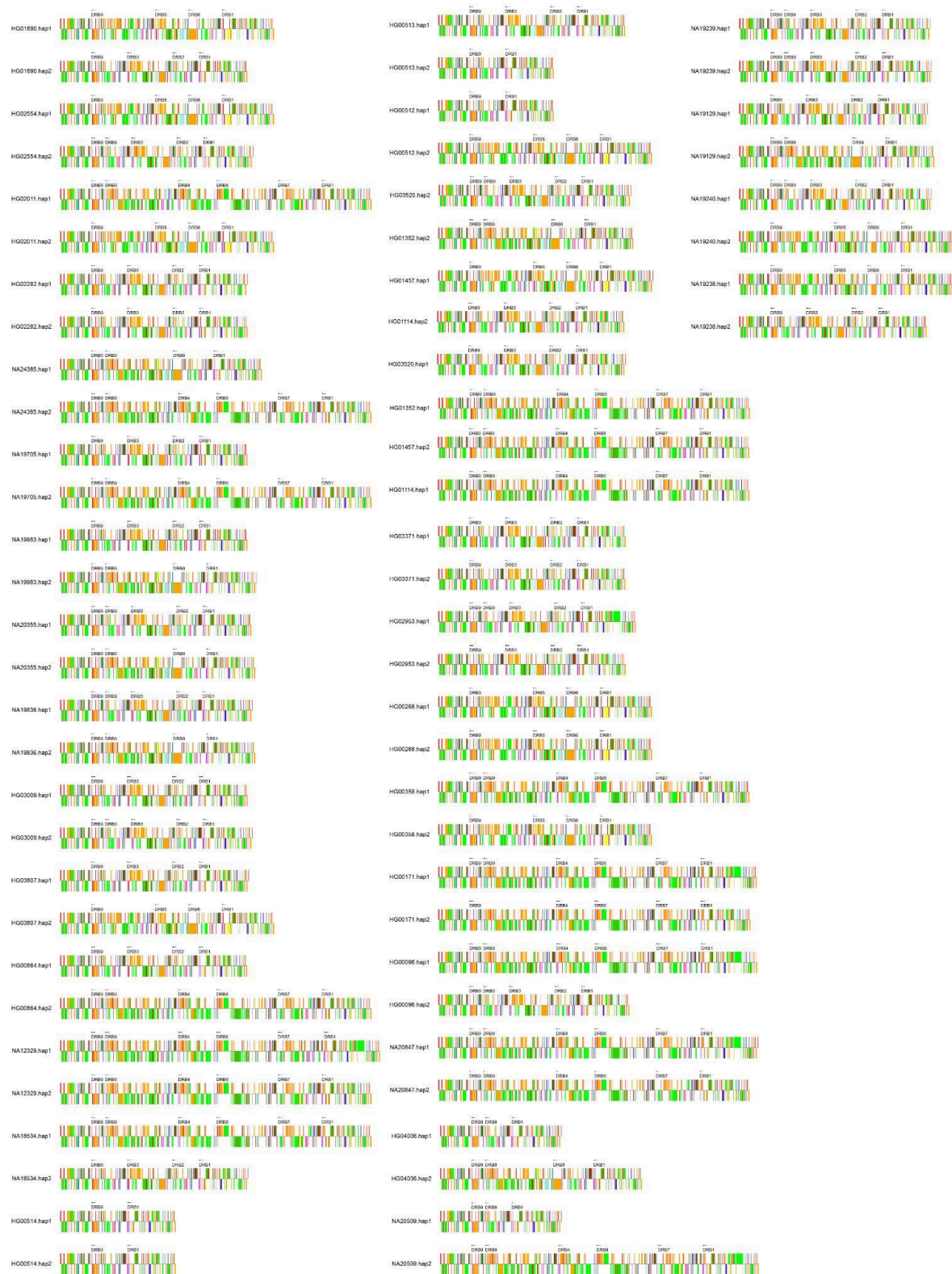
- Variant-based QV estimate: The histogram shows a distribution peaking around 50.22 (indicated by a vertical dashed line).
- Number of variants (> 20bp): The histogram shows a distribution peaking around 0 (indicated by a vertical dashed line).

118

reference, and the second haplotype and both assemblies for HG00096 were aligned for variant calling. **c)** Consensus haplotype HG00096 H1 based on PanGenie-SHAPEIT (CHM13-based) was used as reference, and the second haplotype and both assemblies for HG00096 were aligned for variant calling. **d)** Q100 assembly for HG002 (paternal) was used as reference, and both assembly haplotypes for HG002 and the maternal Q100 haplotype were aligned for variant calling. The median is indicated by a dashed line and shown in bold.

Fig. 46

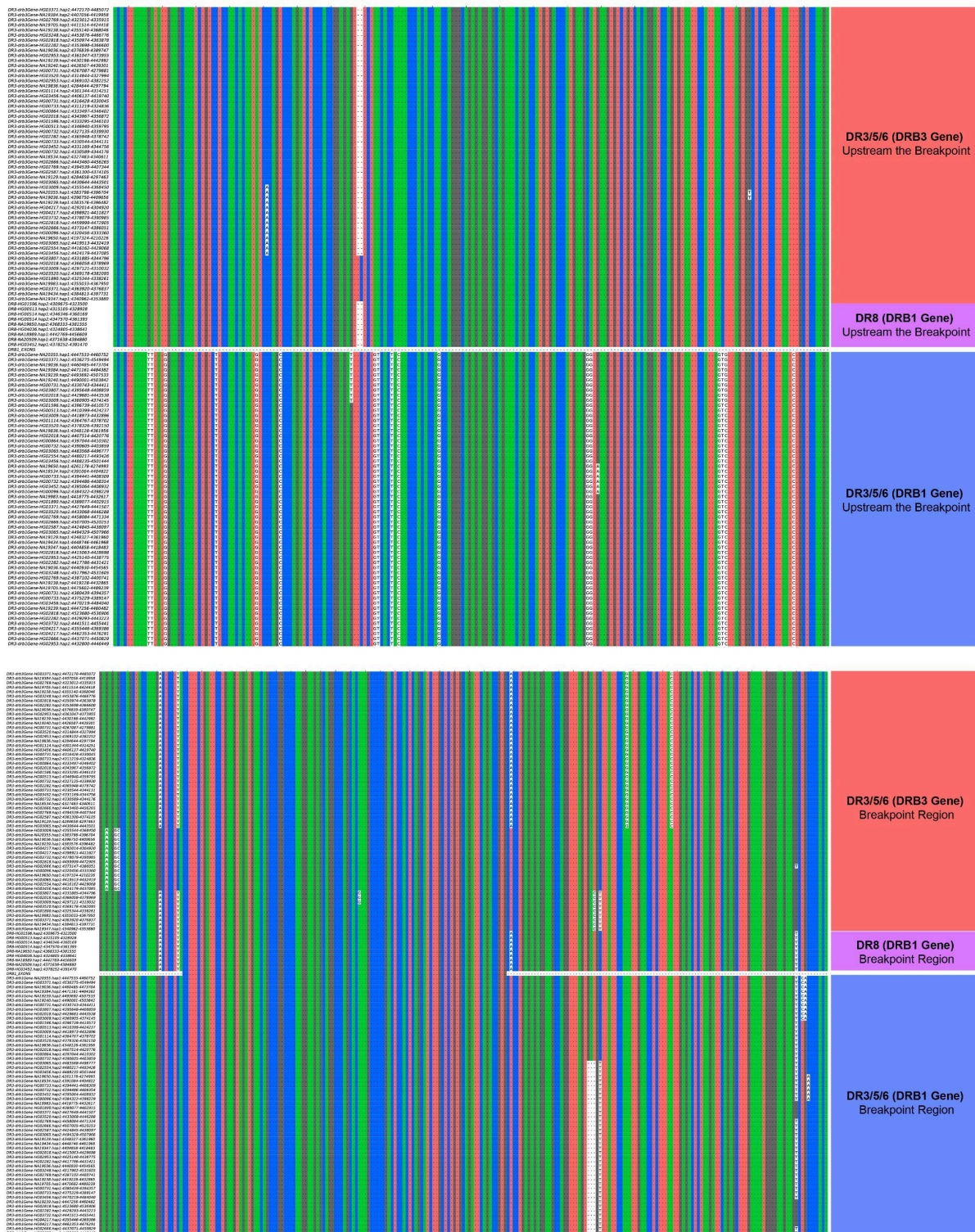


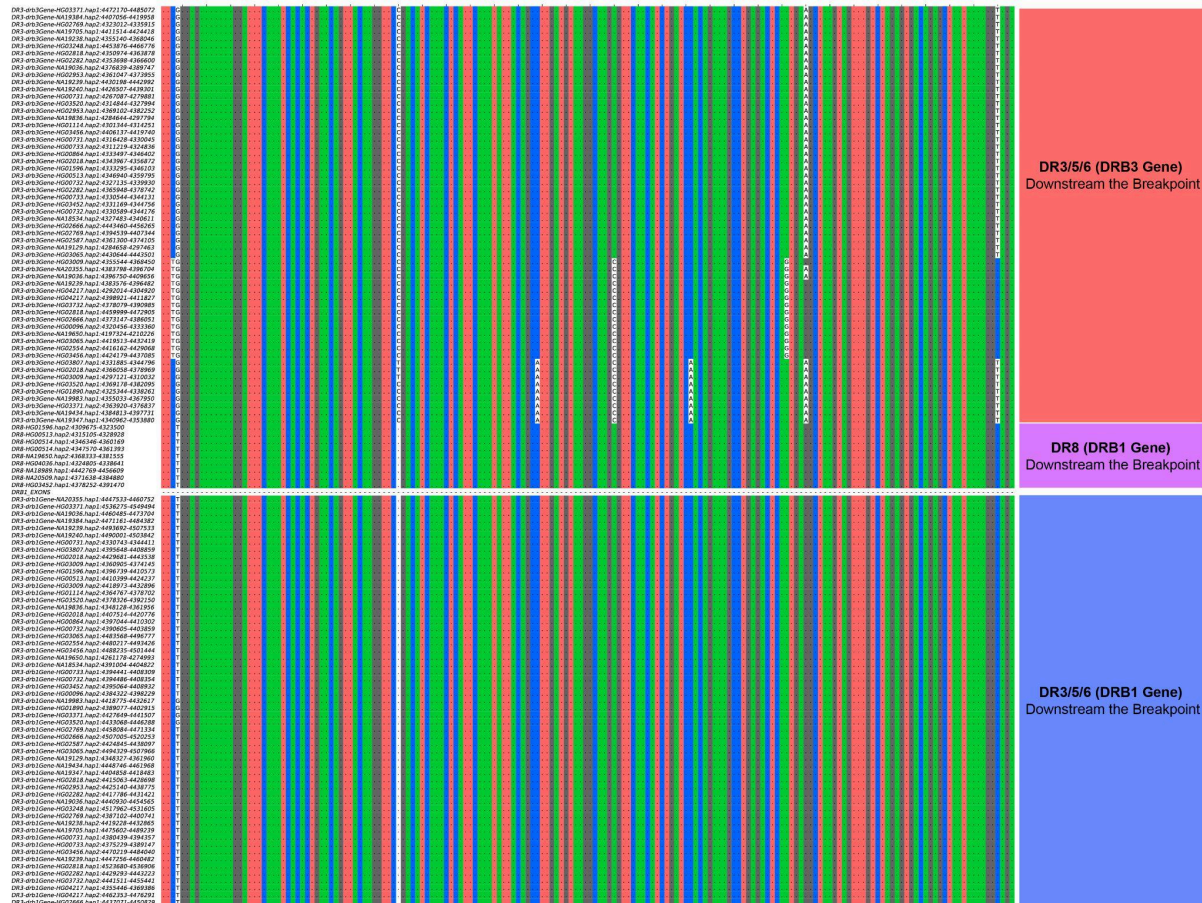


Supplementary Figure 46. DRB exon and repeat visualization. Plots showing the repetitive element and gene/pseudogene exon distribution within all 128 haplotypes analyzed (2/130 haplotypes contained a stretch of Ns in the assembly and were not analyzed). Repetitive

elements, gene exons, and pseudogene exons are uniquely colored. Lines that are oriented above the midline represent sense orientation, while those below the midline are in the antisense orientation. The first exons of each gene/pseudogene are denoted by an arrow with the name of the gene/pseudogene. The direction of the arrow also denotes the orientation of transcription of the gene compared to the region. Some haplotypes contain additional low divergent copies of the first DRB exon (denoted as a second *DRB9* first exon).

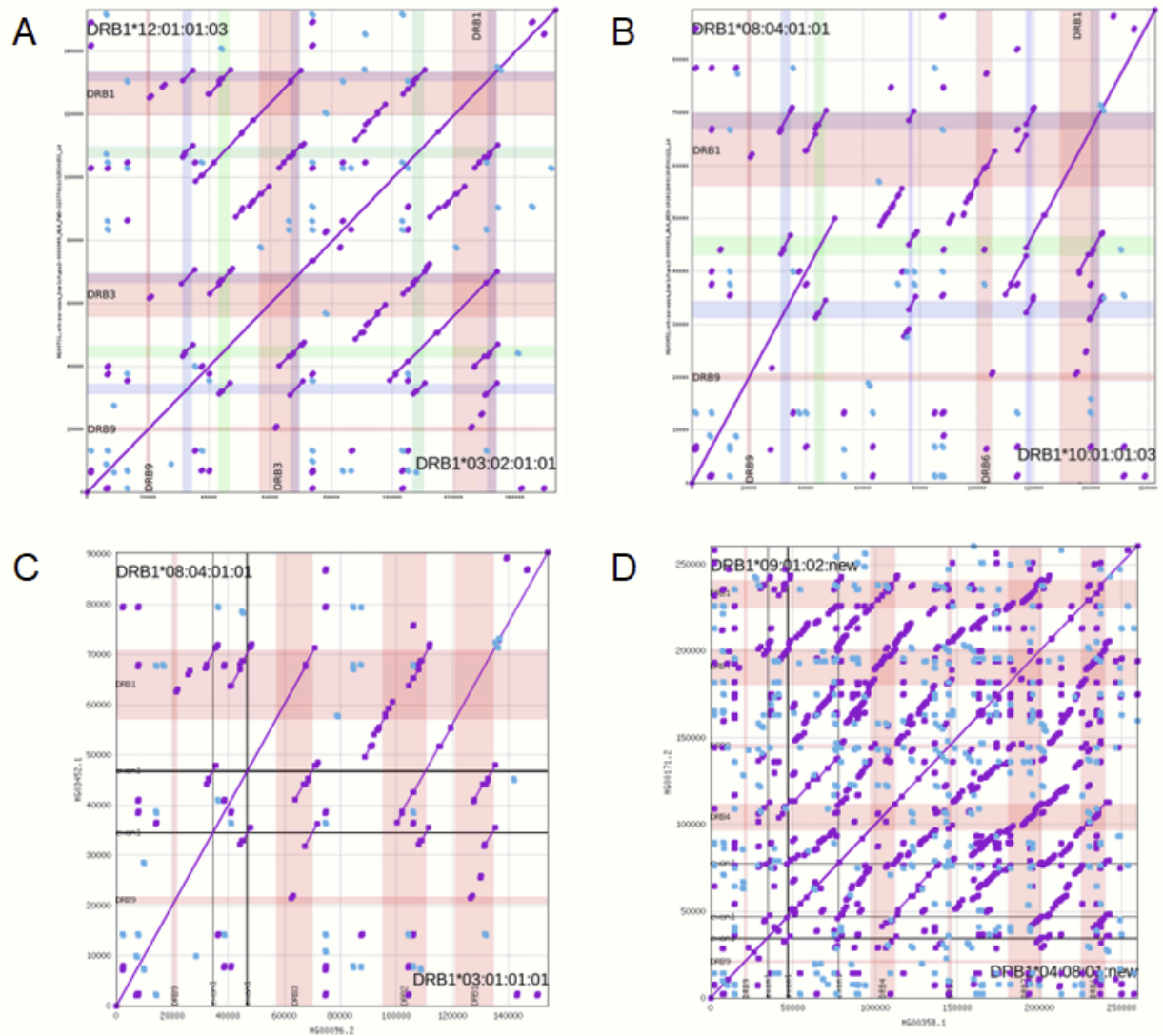
Fig. 47





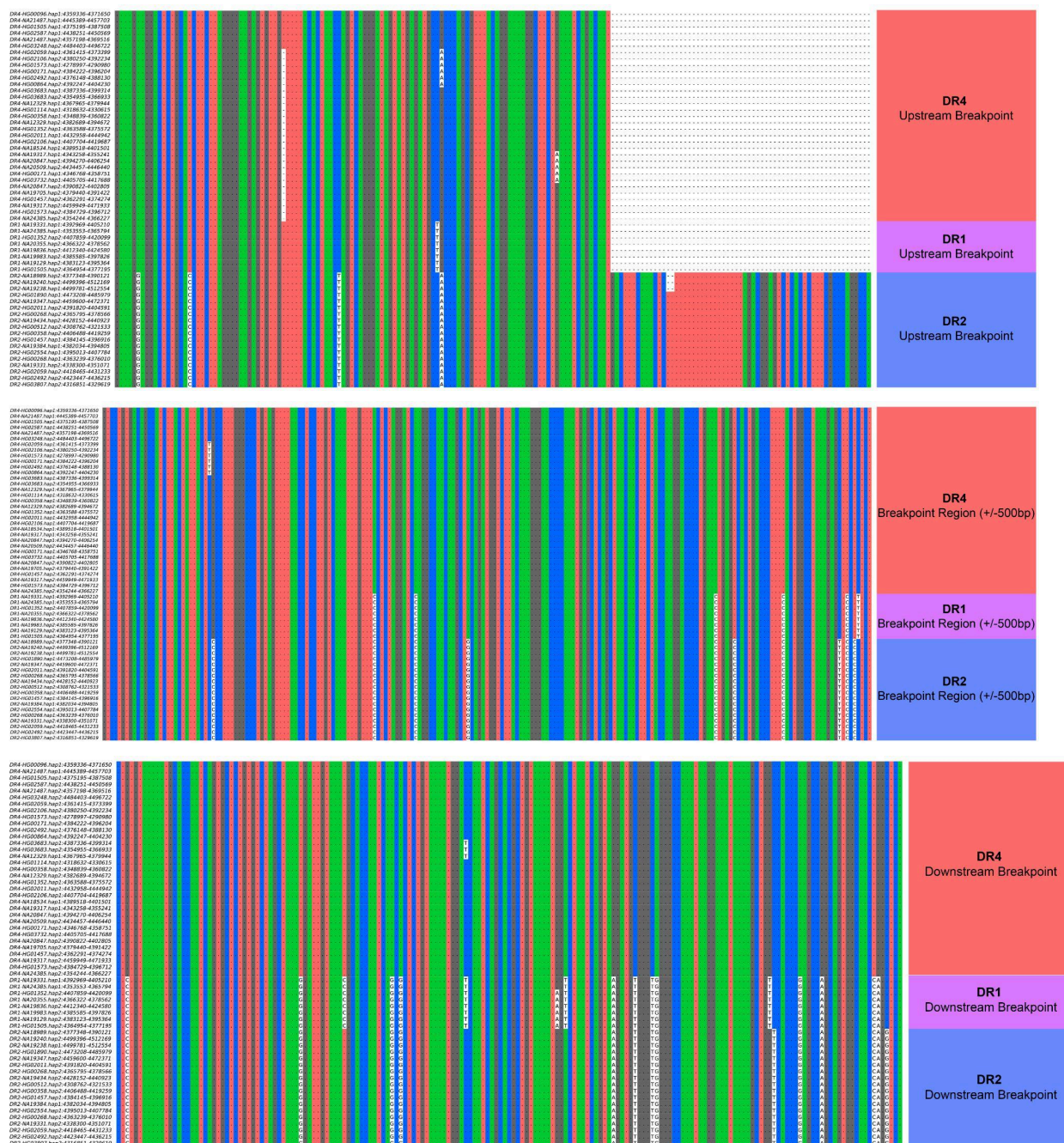
Supplementary Figure 47. DR8 recombination. Plots showing the results of a multiple sequence alignment of all haplotypes containing DR3/5/6 (*HLA-DRB3* and *HLA-DRB1* genes) and DR8 (*DRB1*). DR8 haplotypes are highlighted in purple (on right) and individual exons were placed in the alignment to assist the viewer. The top figure panel shows the alignment upstream of the identified breakpoint region. In this panel DR8 clearly shares more sequence similarity (more SNVs) with the *DRB3* gene of DR3/5/6 than it does to the *DRB1* gene. The middle figure panel shows the breakpoint region around exon 4/intron 3. Multiple haplotypes contain *DRB3* and *DRB1* genes which share perfect homology while others display shorter stretches of sequence homology. After exon 4 is where the *DRB1* gene in DR8 starts to share SNVs with the *DRB1* gene of DR3/5/6. The bottom figure panel shows the sequence immediately downstream of the middle panel and how the SNV pattern has switched from that position onward. The putative breakpoint showed a short length (<150bp) of homology between *HLA-DRB1* and *HLA-DRB3* on DR3/5/6 spanning exon 4 and the surrounding downstream sequence (intron 3) in some DR8 haplotypes suggesting that homology-mediated double strand DNA break repair may be a more likely explanation than NAHR. However, due to the putative ages of these haplotypes²⁰⁷, as well as due to a subset of DR8 haplotypes displaying longer stretches (>200bp) of sequence homology, NAHR cannot be entirely discounted.

Fig. 48



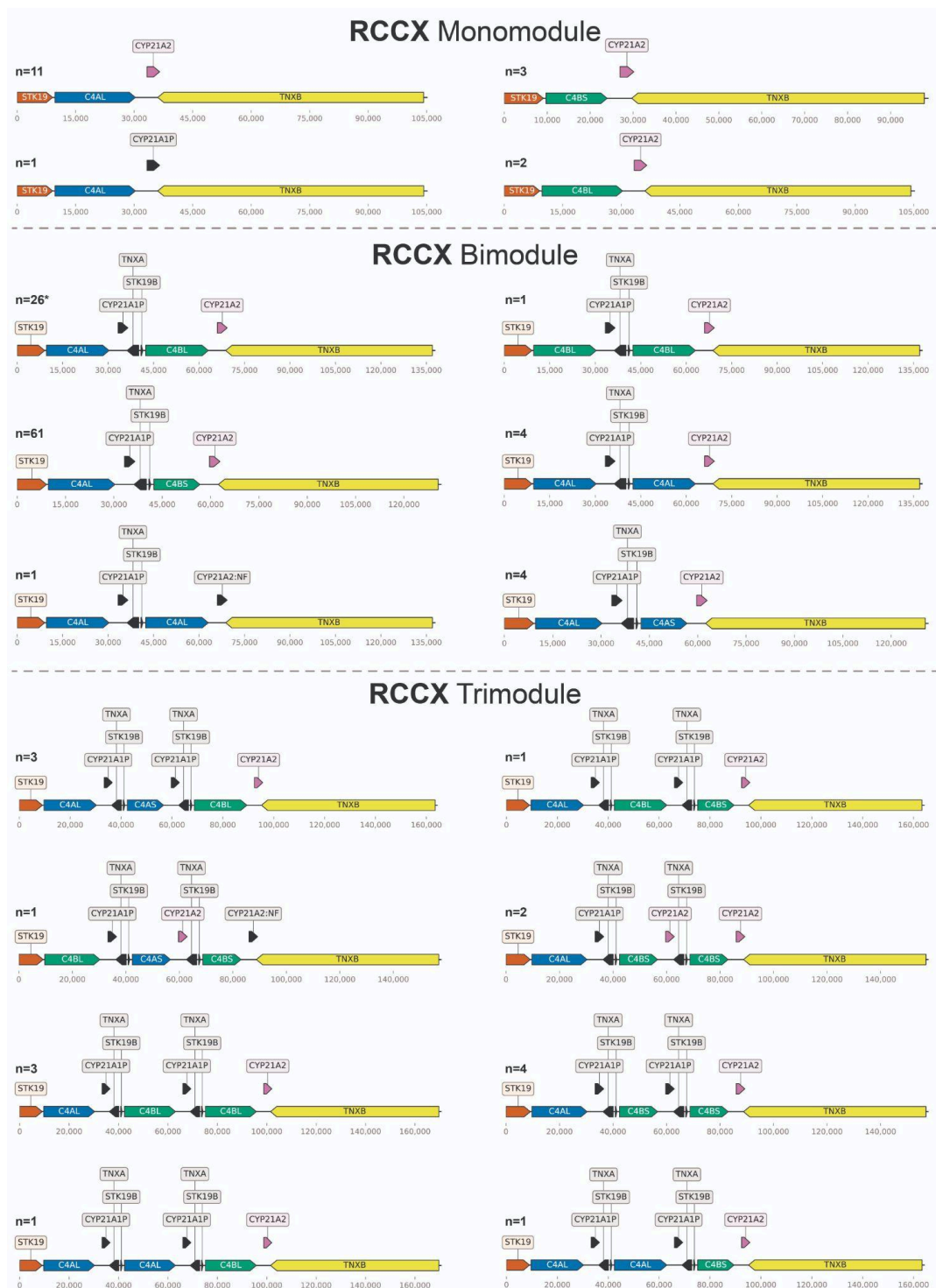
Supplementary Figure 48. Comparison of DR group haplotypes. Representative examples of pairwise alignments between DR group haplotypes DR5 and DRB3/6 (panel A); DR8 and DR1 (panel B); DR8 and DR3/6 (panel C); DR9 and DR4/7 (panel D) respectively. Red shading indicates the positions of HLA-DRB genes, blue/green shading matches to the identified HLA-DR small structural variants carrying solitary HLA-DRB exons.

Fig. 49



Supplementary Figure 49. DR1 recombination. These figures show the results of a multiple sequence alignment around the identified breakpoint region (+/-500bp) within all sample haplotypes containing DR4, DR1, and DR2. The top figure panel starts upstream of the breakpoint region, within a LINE/L1 element. The gap in DR4 and DR1 is the absence of an *A/u* insertion only present in DR2. The middle figure panel is further within the LINE/L1 element where we start to see a switch in the SNV pattern to where DR1 and DR2 are sharing more SNVs with one another. This marks the +/- 500 base pair region where the breakpoint is believed to have occurred. The bottom figure panel is downstream this region showing how DR1 and DR2 continue to share SNVs.

Fig. 50



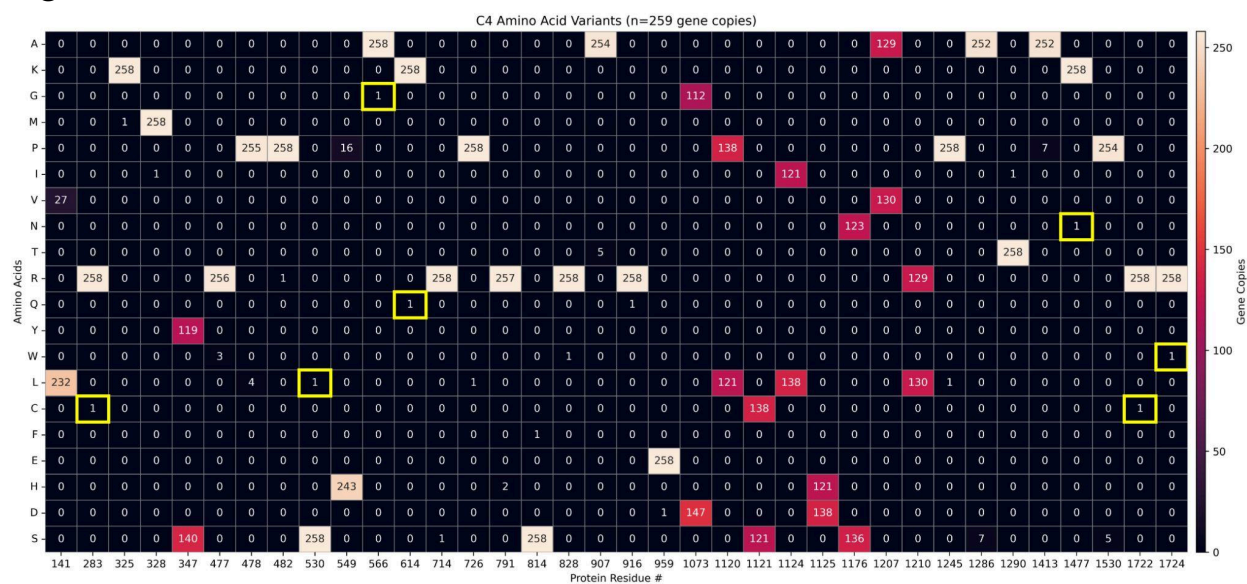
Supplementary Figure 50. RCCX Architecture. This schematic displays all unique RCCX mono/bi/tri module architectures found across the 130 haplotypes. Protein coding genes are *STK19*, *C4*, *CYP21A2*, and *TNXB*. Pseudogenes are *CYP21A1P*, *TNXA*, and *STK19B*. All C4 copies are functional. *CYP21A2:NF* represents a *CYP21A2* gene copy that is no longer functional due to coding for a premature stop codon. The long (L) or short (S) C4 designation

represents if a HERV-K insertion within intron 9 is present. An asterisk (*) denotes the RCCX structure present within the CHM13 reference genome. The most frequently observed representatives of each haplotype configuration carried one functional copy of the *STK19*, *CYP21A2* and *TNXB* genes, as well as *C4AL*, *C4BS* and one copy of the pseudogenes *STK19B*, *CYP21A1P*, and *TNXA* (for n = 61 out of a total of 97 bi-modulesar modules); *C4AL* and no pseudogenes (for n = 11 out of a total of 17 mono-modules); and *C4AL*, 2 copies of *C4BS*, and two copies of the pseudogenes *STK19B*, *CYP21A1P*, and *TNXA* (for n = 4 out of a total of 16 tri-modules).

[illegible]

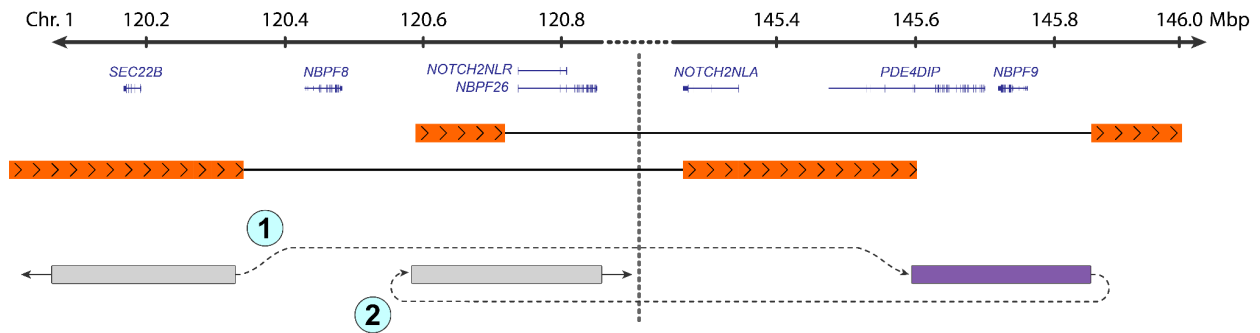
129

Fig. 52



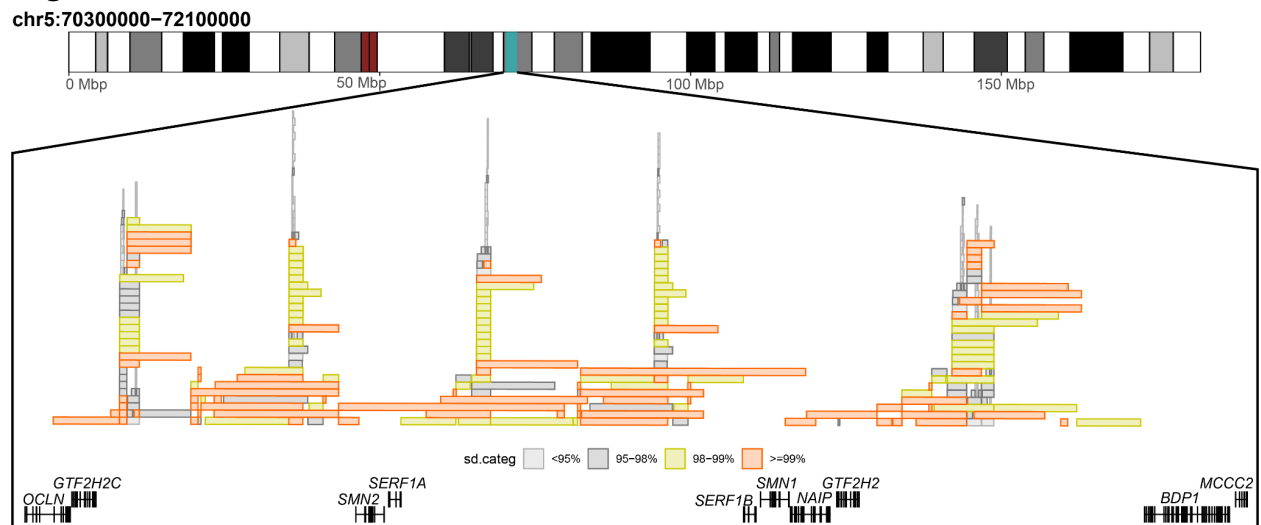
Supplementary Figure 52. C4 Novel Variants. Heatmap showing the variable amino acid positions within our 259 *C4* gene copies (across 130 haplotypes). The x-axis is the specific amino acid position while the y-axis are amino acids represented by a single letter. The seven novel amino acids found in our dataset are outlined in yellow. Each of these variants were only identified in a single *C4* copy.

Fig. 53



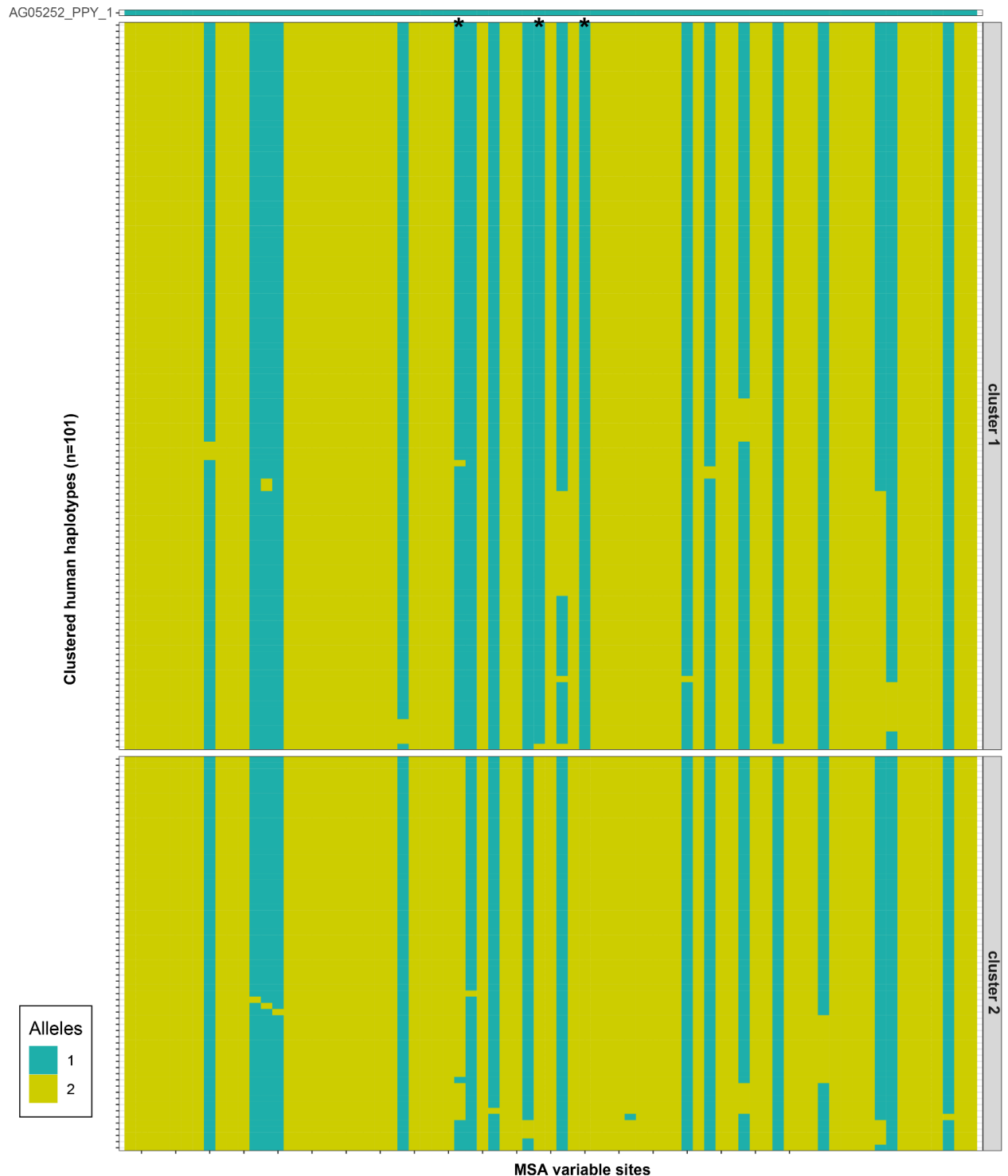
Supplementary Figure 53. SD-mediated CSV substitutes *NBPF8* with *NBPF9*. A pair of SDs each with >99% identity (paired orange boxes) mediate two template switches 25 Mbp apart causing a deletion of *NBPF8* and a duplicate copy of *NBPF9* to be inserted at the *NBPF8* locus. Gray bars: flanking reference alignments. Violet bar: CSV templated site. Coordinates and annotations are derived from T2T-CHM13.

Fig. 54



Supplementary Figure 54. SMN region overview. An overview of the SMN region that is highlighted on the chromosome 5 ideogram as a blue rectangle. Below we plot the extent of segmental duplications in this region color by the sequence identity (sd.cat). Further below we plot the gene model of protein-coding genes in this region.

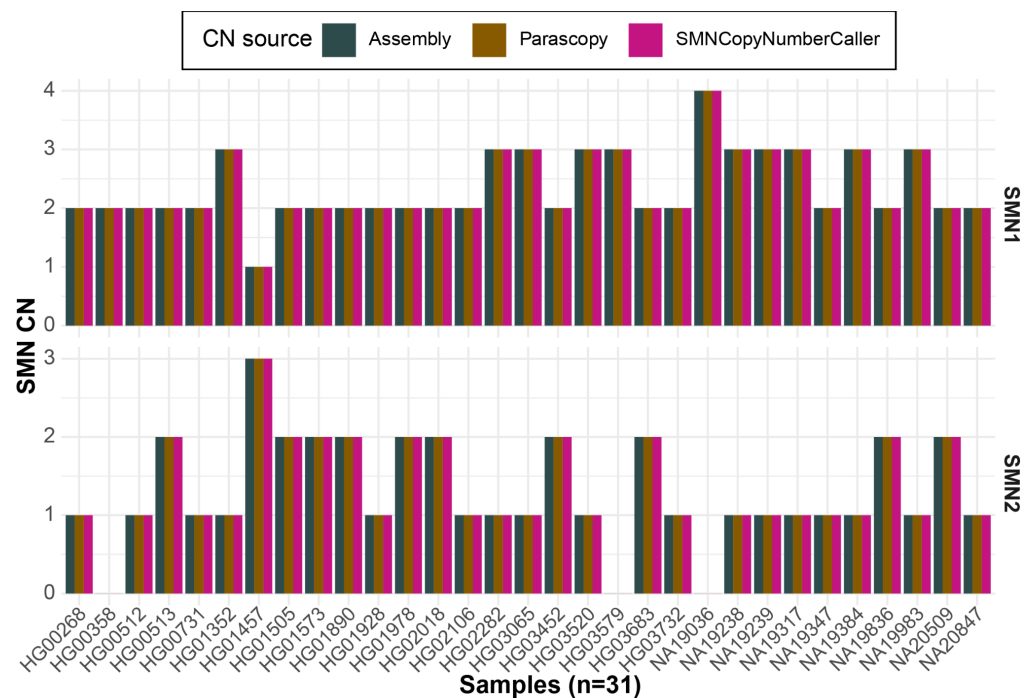
Fig. 55



Supplementary Figure 55. Distinguishing *SMN1* and *SMN2* gene copies. A visualization of a multiple sequence alignment between exonic sequences for all SMN gene copies across all 101 human haplotypes. We used orangutan (PPY) as an outgroup to mark ancestral (green) and derived alleles (yellow). We cluster all haplotypes into two clustered mostly distinguished based

on 3 ancestral alleles present in cluster 1. We mark cluster 1 as *SMN1* gene copy while cluster 2 is marked as *SMN2* gene copy.

Fig. 56

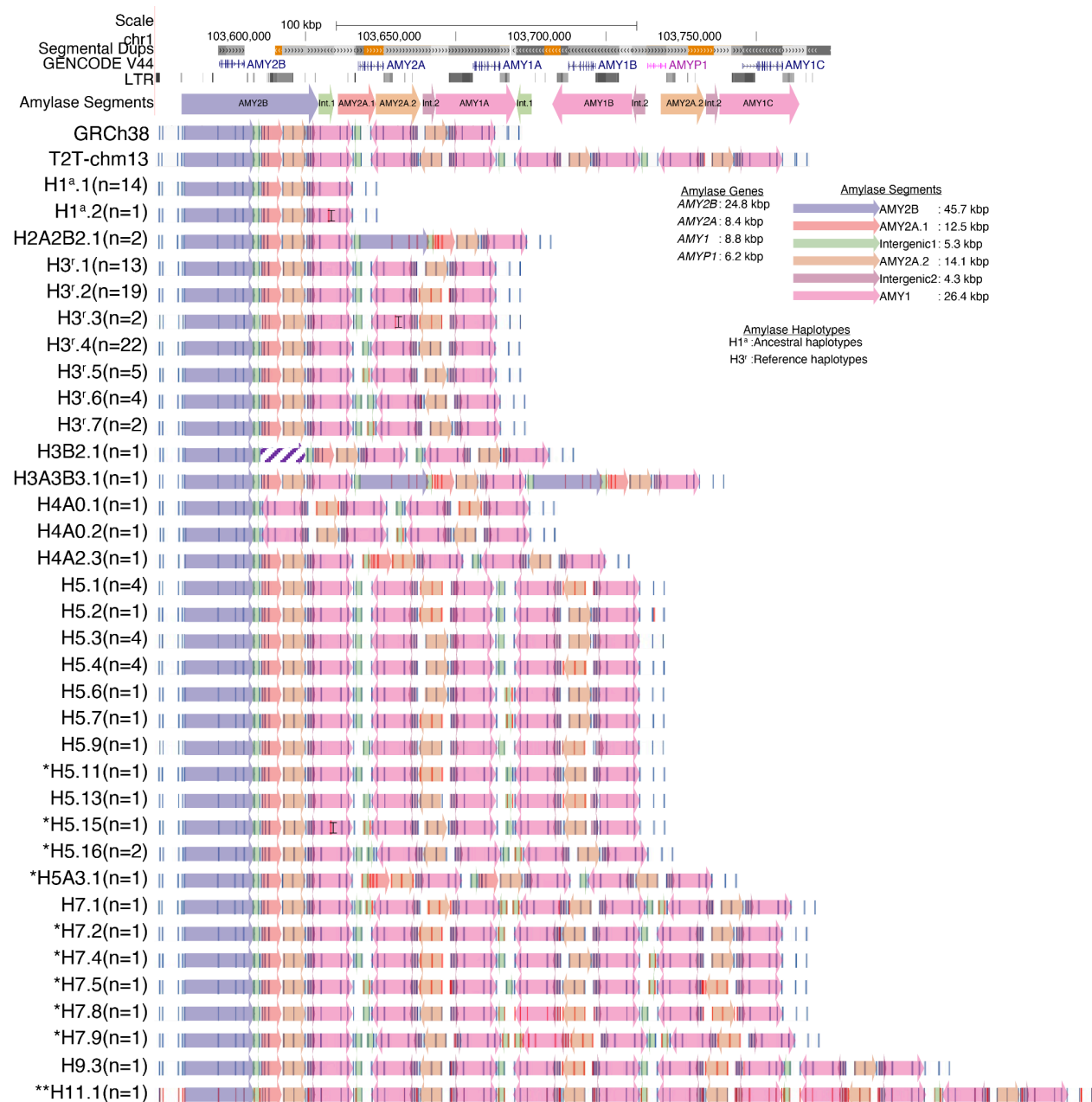


Supplementary Figure 56. Genotyping of *SMN* genes. Comparison of predicted *SMN1* and *SMN2* gene copy numbers defined by assemblies (dark blue) in comparison to short-read based genotyping methods (Parascopy - brown, SMNCopyNumberCaller - pink). We could compare results only for 31 samples which have fully assembled both haplotypes and could be compared to diploid copy number reported by unphased short-reads.

Figure 1 displays a phylogenetic tree and a corresponding haplogroup distribution plot. The tree on the left shows the relationships between various populations, with color-coded nodes indicating superpopulations: AFR (African), AMR (Admixed American), EAS (East Asian), EUR (European), NHP (Non-Human Primate), and SAS (South Asian). The right panel is a haplogroup distribution plot showing the count of SMN1 (yellow) and SMN2 (dark red) haplogroups for each population. The y-axis represents the SMN count (0 to 3), and the x-axis represents the Assembly ID. A legend identifies the superpopulations and SMN haplogroups. An asterisk marks a specific node in the tree.

136

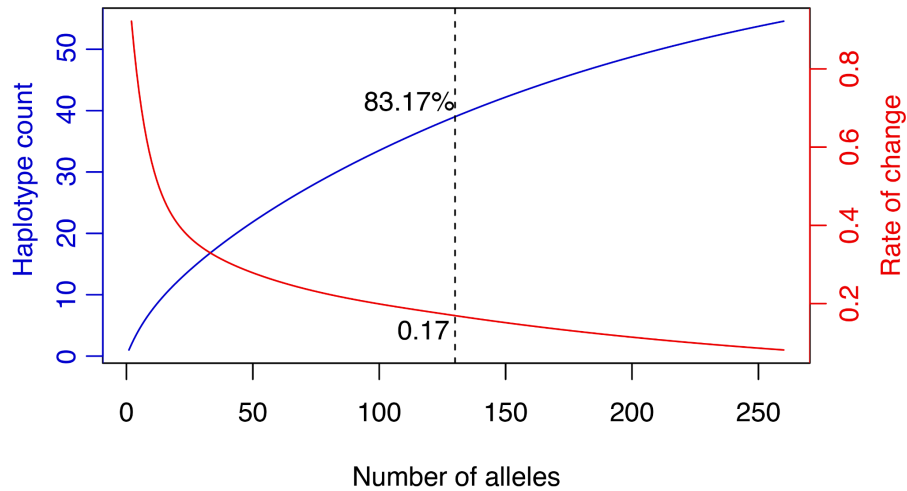
Fig. 58



Supplementary Figure 58. Amylase haplotypes identified in this study. Segmental duplications (light to dark gray: 90-98% similarity, light to dark orange: > 99% similarity), GENCODE V44 gene annotations, and long terminal repeats (LTRs) are represented as tracks. The lower panel shows amylase segments (colored arrows). The AMY2B segment overlaps the *AMY2B* gene, AMY2A.1 and AMY2A.2 segments overlap the *AMY2A* gene, and the AMY1 segment overlaps the *AMY1* gene. The GRCh38 and T2T-chm13 reference genomes and the unique amylase haplotypes resolved in our dataset (n = 35) are represented with *in silico* maps with white background displaying label positions in dark blue and amylase segments with color-coded arrows. The vertical black line in the second AMY1 segment of H3^r.3 represents the polymorphic label present in three alleles. The diagonal stripes in the second AMY2B segment

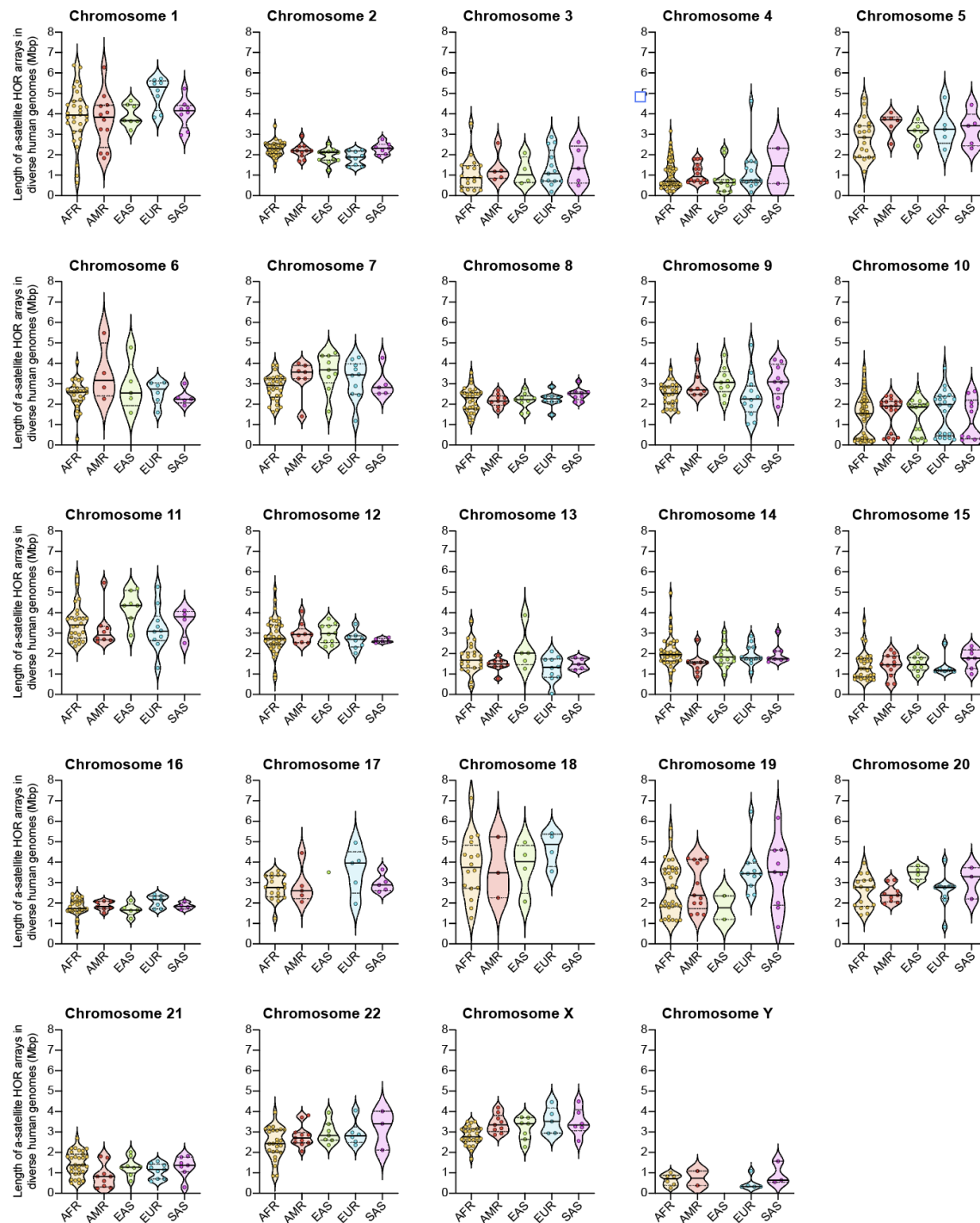
of H3B2.1 indicate that it is a partial copy of the first *AMY2B* segment. Haplotype IDs describe: HX: X denotes the number of *AMY1* copies; AX: X denotes the number of *AMY2A* copies; BX: X denotes the number of *AMY2B* copies. The superscript "a" denotes the ancestral amylase haplotype structure, and the superscript "r" denotes the reference amylase haplotype structure. The number in parentheses indicates the number of alleles. Single asterisk (*) denotes haplotypes that were resolved at the base-pair level for the first time. Double asterisk (**) denotes novel amylase haplotype.

Fig. 59



Supplementary Figure 59. Rarefaction and extrapolation curves based on 39 unique amylase haplotypes. The number of distinct haplotypes (blue line) approaches saturation with increasing alleles. The rate of change (red line) indicates the number of previously unknown haplotypes discovered per unit increase in the number of analyzed alleles. The dashed line represents 83.17% of the estimated haplotypes captured in this study.

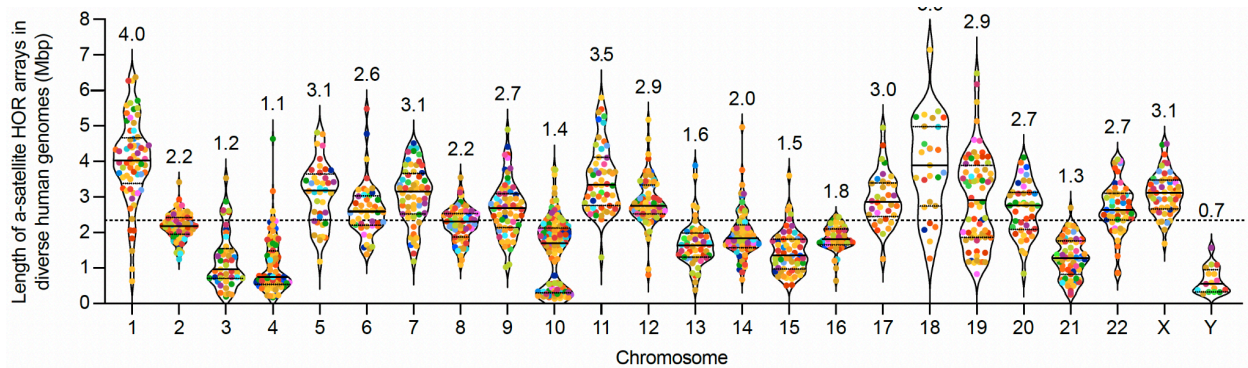
Fig. 60



Supplementary Figure 60. Most continental groups have similar active α -satellite HOR array lengths for each chromosome, with two exceptions on chromosomes 1 and 17. Distribution of active α -satellite HOR array lengths for each chromosome, stratified by continental group. We find no clear population bias in α -satellite HOR array length among most chromosomes, except for chromosomes 1 and 17, where Europeans tend to have larger α -satellite HOR arrays than those from other continental groups. This is statistically significant

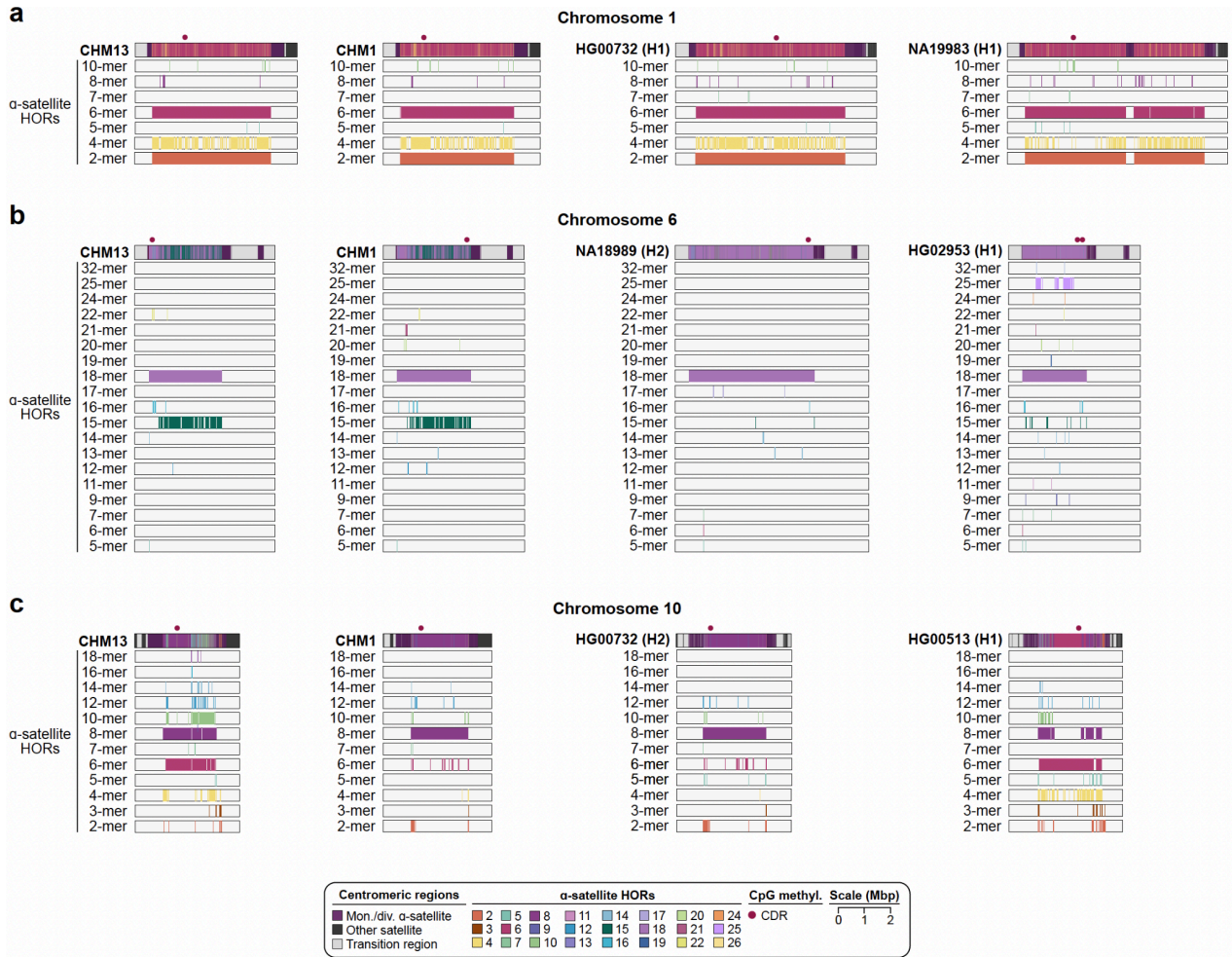
for chromosome 1 when comparing Europeans to Africans and East Asians ($p=0.0438$ and 0.0112 , respectively; two-tailed Kolmogorov-Smirnov test). However, for chromosome 17, this is not statistically significant ($p=0.0765$, 0.3571 , and 0.03571 when comparing Europeans to Africans, Americans, and South Asians, respectively; two-tailed Kolmogorov-Smirnov test).

Fig. 61



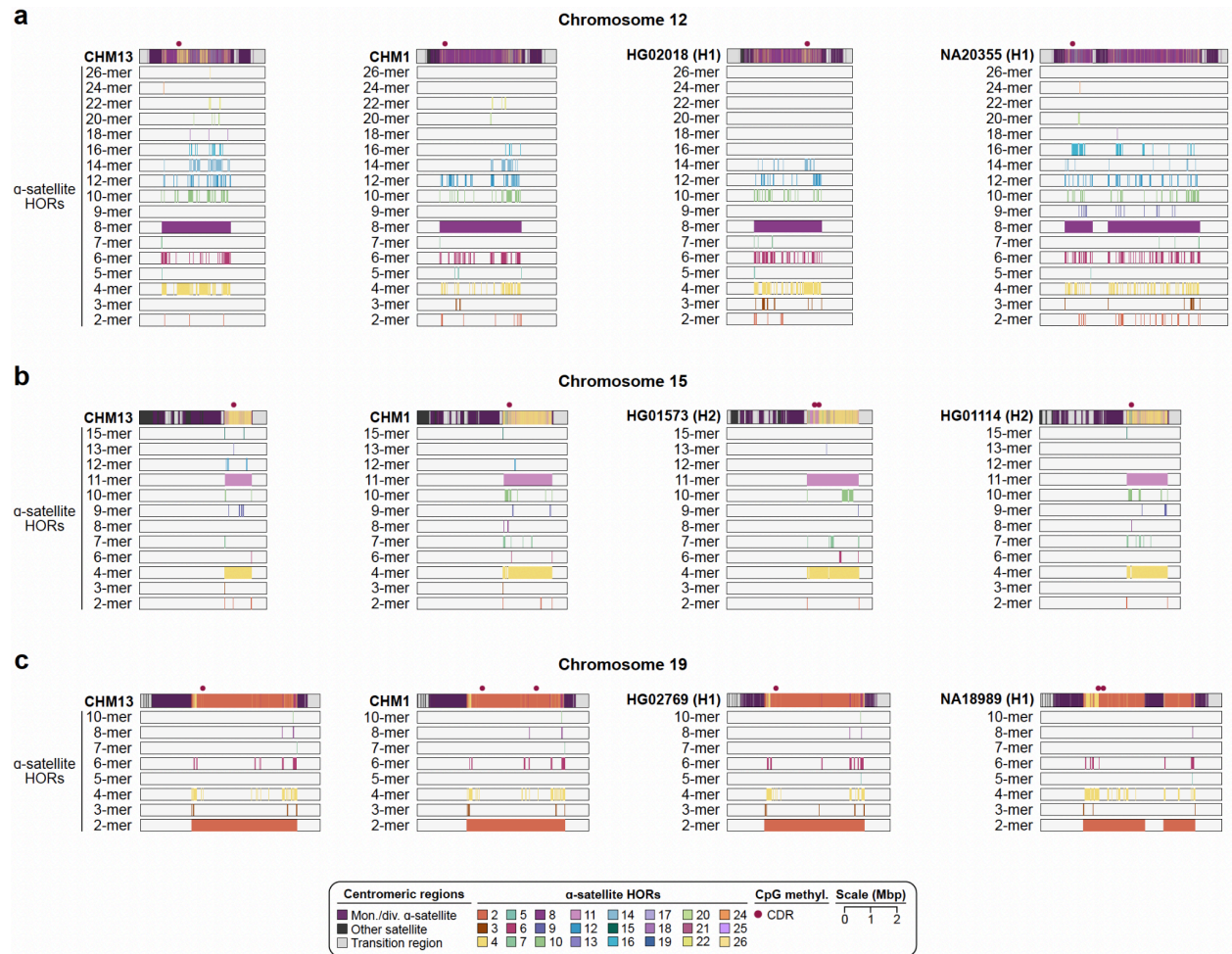
Supplementary Figure 61. Variation in active α -satellite HOR array lengths among 1,246 human centromeres. Length of the active α -satellite higher-order repeat (HOR) array(s) for each complete and accurately assembled centromere in this study. Each data point indicates an active α -satellite HOR array and is colored by population. The median length of all active α -satellite HOR arrays is shown as a dashed line. For each chromosome, the median (solid line) and first and third quartiles (dashed lines) of the active α -satellite HOR arrays are shown, and the mean active α -satellite HOR array length is shown on top in Mbp.

Fig. 62



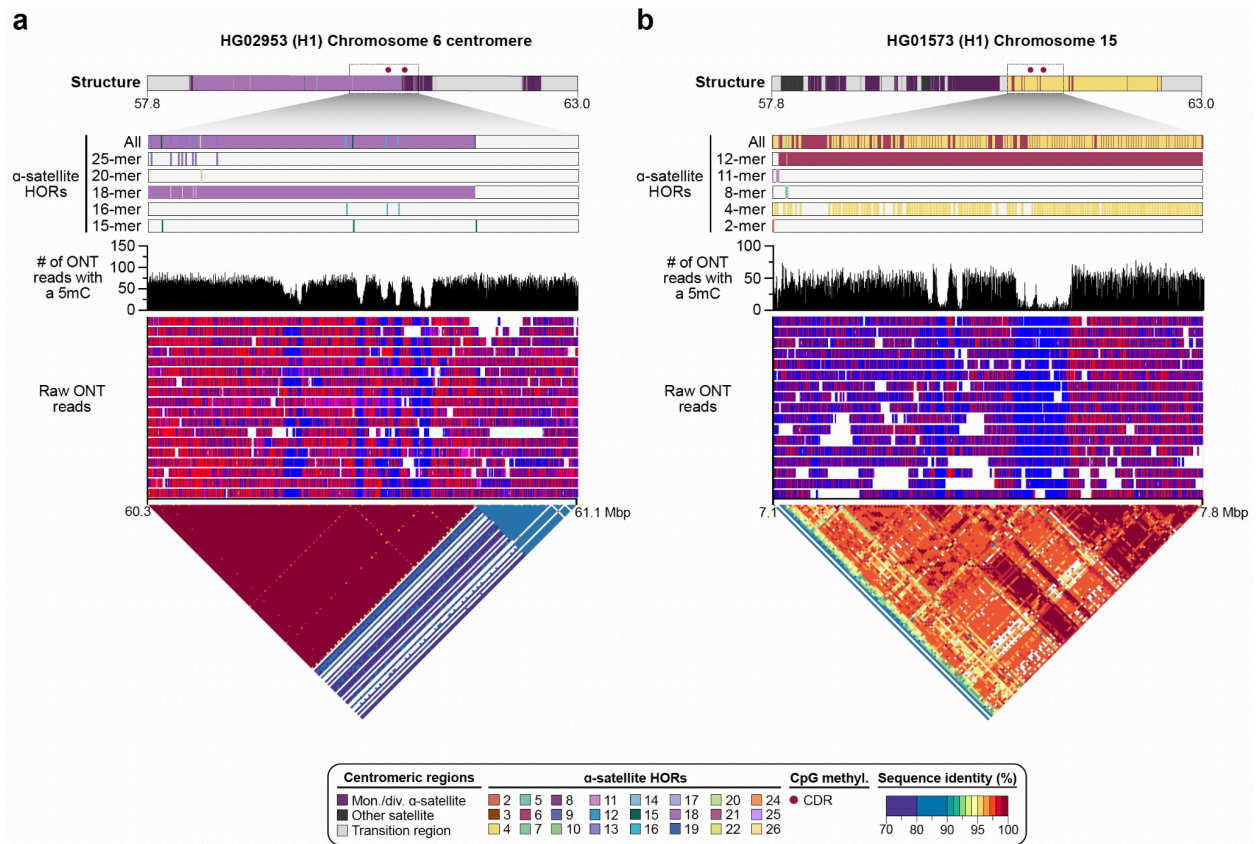
Supplementary Figure 62. Variation in α -satellite HOR array organization for select centromeres. Organization of α -satellite HOR variants for centromeres from chromosomes **a)** 1, **b)** 6, and **c)** 10 as shown in Fig. 6b. Organizations of α -satellite HOR variants for centromeres from chromosomes 12, 15, and 19 are shown in Supplementary Figure 63.

Fig. 63



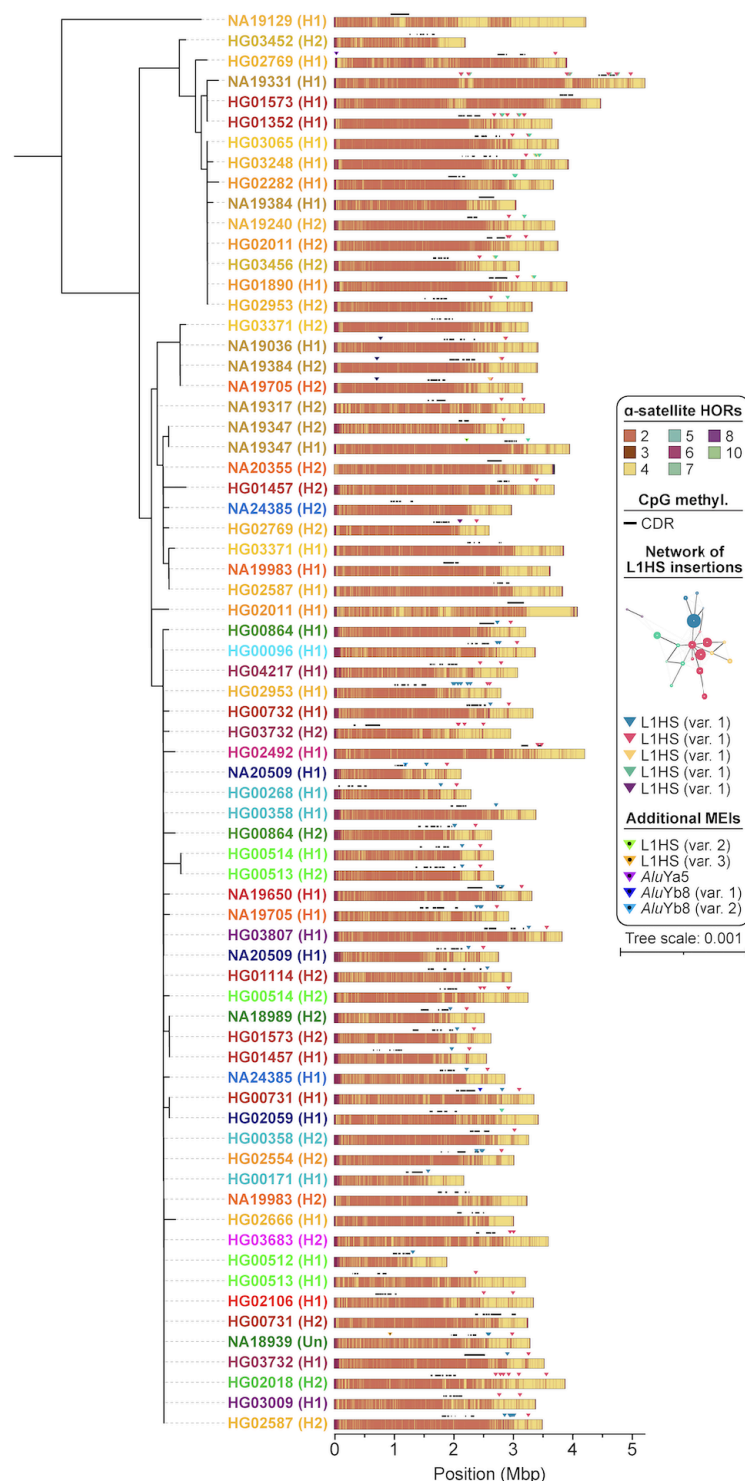
Supplementary Figure 63. Variation in α -satellite HOR array organization for select centromeres. Organization of α -satellite HOR variants for centromeres from chromosomes a) 12, b) 15, and c) 19 as shown in Fig. 6b. Organizations of α -satellite HOR variants for centromeres from chromosomes 1, 6, and 10 are shown in Supplementary Figure 62.

Fig. 64



Supplementary Figure 64. Di-kinetochores reside within highly identical regions of the centromere. Di-kinetochores, defined as two hypomethylated regions present on the same ONT read and located within 80 kbp of each other, reside within highly identical regions of the **a)** HG02953 chromosome 6 centromere and **b)** HG01573 chromosome 15 centromeres.

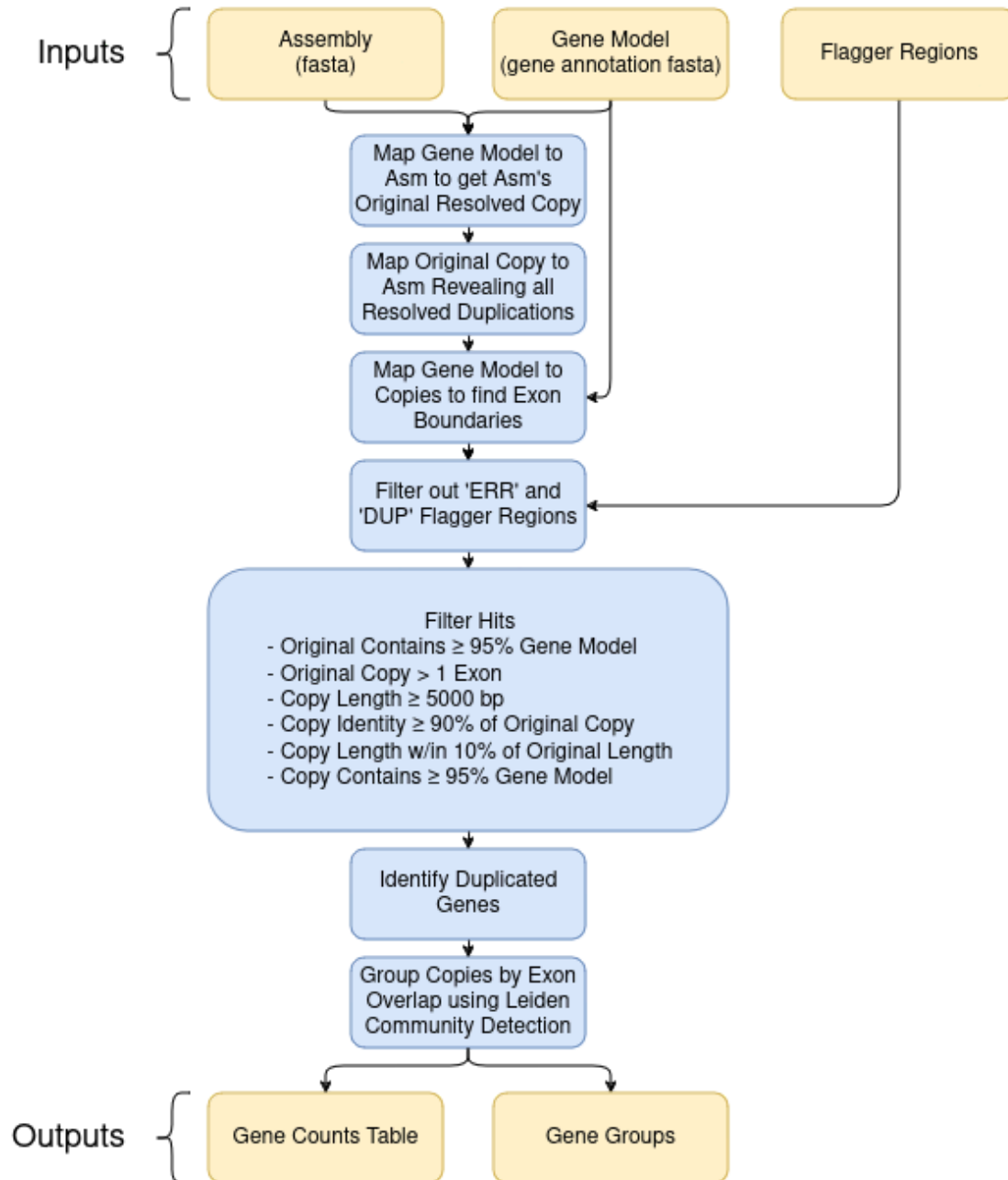
Fig. 65



Supplementary Figure 65. An undirected network of MEIs in the chromosome 2 centromeric α -satellite HOR array reveals sequence similarity between the L1HS (variant 1) insertion and its duplications across haplotypes. MEI network nodes represent unique SNV profiles for the L1HS (variant 1). A solid black edge between two nodes indicates that one node was the other's closest match. A dashed white edge outlined with a black stroke between

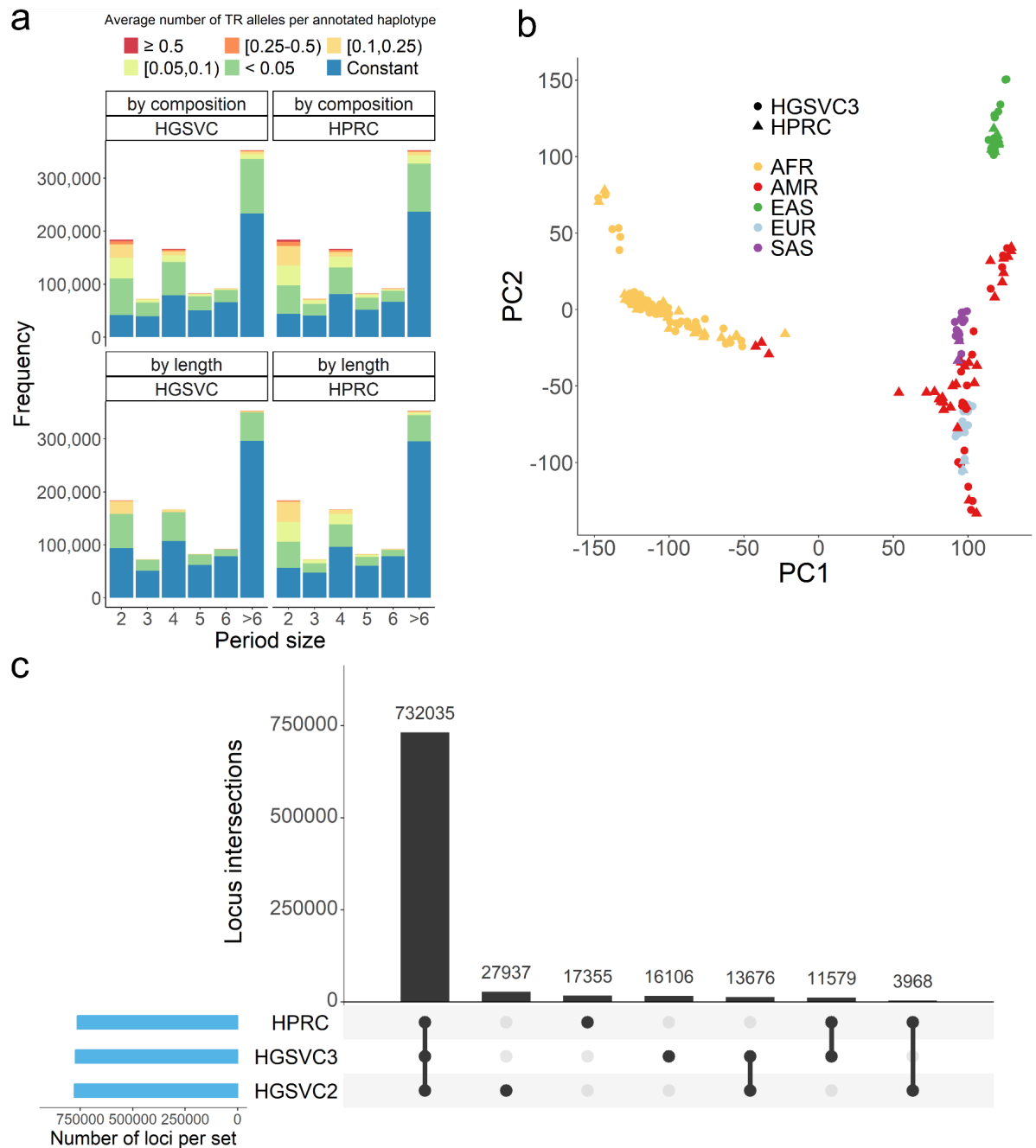
two nodes indicates that one node was the other's second closest match. In both cases, ties are allowed and shown as multiple edges connected to a node. The size of a node is representative of the total L1HS (variant 1) copies with an identical sequence, and the color of the node represents the node community assignment. Most MEIs within the chromosome 2 α -satellite HOR array are consistent with duplications of the same element rather than distinct insertions, and all of them reside outside of the CDR (**Supplementary Methods**).

Fig. 66



Supplementary Figure 66. Segmental duplication annotation pipeline flowchart. Minimap2 was used for all mapping and alignment.

Fig. 67



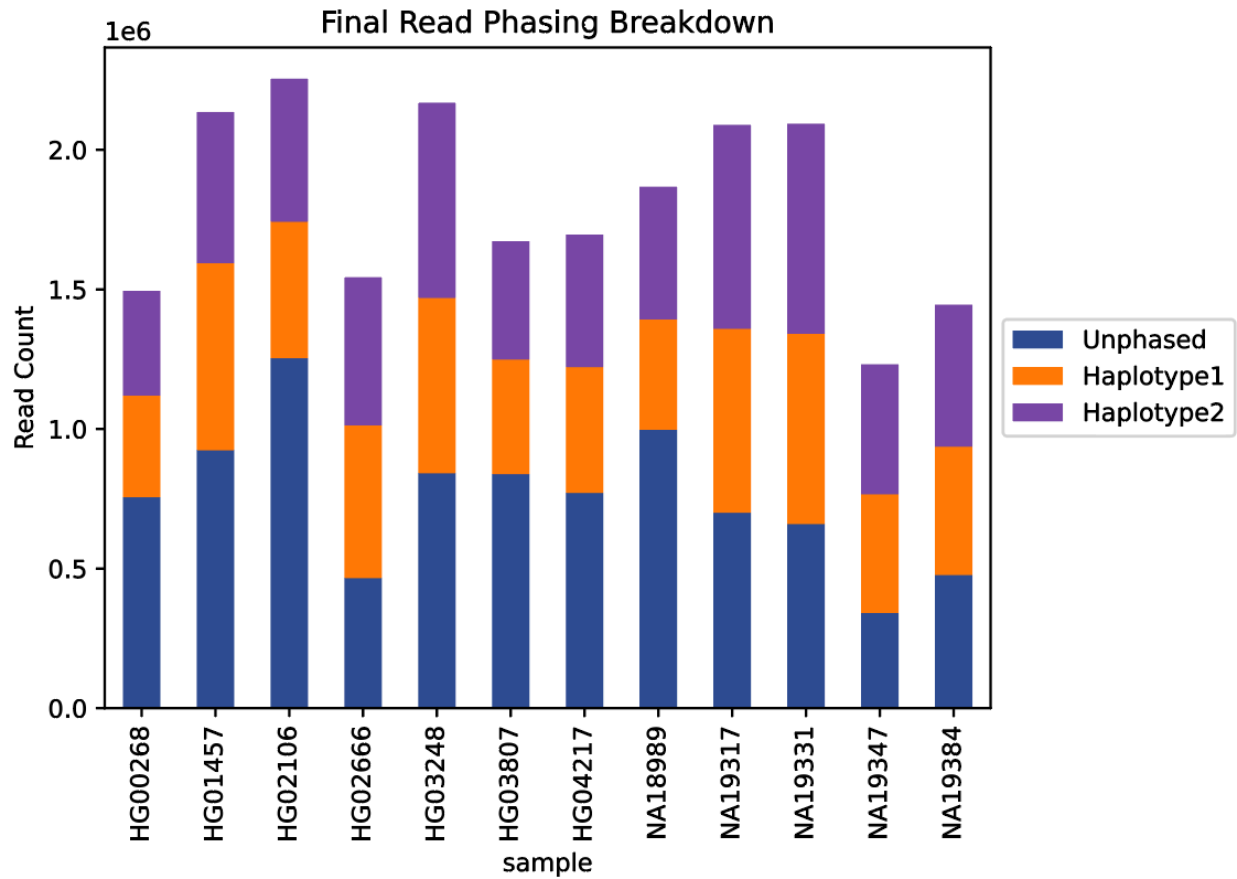
Supplementary Figure 67. VNTR diversity. a: variability of TRs in HPRC and HGSVC3 genomes. b: population clustering of HPRC and HGSVC3 genomes by TRs. c: TRs from Tandem Repeat Finder were merged and compared for the three datasets. Centromere regions were masked.

Fig. 68



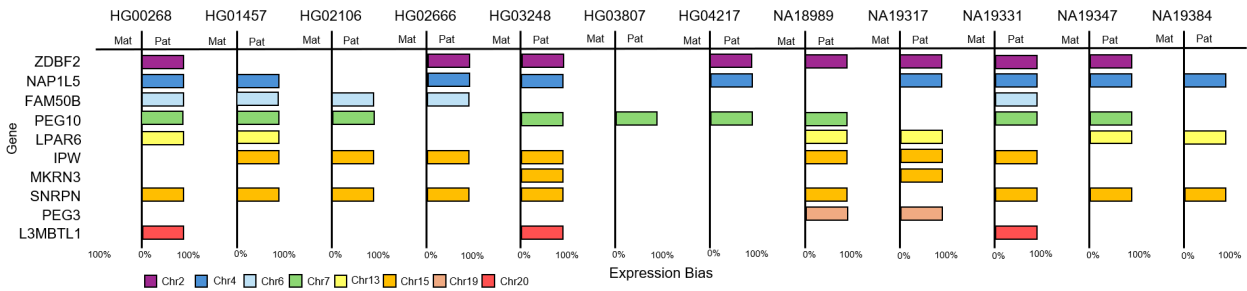
Supplementary Figure 68. Unique TRs in HGSCV3 compared to HPRC. Distribution of unique TRs by Tandem Repeat Finder in HGSCV3 genomes compared to HPRC genomes. Ideograms were generated using a bin size of 1,000,000bp and centromere regions were masked.

Fig. 69



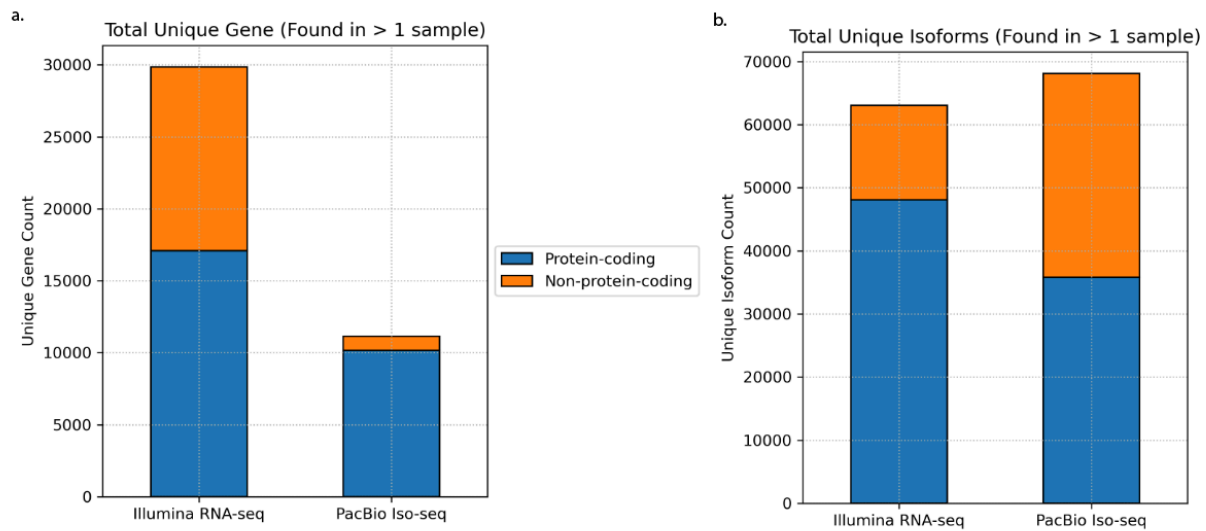
Supplementary Figure 69. Iso-Seq Read Phasing Breakdown. The total number of PacBio Iso-Seq read phasing assignments to haplotype 1 (orange), haplotype2 (purple), and unphased (unable to be phased) (navy) per sample using an ensemble alignment-based phasing approach (**Methods**). Reads designated as unphased aligned to both haplotypes equally well. On average 58.44% of reads were able to be phased to a haplotype of a sample.

Fig. 70



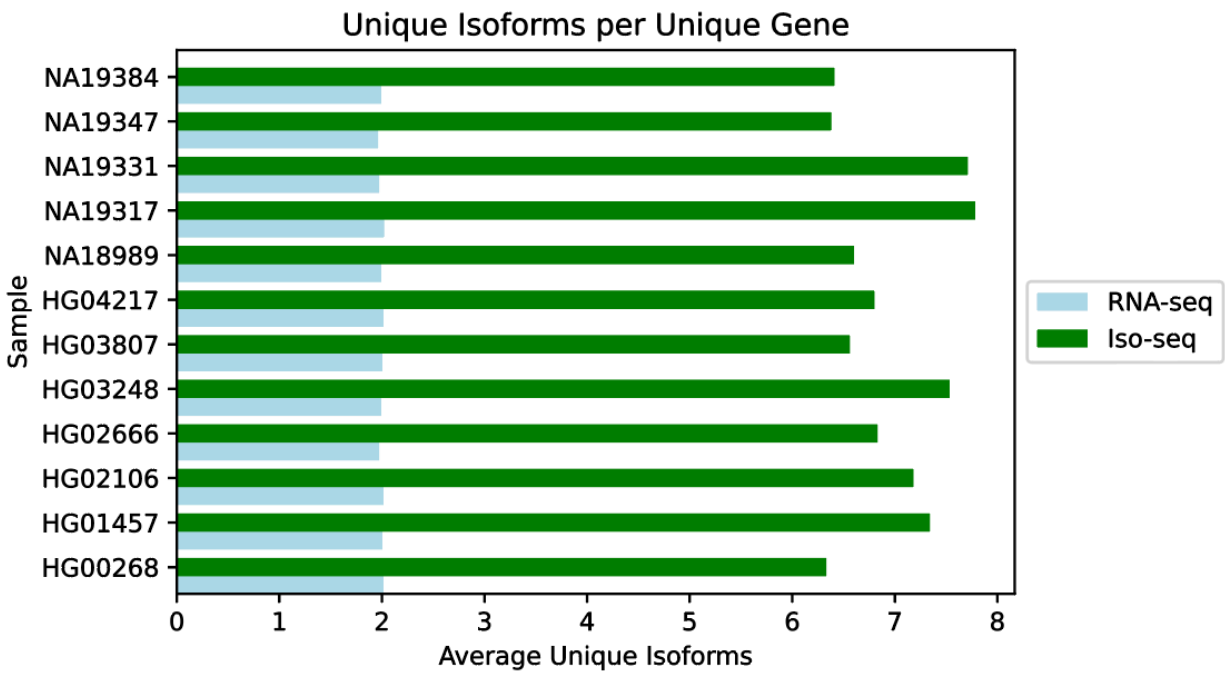
Supplementary Figure 70. Imprinted Genes According to Iso-Seq. Depiction of ten genes (y-axis) expressed in EBV-transformed B-lymphocyte cell lines that were identified as imprinted utilizing phased sample-specific assemblies for mapping of Iso-Seq reads. Mono-haplotype expression was declared if sample-specific reads were phased to a single haplotype. Bar color represents the chromosome on which the gene resides. The imprinted expression pattern observed in our study is in concordance with the previous literature^{169–171}.

Fig. 71



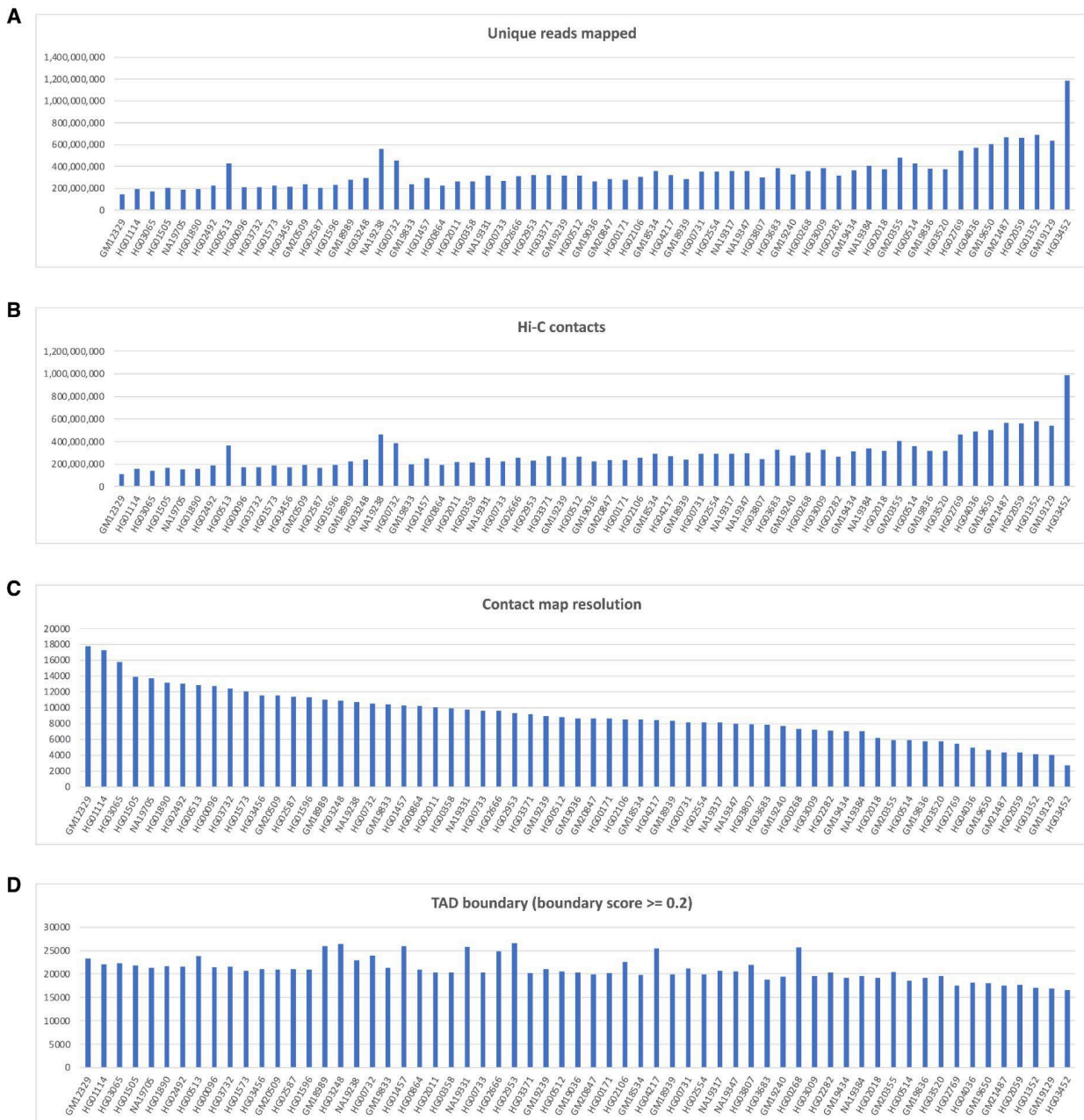
Supplementary Figure 71. Short- vs long-read RNA sequencing total unique genes and isoforms. **a)** Comparison of the total number of unique genes identified across 12 samples using Illumina RNA-seq vs PacBio Iso-Seq. Only genes expressing isoforms identified in two or more samples were included in this analysis. Hue indicates the proportion of genes that are protein coding (blue) or non-protein coding (orange). Results show that a greater total number of unique genes were identified by the Illumina RNA-seq (29,858 unique genes) versus the PacBio Iso-Seq (11,134 unique genes) across 12 samples. **b)** Comparison of the total number of unique isoforms present in at least two out of 12 samples in Illumina RNA-seq and PacBio Iso-Seq. Hue indicates the proportion of isoforms that are from protein-coding (blue) or non-protein-coding genes (orange). This graph indicates that a greater total number of unique isoforms were identified in the the PacBio Iso-Seq dataset (68,097 unique isoforms) compared to the Illumina RNA-seq callset (63,028 unique isoforms) across the 12 samples.

Fig. 72



Supplementary Figure 72. Average unique isoforms per unique gene. Comparison of the average number of unique isoforms identified (present in at least 2 out of 12 samples) per unique gene represented in Illumina RNA-seq (light blue) and PacBio Iso-Seq (green) data. On average, a greater number of unique isoforms per unique gene were found using the Iso-Seq versus the RNA-seq datasets.

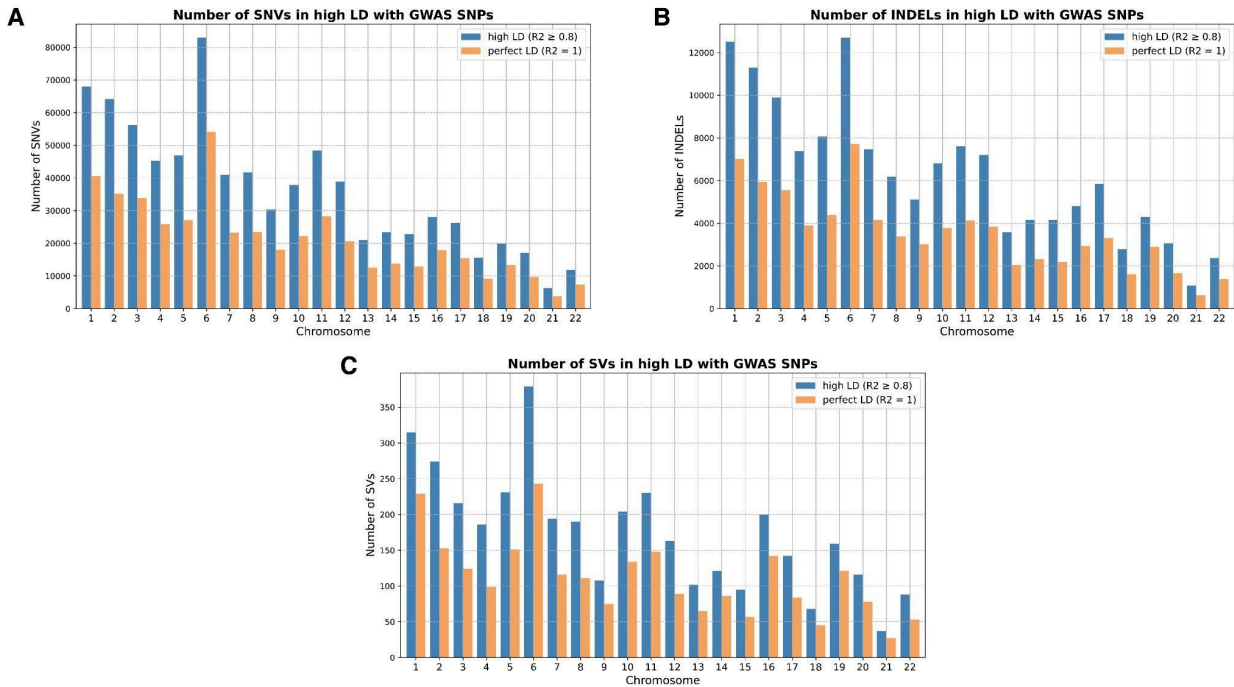
Fig. 73



Supplementary Figure 73. Summary statistics of Hi-C data and chromatin structures. A) Unique reads were mapped to the hg38 reference genome. The x-axis displays sample IDs ordered from lower to higher resolution (calculated based on the GRCh38). The y-axis shows the total number of the mapped reads. **B)** The number of Hi-C contacts for each mapped sample. **C)** The calculated resolution (in base pair) of Hi-C contacts for each sample. **D)** The number of TAD boundaries identified for each sample (filtered by the boundary score as 0.2). The average number of TAD boundaries identified was 21,021 for the hg38 reference genome. No significant correlation was observed between the number of TAD boundaries and contact

map resolution, underscoring the robustness of the /S algorithm, which is not constrained by the data resolution.

Fig. 74



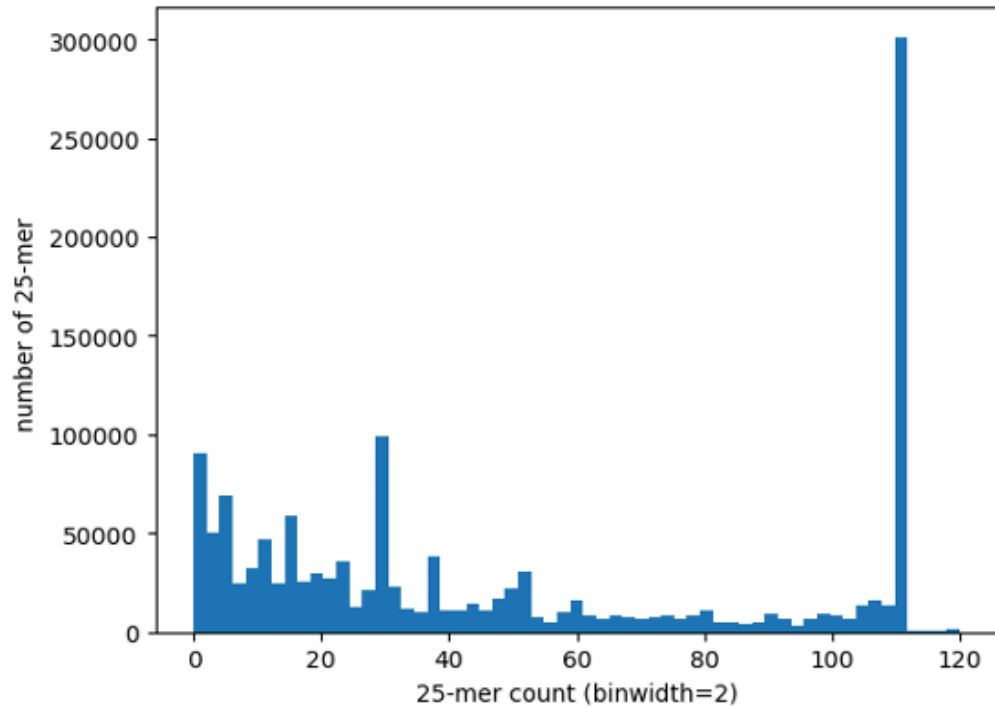
Supplementary Figure 74. SNVs, INDELs, and SVs that are in high LD with GWAS SNPs.
A) Number of SNVs in high LD with GWAS SNPs. X-axis represents chromosomes and the y-axis indicates the number of variants. Blue bars denote variants with an LD threshold of $R^2 \geq 0.8$. Orange bars denote variants in the perfect LD which has $R^2 = 1$. **B)** Number of INDELs in high LD with GWAS SNPs. **C)** Number of SVs in high LD with GWAS SNPs.

Fig. 75



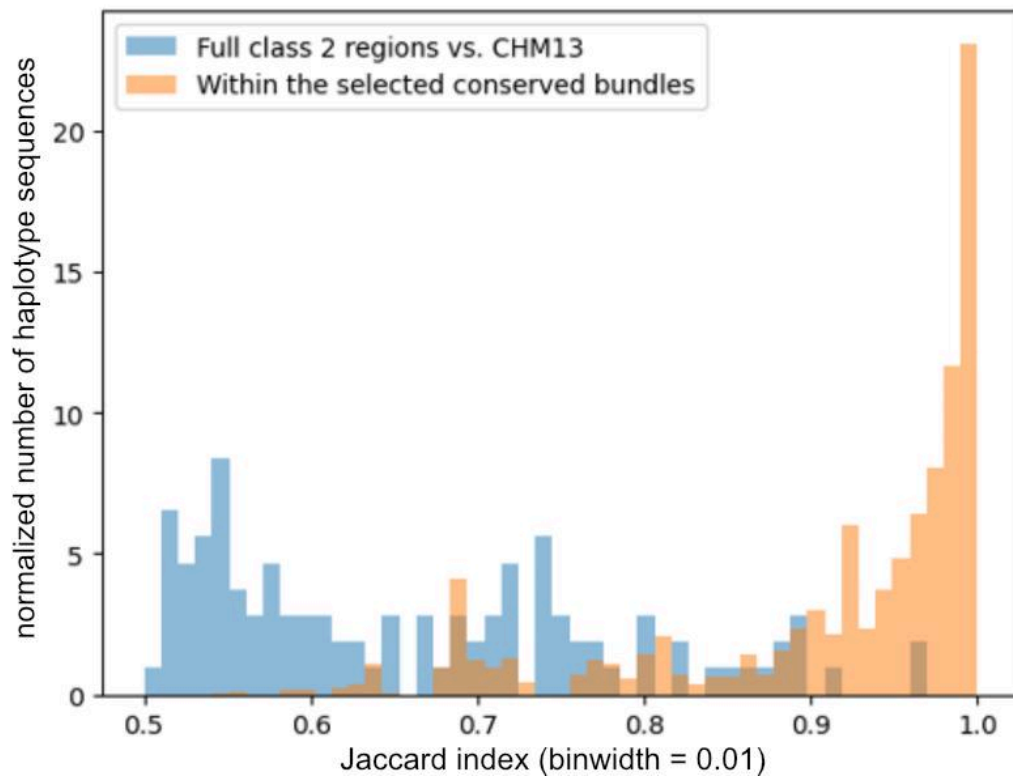
Supplementary Figure 75. MHC C2 alignment. The sequence alignment generated by minimap2 of the HGSVC genomes to the CHM-13 reference. The standard alignment process fails to align the high diversified regions showing as gaps and high density of variations around the gaps.

Fig. 76



Supplementary Figure 76. MHC C2 *k*-mer count histogram. All 25-mer count through the MHC class II region of the pangenome. While the majority of the 25-mers are from conserved (among different genomes) and unqued (in single genome) as shown by the peak around 25-mer count = ~111, there are abundance of less conserved *k*-mers ranged 1 to 50.

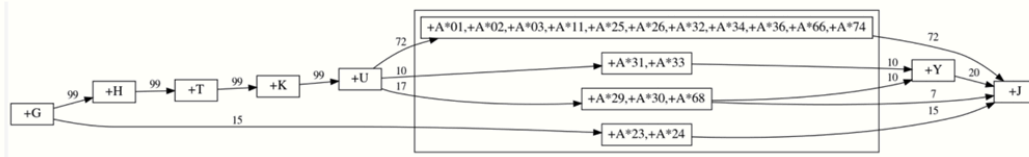
Fig. 77



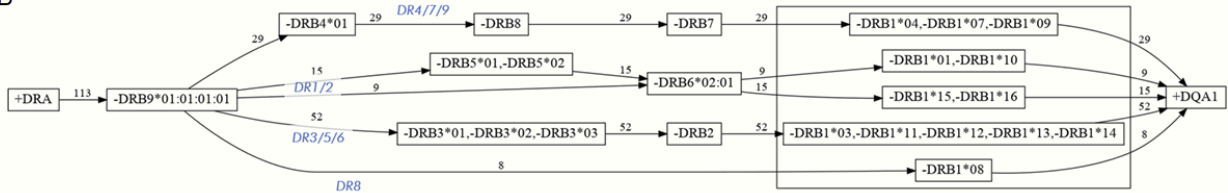
Supplementary Figure 77. Class II MHC vs CHM13 Jaccard index. The Jaccard index comparison of the class II regions vs. CHM13 and within the conserved bundles.

Fig. 78

A

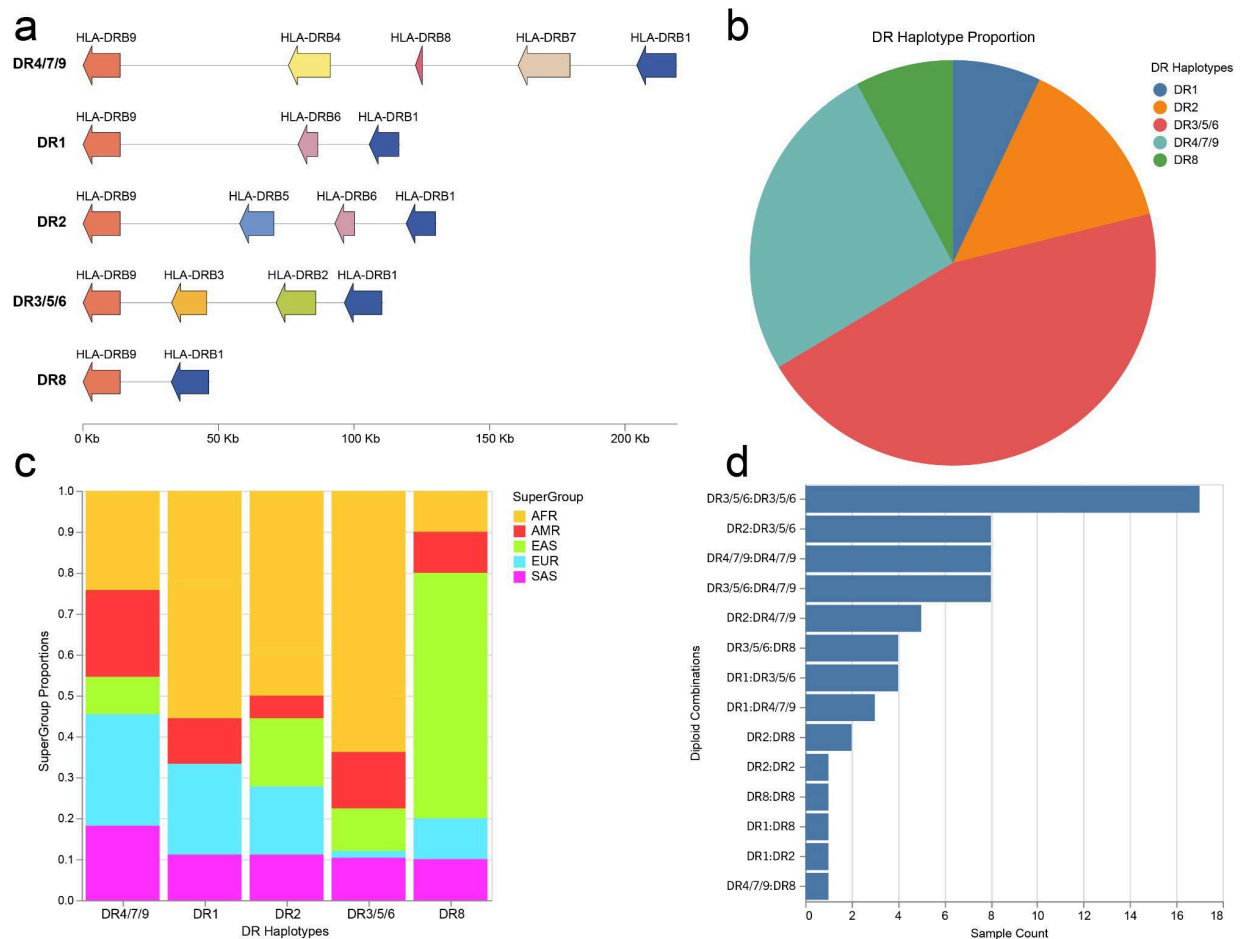


B



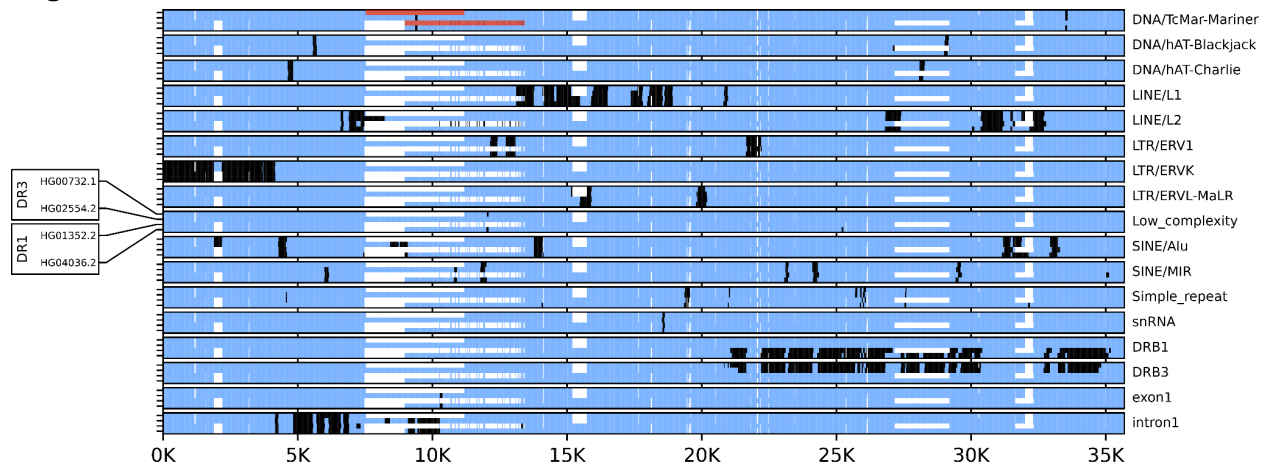
Supplementary Figure 78. MHC gene structure. Visualization of the MHC class I (panel A) and MHC class II (panel B) haplotype structure with respect to linkage equilibrium between HLA gene allele classes and presence/absence patterns, based on annotation with Immuannot²⁰³. Blue labels in panel B indicate which DR group the shown haplotypes belong to.

Fig. 79



Supplementary Figure 79. DR haplotype diagram and statistics. **a)** A diagram showing the typical length of the five unique DR haplotypes and the genes (*HLA-DRB1*, *-DRB3*, *-DRB4*, and *-DRB5*) as well as the pseudogenes (*-DRB2*, *-DRB6*, *-DRB7*, *-DRB8*, and *-DRB9*) which comprise them. **b)** A pie chart showing the frequency of DR haplogroups within the 128 sample haplotypes (2 haplotypes contained N's in the DRB region and were not analyzed). DR3/5/6 was the most frequently observed DR haplotype **c)** A stacked bar graph showing the proportion of sample supergroups (AFR: African, AMR: American, EAS: East Asian, EUR: European, SAS: South Asian) found containing each DR haplotype. Interestingly, no east asians (EAS) in our sample set contained a DR1 haplotype, though they made up the majority of the DR8 haplotype. **d)** A bar graph showing the total # of samples (x-axis) containing DR haplotype diploid combinations (y-axis). The diploid combination of two DR3/5/6 haplotypes within a genome was the most frequently observed across our sample set.

Fig. 80



Supplementary Figure 80. HLA SV insertion visualization. Visualization of haplotype structures and homology for two different loci in the HLA-DR region, both of which were found to contain small (~ 3.7 kbp and 4.4 kbp) structural variants with solitary intron and exon *HLA-DRB* sequences (see main text). The figure shows that the identified structural variants as well as their surrounding sequence regions from different loci in the HLA-DR region exhibit a high degree of homology, and that the structural variants carry solitary *HLA-DRB* intron and exon sequences as well as repeat elements. Shown is a multiple sequence alignment (MSA) of four sequences; each panel shows the same MSA, annotated with a specific sequence feature (e.g., repeat element positions, homologies to *HLA-DRB* intron and exon sequences; sequence features are shown in black, white indicates gaps relative to the longest sequence alignment). The included four sequences comprise a) two representative DR3 haplotypes from the region around the identified structural variant in the *HLA-DRB9* region, one carrying the insertion and one not carrying the insertion; b) two representative DR1 haplotypes from the region around the identified structural variant in the region 10 kbp 3' of *HLA-DRB1*, one carrying the insertion and one not carrying the insertion. The location of the identified structural variants (relative to other sequences from the same locus) is highlighted in red in the top panel.

References

135. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*. 2012;28:3326–3328.
136. Zhang Y, Zhang Y, Burke JM, Gleitsman K, Friedrich SM, Liu KJ, Wang T-H. A Simple Thermoplastic Substrate Containing Hierarchical Silica Lamellae for High-Molecular-Weight DNA Extraction. *Adv Mater*. 2016;28:10630–10636.
137. Logsdon G. HMW gDNA purification and ONT ultra-long-read data generation v3 [Internet]. 2022;Available from: <https://www.protocols.io/view/hmw-gdna-purification-and-ont-ultra-long-read-data-b55tq86n>
138. Gong L, Wong C-H, Idol J, Ngan CY, Wei C-L. Ultra-long Read Sequencing for Whole Genomic DNA Analysis. *JoVE (Journal of Visualized Experiments)*. 2019;e58954.
139. Gorkin DU, Qiu Y, Hu M, Fletez-Brant K, Liu T, Schmitt AD, Noor A, Chiou J, Gaulton KJ, Sebat J, Li Y, Hansen KD, Ren B. Common DNA sequence variation influences 3-dimensional conformation of the human genome. *Genome Biol*. 2019;20:255.
140. Köster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics*. 2012;28:2520–2522.
141. Rautiainen M, Marschall T. MBG: Minimizer-based sparse de Bruijn Graph construction. *Bioinformatics*. 2021;37:2476–2478.
142. Rautiainen M, Marschall T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol*. 2020;21:253.
143. Jain C, Rhie A, Zhang H, Chu C, Walenz BP, Koren S, Phillippy AM. Weighted minimizer sampling improves long read mapping. *Bioinformatics*. 2020;36:i111–i118.
144. Jain C, Rhie A, Hansen NF, Koren S, Phillippy AM. Long-read mapping to repetitive reference sequences using Winnowmap2. *Nat Methods*. 2022;19:705–710.
145. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*. 1998;8:175–185.
146. Henglin M, Ghareghani M, Harvey W, Porubsky D, Koren S, Eichler EE, Ebert P, Marschall T. Phasing Diploid Genome Assembly Graphs with Single-Cell Strand Sequencing. *bioRxiv* [Internet]. 2024;Available from: <http://dx.doi.org/10.1101/2024.02.15.580432>
147. Porubsky D, Sanders AD, Taudt A, Colomé-Tatché M, Lansdorp PM, Guryev V. breakpointR: an R/Bioconductor package to localize strand state changes in Strand-seq data. *Bioinformatics*. 2020;36:1260–1261.
148. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *aos*. 2001;29:1165–1188.
149. Nassar LR, Barber GP, Benet-Pagès A, Casper J, Clawson H, Diekhans M, Fischer C,

- Gonzalez JN, Hinrichs AS, Lee BT, Lee CM, Muthuraman P, Nguy B, Pereira T, Nejad P, Perez G, Raney BJ, Schmelter D, Speir ML, Wick BD, Zweig AS, Haussler D, Kuhn RM, Haeussler M, Kent WJ. The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res.* 2023;51:D1188–D1195.
150. Audano PA, Beck CR. Small polymorphisms are a source of ancestral bias in structural variant breakpoint placement. *Genome Res.* 2024;34:7–19.
 151. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods.* 2020;17:261–272.
 152. Zhou W, Emery SB, Flasch DA, Wang Y, Kwan KY, Kidd JM, Moran JV, Mills RE. Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. *Nucleic Acids Res.* 2020;48:1146–1163.
 153. McDonald TL, Zhou W, Castro CP, Mumm C, Switzenberg JA, Mills RE, Boyle AP. Cas9 targeted enrichment of mobile elements using nanopore sequencing. *Nat Commun.* 2021;12:3586.
 154. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics.* 2009;25:1422–1423.
 155. Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res.* 2002;12:656–664.
 156. Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, Mills RE, 1000 Genomes Project Consortium, Devine SE. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res.* 2017;27:1916–1929.
 157. Weber T, Cosenza MR, Korbel J. MosaiCatcher v2: a single-cell structural variations detection and analysis reference framework based on Strand-seq. *Bioinformatics.* 2023;39:btad633.
 158. Ren J, Gu B, Chaisson MJP. vamos: variable-number tandem repeats annotation using efficient motif sets. *Genome Biol.* 2023;24:175.
 159. Gu B, Chaisson MJP. TRCompDB: A reference of human tandem repeat sequence and composition variation from long-read assemblies [Internet]. *Bioinformatics.* 2024; Available from: <https://www.biorxiv.org/content/10.1101/2024.08.07.607105v1>
 160. Poznik GD, Henn BM, Yee M-C, Sliwerska E, Euskirchen GM, Lin AA, Snyder M, Quintana-Murci L, Kidd JM, Underhill PA, Bustamante CD. Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science.* 2013;341:562–565.
 161. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics.* 2011;27:2987–2993.

162. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics*. 2011;27:2156–2158.
163. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*. 2007;7:214.
164. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30:1312–1313.
165. Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PLF, Aximu-Petri A, Prüfer K, de Filippo C, Meyer M, Zwyns N, Salazar-García DC, Kuzmin YV, Keates SG, Kosintsev PA, Razhev DI, Richards MP, Peristov NV, Lachmann M, Douka K, Higham TFG, Slatkin M, Hublin J-J, Reich D, Kelso J, Viola TB, Pääbo S. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*. 2014;514:445–449.
166. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30:2114–2120.
167. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29:15–21.
168. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*. 2012;7:562–578.
169. Frost JM, Monk D, Stojilkovic-Mikic T, Woodfine K, Chitty LS, Murrell A, Stanier P, Moore GE. Evaluation of allelic expression of imprinted genes in adult human blood. *PLoS One*. 2010;5:e13556.
170. Baran Y, Subramaniam M, Biton A, Tukiainen T, Tsang EK, Rivas MA, Pirinen M, Gutierrez-Arcelus M, Smith KS, Kukurba KR, Zhang R, Eng C, Torgerson DG, Urbanek C, GTEx Consortium, Li JB, Rodriguez-Santana JR, Burchard EG, Seibold MA, MacArthur DG, Montgomery SB, Zaitlen NA, Lappalainen T. The landscape of genomic imprinting across diverse adult human tissues. *Genome Res*. 2015;25:927–936.
171. Glenn CC, Driscoll DJ, Yang TP, Nicholls RD. Genomic imprinting: potential function and mechanisms revealed by the Prader-Willi and Angelman syndromes. *Mol Hum Reprod*. 1997;3:321–332.
172. Meguro M, Mitsuya K, Nomura N, Kohda M, Kashiwagi A, Nishigaki R, Yoshioka H, Nakao M, Oishi M, Oshimura M. Large-scale evaluation of imprinting status in the Prader-Willi syndrome region: an imprinted direct repeat cluster resembling small nucleolar RNA genes. *Hum Mol Genet*. 2001;10:383–394.
173. Tilgner H, Grubert F, Sharon D, Snyder MP. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc Natl Acad Sci U S A*. 2014;111:9869–9874.
174. McArthur E, Capra JA. Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. *Am J Hum Genet*. 2021;108:269–283.

175. Rajderkar S, Barozzi I, Zhu Y, Hu R, Zhang Y, Li B, Alcaina Caro A, Fukuda-Yuzawa Y, Kelman G, Akeza A, Blow MJ, Pham Q, Harrington AN, Godoy J, Meky EM, von Maydell K, Hunter RD, Akiyama JA, Novak CS, Plajzer-Frick I, Afzal V, Tran S, Lopez-Rios J, Talkowski ME, Lloyd KCK, Ren B, Dickel DE, Visel A, Pennacchio LA. Topologically associating domain boundaries are required for normal genome function. *Commun Biol.* 2023;6:435.
176. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* 2016;3:95–98.
177. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25:1754–1760.
178. Li C, Bonder MJ, Syed S, Human Genome Structural Variation Consortium (HGSVC), HGSVC Functional Analysis Working Group, Zody MC, Chaisson MJP, Talkowski ME, Marshall T, Korb J, Eichler EE, Lee C, Shi X. A comprehensive catalog of 3D genome organization in diverse human genomes facilitates understanding of the impact of structural variation on chromatin structure [Internet]. bioRxiv. 2023 [cited 2024 Aug 29];2023.05.15.540856. Available from: <https://www.biorxiv.org/content/10.1101/2023.05.15.540856v1>
179. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst.* 2016;3:99–101.
180. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn A, Machol I, Omer AD, Lander ES, Aiden EL. A three-dimensional map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell.* 2014;159:1665–1680.
181. Kruse K, Hug CB, Vaquerizas JM. FAN-C: a feature-rich framework for the analysis and visualisation of chromosome conformation capture data. *Genome Biol.* 2020;21:303.
182. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science.* 2009;326:289–293.
183. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature.* 2012;485:376–380.
184. Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, Uzawa S, Dekker J, Meyer BJ. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature.* 2015;523:240–244.
185. Wu X, Lu Y, Ding Q, You G, Dai J, Xi X, Wang H, Wang X. Characterisation of large F9 deletions in seven unrelated patients with severe haemophilia B. *Thromb Haemost.* 2014;112:459–465.

186. Wallace MR, Andersen LB, Saulino AM, Gregory PE, Glover TW, Collins FS. A de novo Alu insertion results in neurofibromatosis type 1. *Nature*. 1991;353:864–866.
187. Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L, GTEx Consortium, Montgomery SB, Battle A, Conrad DF, Hall IM. The impact of structural variation on human gene expression. *Nat Genet*. 2017;49:692–699.
188. Hernández HG, Hernández-Castañeda AA, Pieschacón MP, Arboleda H. ZNF718, HOXA4, and ZFP57 are differentially methylated in periodontitis in comparison with periodontal health: Epigenome-wide DNA methylation pilot study. *J Periodontal Res*. 2021;56:710–725.
189. Korbelt JO, Kim PM, Chen X, Urban AE, Weissman S, Snyder M, Gerstein MB. The current excitement about copy-number variation: how it relates to gene duplications and protein families. *Curr Opin Struct Biol*. 2008;18:366–374.
190. Kichaev G, Bhatia G, Loh P-R, Gazal S, Burch K, Freund MK, Schoech A, Pasaniuc B, Price AL. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am J Hum Genet*. 2019;104:65–75.
191. Morgan MD, Pairo-Castineira E, Rawlik K, Canela-Xandri O, Rees J, Sims D, Tenesa A, Jackson IJ. Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability. *Nat Commun*. 2018;9:5271.
192. van der Meer D, Kaufmann T, Shadrin AA, Makowski C, Frei O, Roelfs D, Monereo-Sánchez J, Linden DEJ, Rokicki J, Alnæs D, de Leeuw C, Thompson WK, Loughnan R, Fan CC, Westlye LT, Andreassen OA, Dale AM. The genetic architecture of human cortical folding. *Sci Adv*. 2021;7:eabj9446.
193. Sakaue S, Kanai M, Tanigawa Y, Karjalainen J, Kurki M, Koshiba S, Narita A, Konuma T, Yamamoto K, Akiyama M, Ishigaki K, Suzuki A, Suzuki K, Obara W, Yamaji K, Takahashi K, Asai S, Takahashi Y, Suzuki T, Shinozaki N, Yamaguchi H, Minami S, Murayama S, Yoshimori K, Nagayama S, Obata D, Higashiyama M, Masumoto A, Koretsune Y, FinnGen, Ito K, Terao C, Yamauchi T, Komuro I, Kadowaki T, Tamiya G, Yamamoto M, Nakamura Y, Kubo M, Murakami Y, Yamamoto K, Kamatani Y, Palotie A, Rivas MA, Daly MJ, Matsuda K, Okada Y. A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet*. 2021;53:1415–1424.
194. Massey J, Plant D, Hyrich K, Morgan AW, Wilson AG, Spiliopoulou A, Colombo M, McKeigue P, Isaacs J, Cordell H, Pitzalis C, Barton A, BRAGGSS, MATURA Consortium. Genome-wide association study of response to tumour necrosis factor inhibitor therapy in rheumatoid arthritis. *Pharmacogenomics J*. 2018;18:657–664.
195. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*. 2004;32:D493–6.
196. English AC, Menon VK, Gibbs RA, Metcalf GA, Sedlazeck FJ. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol*. 2022;23:271.
197. Martin M, Patterson M, Garg S, Fischer SO, Pisanti N, Klau GW, Schöenhuth A, Marschall T. WhatsHap: fast and accurate read-based phasing [Internet]. bioRxiv. 2016 [cited 2024 Aug 29];085050. Available from:

198. Marco-Sola S, Eizenga JM, Guarracino A, Paten B, Garrison E, Moreto M. Optimal gap-affine alignment in O(s) space. *Bioinformatics* [Internet]. 2023;39. Available from: <http://dx.doi.org/10.1093/bioinformatics/btad074>
199. de Bakker PIW, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, Ke X, Monsuur AJ, Whittaker P, Delgado M, Morrison J, Richardson A, Walsh EC, Gao X, Galver L, Hart J, Hafler DA, Pericak-Vance M, Todd JA, Daly MJ, Trowsdale J, Wijmenga C, Vyse TJ, Beck S, Murray SS, Carrington M, Gregory S, Deloukas P, Rioux JD. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat Genet*. 2006;38:1166–1172.
200. Dillthey A, Cox C, Iqbal Z, Nelson MR, McVean G. Improved genome inference in the MHC using a population reference graph. *Nat Genet*. 2015;47:682–688.
201. Moutsianas L, Jostins L, Beecham AH, Dillthey AT, Xifara DK, Ban M, Shah TS, Patsopoulos NA, Alfredsson L, Anderson CA, Attfield KE, Baranzini SE, Barrett J, Binder TMC, Booth D, Buck D, Celius EG, Cotsapas C, D'Alfonso S, Dendrou CA, Donnelly P, Dubois B, Fontaine B, Fugger L, Goris A, Gourraud P-A, Graetz C, Hemmer B, Hillert J, International IBD Genetics Consortium (IIBDGC), Kockum I, Leslie S, Lill CM, Martinelli-Boneschi F, Oksenberg JR, Olsson T, Oturai A, Saarela J, Söndergaard HB, Spurkland A, Taylor B, Winkelmann J, Zipp F, Haines JL, Pericak-Vance MA, Spencer CCA, Stewart G, Hafler DA, Ivinson AJ, Harbo HF, Hauser SL, De Jager PL, Compston A, McCauley JL, Sawcer S, McVean G. Class II HLA interactions modulate genetic risk for multiple sclerosis. *Nat Genet*. 2015;47:1107–1113.
202. Dillthey AT. State-of-the-art genome inference in the human MHC. *Int J Biochem Cell Biol*. 2021;131:105882.
203. Zhou Y, Song L, Li H. Full resolution HLA and KIR gene annotations for human genome assemblies. *Genome Res* [Internet]. 2024;Available from: <http://dx.doi.org/10.1101/gr.278985.124>
204. Alexandrov N, Wang T, Blair L, Nadon B, Sayer D. HLA-OLI: A new MHC class I pseudogene and HLA-Y are located on a 60 kb indel in the human MHC between HLA-W and HLA-J. *Hladnikia* . 2023;102:599–606.
205. Klussmeier A, Putke K, Klasberg S, Kohler M, Sauter J, Schefzyk D, Schöfl G, Massalski C, Schäfer G, Schmidt AH, Roers A, Lange V. High population frequencies of MICA copy number variations originate from independent recombination events. *Front Immunol*. 2023;14:1297589.
206. Carapito R, Aouadi I, Verniquet M, Untrau M, Pichot A, Beaudrey T, Bassand X, Meyer S, Faucher L, Posson J, Morlon A, Kotova I, Delbos F, Walencik A, Aarnink A, Kennel A, Suberbielle C, Taupin J-L, Matern BM, Spierings E, Congy-Jolivet N, Essaydi A, Perrin P, Blancher A, Charron D, Cereb N, Maumy-Bertrand M, Bertrand F, Garrigue V, Pernin V, Weekers L, Naesens M, Kamar N, Legendre C, Glotz D, Caillard S, Ladrière M, Giral M, Anglicheau D, Süsal C, Bahram S. The MHC class I MICA gene is a histocompatibility antigen in kidney transplantation. *Nat Med*. 2022;28:989–998.
207. Gongora R, Figueroa F, Klein J. The HLA-DRB9 gene and the origin of HLA-DR

- haplotypes. *Hum Immunol*. 1996;51:23–31.
208. Svensson AC, Setterblad N, Pihlgren U, Rask L, Andersson G. Evolutionary relationship between human major histocompatibility complex HLA-DR haplotypes. *Immunogenetics*. 1996;43:304–314.
 209. Gongora R, Figueroa F, O’Huigin C, Klein J. HLA-DRB9--possible remnant of an ancient functional DRB subregion. *Scand J Immunol*. 1997;45:504–510.
 210. Yang Y, Chung EK, Wu YL, Savelli SL, Nagaraja HN, Zhou B, Hebert M, Jones KN, Shu Y, Kitzmiller K, Blanchong CA, McBride KL, Higgins GC, Rennebohm RM, Rice RR, Hackshaw KV, Roubey RAS, Grossman JM, Tsao BP, Birmingham DJ, Rovin BH, Hebert LA, Yu CY. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *Am J Hum Genet*. 2007;80:1037–1054.
 211. Blanchong CA, Zhou B, Rupert KL, Chung EK, Jones KN, Sotos JF, Zipf WB, Rennebohm RM, Yung Yu C. Deficiencies of human complement component C4A and C4B and heterozygosity in length variants of RP-C4-CYP21-TNX (RCCX) modules in caucasians. The load of RCCX genetic diversity on major histocompatibility complex-associated disease. *J Exp Med*. 2000;191:2183–2196.
 212. Belt KT, Yu CY, Carroll MC, Porter RR. Polymorphism of human complement component C4. *Immunogenetics*. 1985;21:173–180.
 213. Yu CY, Belt KT, Giles CM, Campbell RD, Porter RR. Structural basis of the polymorphism of human complement components C4A and C4B: gene size, reactivity and antigenicity. *EMBO J*. 1986;5:2873–2881.
 214. Carrozza C, Foca L, De Paolis E, Concolino P. Genes and Pseudogenes: Complexity of the RCCX Locus and Disease. *Front Endocrinol* . 2021;12:709758.
 215. Bánlaki Z, Doleschall M, Rajczy K, Fust G, Szilágyi A. Fine-tuned characterization of RCCX copy number variants and their relationship with extended MHC haplotypes. *Genes Immun*. 2012;13:530–535.
 216. Marin WM, Augusto DG, Wade KJ, Hollenbach JA. High-throughput complement component 4 genomic sequence analysis with C4Investigator. *HLA*. 2024;103:e15273.
 217. Sekar A, Bialas AR, de Rivera H, Davis A, Hammond TR, Kamitaki N, Tooley K, Presumey J, Baum M, Van Doren V, Genovese G, Rose SA, Handsaker RE, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Daly MJ, Carroll MC, Stevens B, McCarroll SA. Schizophrenia risk from complex variation of complement component 4. *Nature*. 2016;530:177–183.
 218. Yang Z, Mendoza AR, Welch TR, Zipf WB, Yu CY. Modular variations of the human major histocompatibility complex class III genes for serine/threonine kinase RP, complement component C4, steroid 21-hydroxylase CYP21, and tenascin TNX (the RCCX module). A mechanism for gene deletions and disease associations. *J Biol Chem*. 1999;274:12147–12156.

219. Koppens PFJ, Smeets HJM, de Wijs IJ, Degenhart HJ. Mapping of a de novo unequal crossover causing a deletion of the steroid 21-hydroxylase (CYP21A2) gene and a non-functional hybrid tenascin-X (TNXB) gene. *J Med Genet.* 2003;40:e53.
220. Yang Y, Chung EK, Zhou B, Lhotta K, Hebert LA, Birmingham DJ, Rovin BH, Yu CY. The intricate role of complement component C4 in human systemic lupus erythematosus. *Curr Dir Autoimmun.* 2004;7:98–132.
221. Mc Cartney AM, Shafin K, Alonge M, Bzikadze AV, Formenti G, Fungtammasan A, Howe K, Jain C, Koren S, Logsdon GA, Miga KH, Mikheenko A, Paten B, Shumate A, Soto DC, Sović I, Wood JMD, Zook JM, Phillippy AM, Rhie A. Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nat Methods.* 2022;19:687–695.
222. Ruiz JL, Reimering S, Escobar-Prieto JD, Brancucci NMB, Echeverry DF, Abdi AI, Marti M, Gómez-Díaz E, Otto TD. From contigs towards chromosomes: automatic improvement of long read assemblies (ILRA). *Brief Bioinform* [Internet]. 2023;24. Available from: <http://dx.doi.org/10.1093/bib/bbad248>
223. Zhang H, Jain C, Aluru S. A comprehensive evaluation of long read error correction methods. *BMC Genomics.* 2020;21:889.
224. Mahmoud M, Huang Y, Garimella K, Audano PA, Wan W, Prasad N, Handsaker RE, Hall S, Pionzio A, Schatz MC, Talkowski ME, Eichler EE, Levy SE, Sedlazeck FJ. Utility of long-read sequencing for All of Us. *Nat Commun.* 2024;15:837.
225. Jiang Z, Hubley R, Smit A, Eichler EE. DupMasker: a tool for annotating primate segmental duplications. *Genome Res.* 2008;18:1362–1368.
226. Wright E. Using DECIPHER v2.0 to analyze big biological sequence data in R. *R J.* 2016;8:352.
227. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics.* 2019;35:526–528.
228. Schliep KP. phangorn: phylogenetic analysis in R. *Bioinformatics.* 2011;27:592–593.
229. Yilmaz F, Gurusamy U, Mosley TJ, Hallast P, Kim K, Mostovoy Y, Purcell RH, Shaikh TH, Zwick ME, Kwok P-Y, Lee C, Mülle JG. High level of complexity and global diversity of the 3q29 locus revealed by optical mapping and long-read sequencing. *Genome Med.* 2023;15:35.
230. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215:403–410.
231. Chao A, Gotelli NJ, Hsieh TC, Sander EL, Ma KH, Colwell RK, Ellison AM. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs.* 2014;84:45–67.
232. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. Recent segmental duplications in the human genome. *Science.* 2002;297:1003–1007.

233. Inoue K, Lupski JR. Molecular mechanisms for genomic disorders. *Annu Rev Genomics Hum Genet.* 2002;3:199–242.
234. Zody MC, Jiang Z, Fung H-C, Antonacci F, Hillier LW, Cardone MF, Graves TA, Kidd JM, Cheng Z, Abouelleil A, Chen L, Wallis J, Glasscock J, Wilson RK, Reily AD, Duckworth J, Ventura M, Hardy J, Warren WC, Eichler EE. Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat Genet.* 2008;40:1076–1083.
235. Vollger MR, Dishuck PC, Harvey WT, DeWitt WS, Guitart X, Goldberg ME, Rozanski AN, Lucas J, Asri M, Human Pangenome Reference Consortium, Munson KM, Lewis AP, Hoekzema K, Logsdon GA, Porubsky D, Paten B, Harris K, Hsieh P, Eichler EE. Increased mutation and gene conversion within human segmental duplications. *Nature.* 2023;617:325–334.
236. Schmutz J, Martin J, Terry A, Couronne O, Grimwood J, Lowry S, Gordon LA, Scott D, Xie G, Huang W, Hellsten U, Tran-Gyamfi M, She X, Prabhakar S, Aerts A, Altherr M, Bajorek E, Black S, Branscomb E, Caoile C, Challacombe JF, Chan YM, Denys M, Detter JC, Escobar J, Flowers D, Fotopulos D, Glavina T, Gomez M, Gonzales E, Goodstein D, Grigoriev I, Groza M, Hammon N, Hawkins T, Haydu L, Israni S, Jett J, Kadner K, Kimball H, Kobayashi A, Lopez F, Lou Y, Martinez D, Medina C, Morgan J, Nandkeshwar R, Noonan JP, Pitluc S, Pollard M, Predki P, Priest J, Ramirez L, Retterer J, Rodriguez A, Rogers S, Salamov A, Salazar A, Thayer N, Tice H, Tsai M, Ustaszewska A, Vo N, Wheeler J, Wu K, Yang J, Dickson M, Cheng J-F, Eichler EE, Olsen A, Pennacchio LA, Rokhsar DS, Richardson P, Lucas SM, Myers RM, Rubin EM. The DNA sequence and comparative analysis of human chromosome 5. *Nature.* 2004;431:268–274.
237. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics.* 2004;5:113.
238. Hagberg AA, Schult DA, Swart PJ. Exploring Network Structure, Dynamics, and Function using NetworkX. *scipy* [Internet]. 2008 [cited 2024 Dec 20];Available from: <https://proceedings.scipy.org/articles/TCWV9851>