Article

Complex genetic variation in nearly complete human genomes

https://doi.org/10.1038/s41586-025-09140-6

Received: 23 September 2024

Accepted: 12 May 2025

Published online: 23 July 2025

Open access

Check for updates

Glennis A. Logsdon[1,2,42], Peter Ebert[3,4,42], Peter A. Audano[5,42], Mark Loftus[6,7,41,42], David Porubsky[1], Jana Ebler[4,8], Feyza Yilmaz[5], Pille Hallast[5], Timofey Prodanov[4,8], DongAhn Yoo[1], Carolyn A. Paisie[5], William T. Harvey[1], Xuefang Zhao[9,10,11], Gianni V. Martino[6,7,12], Mir Henglin[4,8], Katherine M. Munson[1], Keon Rabbani[13], Chen-Shan Chin[14], Bida Gu[13], Hufsah Ashraf[4,8], Stephan Scholz[4,15], Olanrewaju Austine-Orimoloye[16], Parithi Balachandran[5], Marc Jan Bonder[17,18,19], Haoyu Cheng[20], Zechen Chong[21], Jonathan Crabtree[22], Mark Gerstein[23,24], Lisbeth A. Guethlein[25], Patrick Hasenfeld[26], Glenn Hickey[27], Kendra Hoekzema[1], Sarah E. Hunt[16], Matthew Jensen[23,24], Yunzhe Jiang[23,24], Sergey Koren[28], Youngjun Kwon[1], Chong Li[29,30], Heng Li[31,32], Jiaqi Li[23,24], Paul J. Norman[33,34], Keisuke K. Oshima[2], Benedict Paten[27], Adam M. Phillippy[28], Nicholas R. Pollock[33], Tobias Rausch[26], Mikko Rautiainen[35], Yuwei Song[21], Arda Söylev[4,8], Arvis Sulovari[1], Likhitha Surapaneni[16], Vasiliki Tsapalou[26], Weichen Zhou[36], Ying Zhou[31], Qihui Zhu[5,37], Michael C. Zody[38], Ryan E. Mills[36], Scott E. Devine[22], Xinghua Shi[29,30], Michael E. Talkowski[9,10,11], Mark J. P. Chaisson[13], Alexander T. Dilthey[4,15], Miriam K. Konkel[6,7]✉, Jan O. Korbel[26]✉, Charles Lee[5]✉, Christine R. Beck[5,39]✉, Evan E. Eichler[1,40]✉ & Tobias Marschall[4,8]✉

Diverse sets of complete human genomes are required to construct a pangenome reference and to understand the extent of complex structural variation. Here we sequence 65 diverse human genomes and build 130 haplotype-resolved assemblies (median continuity of 130 Mb), closing 92% of all previous assembly gaps[1,2] and reaching telomere-to-telomere status for 39% of the chromosomes. We highlight complete sequence continuity of complex loci, including the major histocompatibility complex (MHC), *SMN1*/*SMN2*, *NBPF8* and *AMY1*/*AMY2*, and fully resolve 1,852 complex structural variants. In addition, we completely assemble and validate 1,246 human centromeres. We find up to 30-fold variation in α-satellite higher-order repeat array length and characterize the pattern of mobile element insertions into α-satellite higher-order repeat arrays. Although most centromeres predict a single site of kinetochore attachment, epigenetic analysis suggests the presence of two hypomethylated regions for 7% of centromeres. Combining our data with the draft pangenome reference[1] significantly enhances genotyping accuracy from short-read data, enabling whole-genome inference[3] to a median quality value of 45. Using this approach, 26,115 structural variants per individual are detected, substantially increasing the number of structural variants now amenable to downstream disease association studies.

Long-read sequencing (LRS) technologies were critical to the completion of the first human genome[4]. LRS technologies significantly increase the sensitivity to detect structural variants (SVs), defined as variants 50 bp in length or longer, and coupling LRS data with Hi-C[5], single-cell template strand sequencing (Strand-seq)[6] or trio data[7] provided the necessary short-range and long-range phasing data to assemble both haplotypes. The high sequence quality and contiguity of such diploid genome assemblies have made the first draft human pangenome reference possible[1].

Despite these advances, gaps remain, especially at genetically complex loci[2]. For example, in our previous assembly of 32 human genomes as part of the Human Genome Structural Variation Consortium (HGSVC)[8], we found that most centromeres and more than half of the large, highly identical segmental duplications (SDs) were incomplete, resulting in missing protein-coding genes[2]. Closing these gaps in the first complete human genome[4] required combining the complementary strengths of PacBio high-fidelity (HiFi) reads (approximately 18 kb in length and high base-level accuracy) and ultra-long Oxford Nanopore Technologies (ONT) reads (more than 100 kb in length but with lower base-level accuracy). Computational tools such as Verkko[9] and hifiasm (ultra-long)[10] have automated this process. Here we present new resources and results from the HGSVC (Supplementary Fig. 1), targeting a diverse set of 65 humans predominantly from the 1000 Genomes Project (1kGP) cohort[11] with the goal of producing a genetically diverse sampling of nearly gapless chromosomes, including the centromeres and complex SDs.

A list of affiliations appears at the end of the paper.

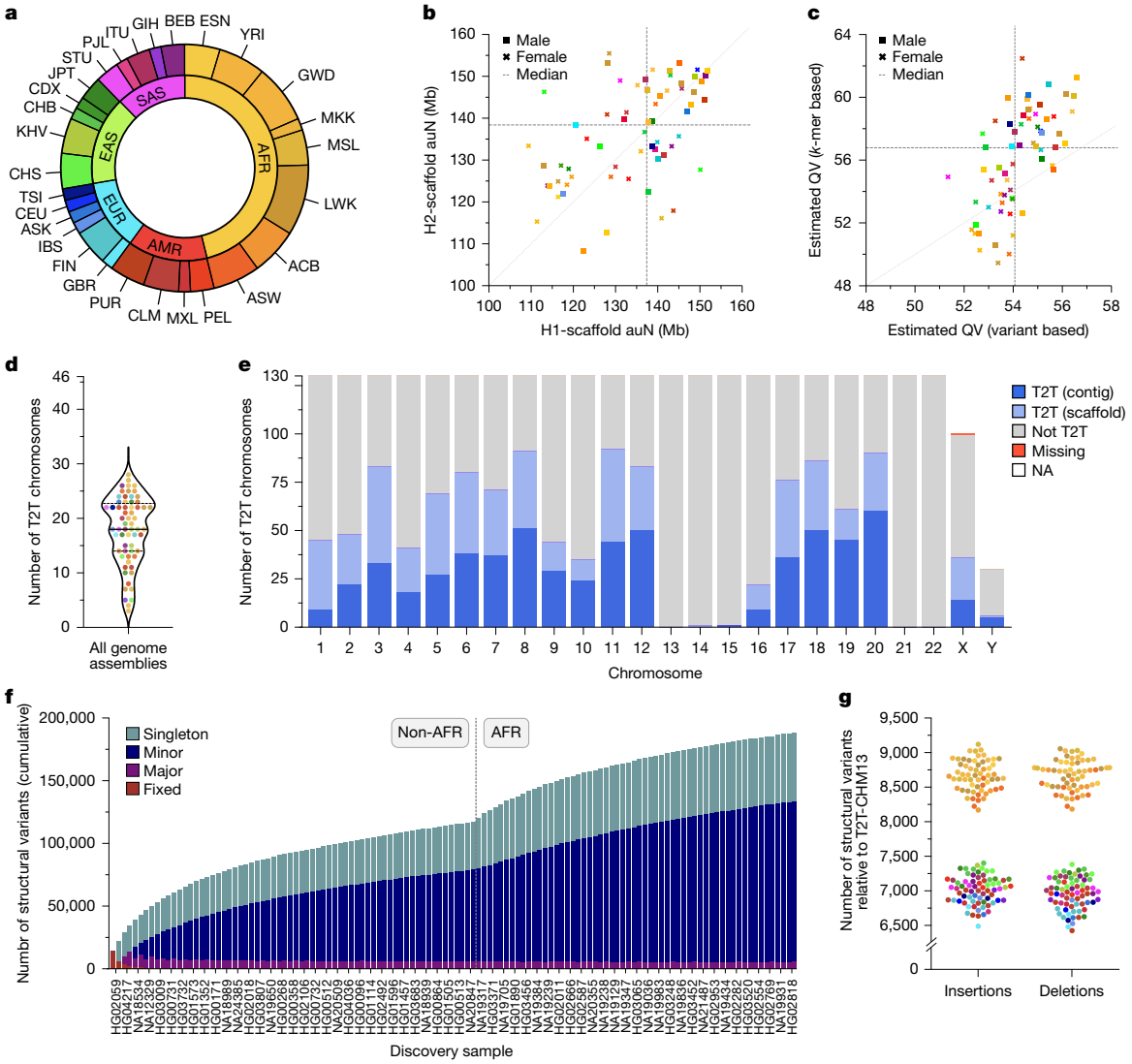


Fig. 1 | LRS, assembly and variant calling of 65 diverse humans. a, Continental group (inner ring) and population group (outer ring) of the 65 diverse humans analysed in this study. AFR, African; AMR, American; EAS, East Asian; EUR, European; SAS, South Asian. Population groups are labelled according to the 1000 Genomes Project[11], along with the added Maasai in Kinyawa, Kenya (MKK) and Ashkenazim (ASK) labels. **b**, Scaffold auN for haplotype 1 (H1) and haplotype 2 (H2) contigs from each genome assembly. Data points are coloured by population group. The dashed lines indicate the median auN per haplotype. The dotted line indicates the unit diagonal. **c**, Quality value (QV) estimates for each genome assembly derived from variant calls or *k*-mer statistics (Methods). **d**, The number of chromosomes assembled from T2T for each genome assembly, including both single contigs and scaffolds (Methods). The median (solid line) and first and third quartiles (dotted lines) are shown. **e**, The number of T2T chromosomes in a single contig (dark blue, T2T contig) or in a single scaffold (light blue, T2T scaffold). Incomplete chromosomes are labelled as 'not T2T' or 'missing' if missing entirely. Sex chromosomes not present in the respective haploid assembly are labelled as 'NA'. **f**, Cumulative non-redundant SVs across the diverse haplotypes in this study called with respect to the T2T-CHM13 reference genome (three trio children excluded). **g**, Number of SVs detected for each haplotype relative to the T2T-CHM13 reference genome, coloured by population. Insertions and deletions are balanced when called against the T2T-CHM13 reference genome but imbalanced when called against the GRCh38 reference genome (Extended Data Fig. 1d).

Production of 130 haplotype assemblies

Data production

We selected 65 human lymphoblastoid cell lines representing individuals spanning five continental groups and 28 population groups for sequencing (Fig. 1a and Supplementary Table 1). We generated approximately 47-fold coverage of PacBio HiFi and approximately 56-fold coverage of ONT (approximately 36-fold ultra-long) long reads on average per individual (Extended Data Fig. 1a,b and Supplementary Table 2; see Methods). In addition, we performed Strand-seq (Supplementary Table 2), Bionano Genomics optical mapping (Supplementary Table 3), Hi-C sequencing (Supplementary Tables 4 and 5), isoform sequencing (Iso-Seq; Supplementary Table 6) and RNA sequencing (RNA-seq; Supplementary Table 7).

Assembly

We generated haplotype-resolved assemblies from all 65 diploid individuals using Verkko[9] (Fig. 1a and Supplementary Tables 1 and 2; see Methods). The phasing signal was produced with Graphasing[12], leveraging Strand-seq to globally phase assembly graphs at a quality on par with trio-based workflows[12] (Methods). This approach enabled us to cover all 26 populations from the 1kGP by including individuals that are not part of a family trio. The resulting set of 130 haploid assemblies is highly contiguous (median area under the Nx curve (auN) of 137 Mb; Fig. 1b and Supplementary Table 8) and accurate at the base-pair level (median quality value between 54 and 57; Fig. 1c and Supplementary Table 9; see Methods). We estimated the assemblies to be 99% complete (median) for known single-copy genes (Extended Data Fig. 1c

and Supplementary Table 10) and to close 92% of previously reported gaps in PacBio HiFi-only assemblies[2] (Supplementary Figs. 2 and 3 and Supplementary Table 11; see Methods).

We integrated a range of quality control annotations for each assembly using established tools such as Flagger, NucFreq, Merqury and Inspector (Supplementary Tables 12 and 13 and Figs. 4 and 5) to compute robust error estimates for each assembled base (Supplementary Tables 14–17; see Methods). We estimated that 99.6% of the phased sequence (median) has been assembled correctly (Supplementary Table 18). For the three family trios in our dataset (SH032, Y117 and PR05 (ref. 11)), we assessed the parental support for the respective haplotypes in the child's assembly via assembly-to-assembly alignments and found that a median of 99.9% of all sequence assembled in contigs of more than 100 kb are supported by one parent assembly (Supplementary Table 19; see Methods). In total, Verkko assembled 602 chromosomes as a single gapless contig from telomere to telomere (T2T; median of 10 per genome) and an additional 559 as a single scaffold (median of 8 per genome), that is, in a connected sequence containing one or more N-gaps (Fig. 1d,e, Supplementary Table 20 and Supplementary Fig. 6; see Methods).

Certain regions, such as centromeres or the Yq12 region, remained challenging to assemble and evaluate. We therefore complemented our assembly efforts by running hifiasm (ultra-long)[10] on the same dataset (Supplementary Tables 21–23 and Supplementary Figs. 7 and 8; see Methods), but restricted the use of the resulting assemblies to extending our analysis set for centromeres and the Yq12 region after manual curation of the relevant sequences.

Variant calling

From our phased assemblies, we identified 188,500 SVs, 6.3 million indels and 23.9 million single-nucleotide variants (SNVs) against the T2T-CHM13v.2.0 (T2T-CHM13) reference (Fig. 1f). Against GRCh38-NoALT (GRCh38), we identified 176,531 SVs, 6.2 million indels and 23.5 million SNVs (Supplementary Table 24; see Data availability). Callsets for both references were led by PAV (v.2.4.0.1)[8] with orthogonal support from 10 other independent callers with sensitivity for SVs, indels and SNVs (Supplementary Table 25; see Methods). We found higher support for PAV calls across all callers (99.7%) than other methods (99.7% to 67.9%; Extended Data Fig. 1d and Supplementary Fig. 9), with one exception for SVIM-asm, when run using the alignment parameters for PAV (99.70% SVIM-asm versus 99.66% PAV; Supplementary Table 26). With our current assemblies and this approach, we increased the size of the SV callset by 59% and reduced false discovery by 55% on average compared with previous callsets[8] (Supplementary Tables 27 and 28 and Supplementary Methods). With one additional individual, we estimated that our callset would increase by 842 SV insertions and deletions with a 1.86× enrichment for an African versus a non-African individual (1,117 versus 599; Supplementary Methods).

Per assembled haplotype, we identified 7,772 SV insertions (12,903 per genome) and 7,745 SV deletions (12,505 per genome) on average in the T2T-CHM13 reference (Fig. 1g). As expected, GRCh38 SVs are unbalanced[8,13] with 11,275 SV insertions per haplotype (17,458 per genome) and 6,972 SV deletions per haplotype (10,868 per genome) on average (Extended Data Fig. 1e and Supplementary Tables 29 and 30), with excess insertions occurring in high-allele-frequency variants, which can be largely explained by reference errors[14]. As expected, a distinct peak for fixed SVs (100% allele frequency) is apparent for GRCh38 SV insertions composed of variants in GRCh38 with no representation in T2T-CHM13 (Extended Data Fig. 1f).

An improved genomic resource

Mobile element insertions

Mobile element insertions (MEIs)[15] constitute 8.2% of all SVs (relative to T2T-CHM13). We identified 12,919 putative MEIs from the 130 haplotype assemblies (Supplementary Table 31 and Supplementary Fig. 10; see Methods; for the GRCh38 union callset, see Supplementary Table 32 and Supplementary Fig. 11). Comparison with an orthogonal MEI callset showed a high concordance of 92.1% (Supplementary Tables 33 and 34; see Methods). Of note, we found 559 full-length L1 insertions (L1HS and L1PA2), with 96.1% possessing at least one intact open reading frame (ORF) and 82.3% harbouring two intact ORFs. Therefore, the vast majority of full-length L1 MEIs appear to retain the potential to retrotranspose. Compared with our previous study[8] (n = 9,453 MEIs; 7,738 for *Alu*, 1,775 for L1 and 540 for SINE-VNTR-Alu (SVA)), the total number of MEIs increased by 36.65% primarily due to an increase in individuals of African descent (Supplementary Fig. 10d). Finally, we screened the PAV deletion callset and identified 2,450 polymorphic MEIs present in T2T-CHM13 (Supplementary Tables 35 and 36 and Supplementary Fig. 12).

Inversions

Identifying inversions is challenging due to the frequent location of their boundaries in long, highly identical repeat sequences. We identified 276 T2T-CHM13-based and 298 GRCh38-based inversions in the main callset and performed quality control by re-genotyping these calls using ArbiGent on Strand-seq data[16] (Supplementary Tables 37 and 38 and Supplementary Methods) as well as manual inspection (Supplementary Table 37, Supplementary Figs. 13 and 14 and Supplementary Methods). Of note, we found 21 novel inversions in the PAV callset, of which 18 were detected among 24 new individuals added in the current study. These include a large (1.8 Mb) inversion at chromosome 5q35 that overlaps with the Sotos syndrome critical region[17].

Segmental duplications

SDs are defined independently for each haplotype as segments occurring more than once with more than 1 kb in length and more than 90% identity. Owing to their propensity to undergo non-allelic homologous recombination, they are enriched tenfold for copy number variation and are the source of some of the most complex forms of genic structural polymorphism in the human genome[18,19]. Overall, we found an average of 168.1 Mb (s.d. of 9.2 Mb) of SDs per human genome and observed an improved representation of interchromosomal SDs (Supplementary Figs. 15 and 16) when compared with the Human Pangenome Reference Consortium (HPRC) release[1]. Using T2T-CHM13 as a gauge of completeness (193.7 Mb), we estimated that 25.6 Mb of SDs still remain unresolved per haploid genome (Extended Data Fig. 2a). Most of these unresolved SDs (21.2 Mb) correspond to the acrocentric short arms of chromosomes 13, 14, 15, 21 and 22 (refs. 4,20). We found that 80–90% of SDs are accurately assembled depending on the genome (Supplementary Figs. 17 and 18; see Methods).

When analysing SDs outside of acrocentric regions and where the copy number was supported by fastCN (Supplementary Fig. 19; see Methods), we classified at least 92.8 Mb of the SDs as shared among most humans (present in at least 90% of individuals) and 61.0 Mb as variable across the human population (Extended Data Fig. 2b). In addition, we identified 33 Mb of the SD sequence present in a single copy or not annotated as SDs in T2T-CHM13 (Extended Data Fig. 2c,d). The majority of these (23.6 Mb, including 2.4 Mb of X chromosome SDs) are novel when compared with a recent analysis of 170 human genomes[21] and completely or partially overlap with 167 protein-coding genes (Supplementary Fig. 20). Of note, 31 loci (0.4 Mb) are shared among most humans but not classified as duplicated in the T2T-CHM13 human genome, suggesting that this unique status in the reference represents the minor allele in the human population, a cell line artefact or, less likely, an error in the assembly. Examining genomes by continental group, both the absolute SD content[21] (Supplementary Figs. 21 and 22) and the number of new SDs added per genome is highest for African individuals (3.97 Mb per individual) when compared with genomes of non-African individuals (2.88 Mb per individual).

Genomes with African ancestry have, on average, 468 additional paralogous genes (n = 21,595 total genes) when compared with genomes of non-African individuals (n = 21,127 total genes; Methods). We identified a total of 727 multi-copy genes that have SDs spanning at least 90% of the gene body, with a large proportion corresponding to shared (n = 335 or 46.1%) and variable (n = 292 or 40.2%) SDs (Supplementary Table 39). Comparing the copy numbers to the HPRC assemblies[1], we discovered a similar distribution of genes (Supplementary Fig. 23). Among copy number polymorphic genes, we identified 16 gene families in which the distribution significantly differs between the HPRC and our data (Supplementary Fig. 23; adjusted $P < 0.05$, two-sided Welch's t-test); however, the contiguity for copy number variant genes was considerably greater in our assemblies versus HPRC; 5.88% of duplicated genes in our assemblies are within 200 kb of a contig break or unknown base ('N') compared with 13.95% of duplicated genes in HPRC assemblies (Supplementary Fig. 24).

Y chromosome variation

The Y chromosome remains among the most challenging of human chromosomes to fully assemble due to its highly repetitive sequence composition[20] (Fig. 2a). Our resource provides highly contiguous Y assemblies for 30 male individuals. Seven of these (23%) assembled without breaks across the male-specific Y region (excluding the pseudoautosomal regions, six assembled as T2T scaffolds and one that has a break in the pseudoautosomal region 1; Supplementary Figs. 25 and 26). Of these seven, four are novel fully assembled human Y chromosomes representing E1b1a, R2a and R1b1a Y lineages prevalent in populations of African, Asian and European descent[22–24] (Supplementary Fig. 27).

Our assemblies enable the investigation of the largest heterochromatic region in the human genome, Yq12, mostly composed of highly similar (but size variable) alternating arrays of *DYZ1* (*HSat3A6*, approximately 3.5-kb unit size) and *DYZ2* (*HSat1B*, approximately 2.4-kb unit size) repeats (Fig. 2a). The Yq12 regions across 16 individuals (9 novel and 7 previously published) range from 17.85 to 37.39 Mb (mean of 27.25 Mb, median of 25.62 Mb), with high levels of variation in the number (34–86 arrays; mean of 60, median of 58) and length of *DYZ1* (24.4 kb to 3.59 Mb; mean of 525.7 kb, median of 455.0 kb) and *DYZ2* (11.2 kb to 2.20 Mb; mean of 358.0 kb, median of 273.3 kb) repeat arrays[23,24] (Supplementary Table 40 and Supplementary Fig. 28). Investigating the dynamics of Yq12 remains challenging[25]; however, using the duplication and deletion patterns of four unique *Alu* insertions, we can examine this genomic region over time (Fig. 2a and Supplementary Fig. 28). For example, in NA19239, the presence of two unique *Alu*Y retrotransposon insertions allows clear visualization of a tandem duplication in the region.

Functional effects of SVs

To identify SVs disrupting protein-coding genes under selective constraint[26], we intersected all 176,531 GRCh38-based SVs with coding exons from GENCODE v.45. We found 1,535 SVs, including 938 deletions, 80 inversions, 504 insertions and 13 MEIs, that disrupt 985 unique genes (Supplementary Table 41). A mean of 368 genes per genome have an SV breakpoint altering the coding sequence. On average, only 11.7 genes (3.2%) were disrupted by a singleton variant unique to that individual, whereas 96.8% of genes were disrupted by polymorphic SVs, and 27.8% were disrupted by major-allele SVs (more than 50% allele frequency). Of the 1,535 genes affected by SVs, only 37 were predicted to be intolerant to loss of function in humans (loss-of-function observed/expected upper bound fraction (LOEUF) < 0.35)[27]. Polymorphic SVs altered 16 constrained genes, suggesting that the SVs did not result in loss of function. Indeed, we found that tandem repeat unit variants in coding sequences of four constrained genes were in frame (*MUC5B*, *ACAN*, *FMN2* and *ARMCX4*). Deletion of one or more 59-bp VNTR units overlapping the last 8 bp of *MUC5B* exon 37 left coding sequences and splice sites intact.

To assess isoform differences and detect imprinted genes, we generated long-read Iso-Seq data for 12 of the 65 individuals (EBV-transformed lymphoblastoid B cell lines) and aligned these to donor-matched haplotype assemblies (Fig. 2b, Extended Data Fig. 3a and Supplementary Methods). Using our SV callset (Methods), we identified 136 structurally variable protein-coding gene sequences (Supplementary Table 42 and Supplementary Methods). Of these 136 genes, 58% (n = 79) contained a common SV (allele frequency > 0.05; Extended Data Fig. 3b). One example, *ZNF718*, creates nine unique isoforms (Fig. 2c) due to a common (allele frequency = 0.55) 6,142-bp polymorphic deletion that removes exons 2 and 3 from the canonical transcript as well as the 3′ part of an exon annotated as an alternate first exon (Extended Data Fig. 3b). Across the 14 wild-type *ZNF718* haplotypes, we found three known isoforms and four previously unreported isoforms (Methods). In contrast to other protein-coding genes with a single SV (Extended Data Fig. 3c), we found greater transcript diversity among the variant haplotypes of *ZNF718* than wild-type haplotypes. We also searched for SVs affecting nearby gene expression (RNA-seq) and identified 122 unique SVs proximal (less than 50 kb) to 98 differentially expressed genes across the 12 individuals, representing an enrichment compared with randomly permuted SVs (Extended Data Fig. 3d; empirical P = 0.001; Supplementary Table 43 and Supplementary Fig. 29; see Methods). Genome-wide, SVs were depleted across protein-coding genes and regulatory regions in the genome, as expected[28] (Extended Data Fig. 3e,f and Supplementary Fig. 30). By intersecting these 122 SVs with Hi-C data from the same individuals, we found that 29 of the SVs (associated with 24 genes) correspond to contact density changes in chromatin conformation regions (Extended Data Fig. 3g, Supplementary Table 44 and Supplementary Methods). Finally, we identified 3,818 SVs in high linkage disequilibrium with single-nucleotide polymorphism (SNP) loci from genome-wide association studies (GWAS) of human disease (Extended Data Fig. 3h and Supplementary Table 45; see Methods).

Genotyping and integrated reference panel

Genome-wide genotyping with PanGenie

Pangenome references have enabled genome inference, a process leveraging haplotype structures to genotype all variation encoded within a pangenome in a new individual from short-read whole-genome sequencing data[3]. We therefore constructed a pangenome graph containing all 65 genomes assembled here as well as 42 HPRC genome assemblies[1] with Minigraph-Cactus and detected variants by identifying graph bubbles relative to T2T-CHM13 (Methods). We used PanGenie to genotype bubbles across all 3,202 individuals from the 1kGP cohort based on Illumina data[29] and decomposed the 30,490,169 bubbles into 28,343,728 SNPs, 10,421,787 indels and 547,663 SV alleles[1] (Supplementary Fig. 31; see Methods). Leave-one-out experiments confirmed high genotype concordance of up to approximately 94% for biallelic SVs (Supplementary Figs. 32–34), and filtering the genotypes[1,8] resulted in a set of reliably genotypable variants comprising 25,695,951 SNPs, 5,774,201 indels and 478,587 SV alleles (Supplementary Table 46, Supplementary Figs. 35 and 36 and Supplementary Methods). We note that this set of SV alleles is larger than our main PAV callset (188,500 SVs) because it includes the HPRC genome assemblies and at the same time retains all SV alleles at multi-allelic sites (Supplementary Fig. 37 and Supplementary Methods).

We compared our genotyped set to other SV sets for the 1kGP cohort, including the HPRC PanGenie genotypes that we produced previously[1], as well as the 1kGP short-read high-coverage SV callset (1kGP-HC)[29] (Supplementary Figs. 38 and 39). On average, we found 26,115 SVs per genome, whereas this number was 18,462 for the HPRC genotypes and 9,596 for the 1kGP-HC SV calls. We specifically observed increases for rare variants (allele frequency < 1%; Fig. 2d). While the average number of rare SVs per genome was 87 for non-African individuals in the HPRC set and 169 in the 1kGP-HC set, we can now access on average

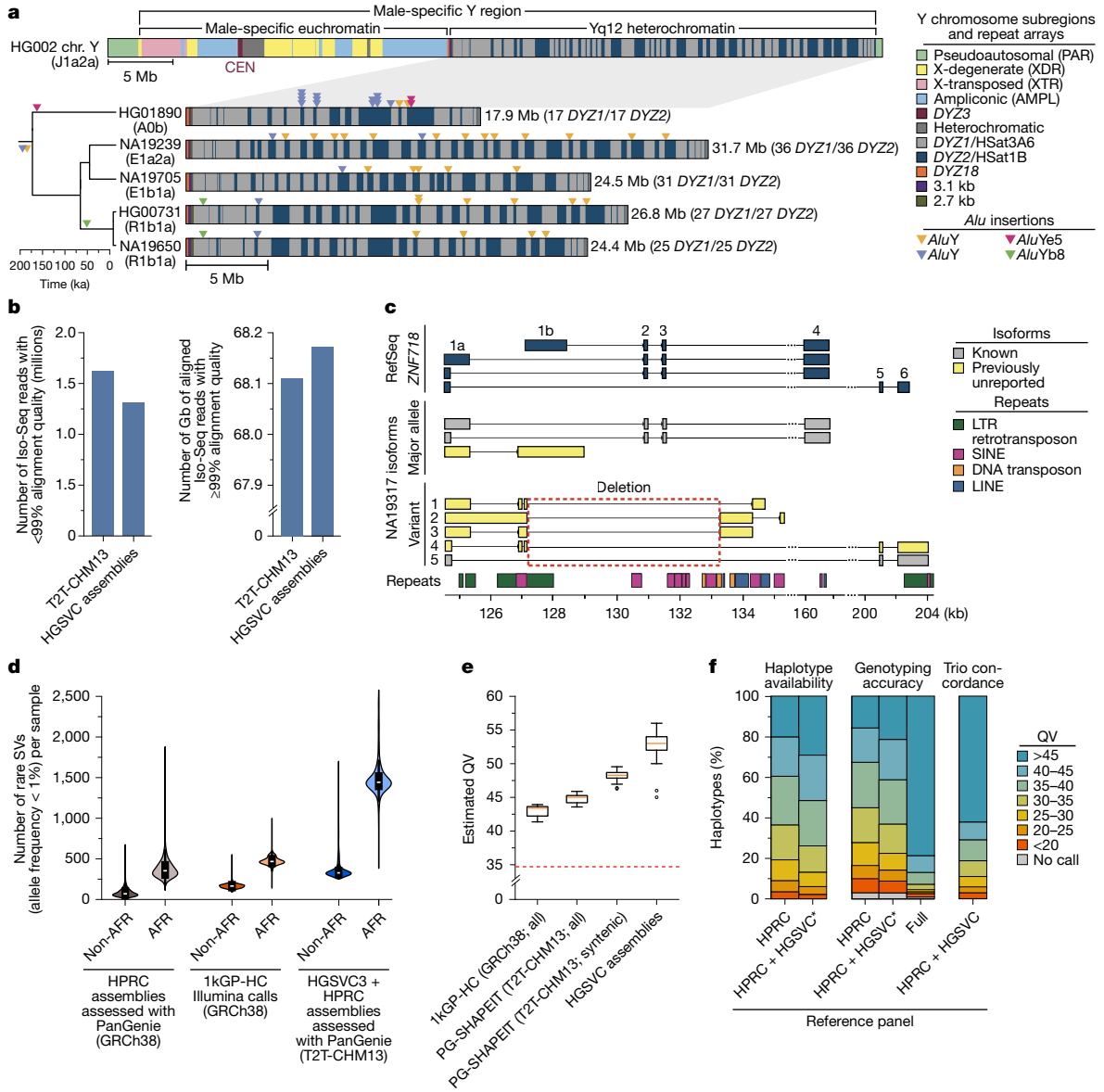


Fig. 2 | An improved genomic resource for challenging loci. a, Structure of a human Y chromosome, including the centromere (CEN; top), and repeat composition of five contiguously assembled Yq12 heterochromatic regions with their phylogenetic relationships (bottom left), size or number of *DYZ1* and *DYZ2* repeat array blocks (bottom right), and *Alu* insertion locations (triangles). ka, thousand years ago. **b**, Number of Iso-Seq reads that fail to align with 99% or less accuracy (left), and number of gigabases (Gb) of Iso-Seq reads that align with 99% or more accuracy (right) to the T2T-CHM13 reference genome versus the assemblies in this study. **c**, Expressed isoforms of *ZNF718* in NA19317. This individual is heterozygous for a deletion (red box, chr. 4: 127125–133267) that affects the *ZNF718* exon–intron structure. Isoforms not previously annotated in RefSeq, GENCODE or CHESS (Methods) are shown (yellow). LTR, long terminal repeat; SINE, short interspersed nuclear element; LINE, long interspersed nuclear element. **d**, Number of rare (allele frequency < 1%) SVs per sample in the HPRC-genotyped callset (grey), Illumina-based 1kGP-HC SV callset (orange), and combined HPRC and HGSVC-genotyped callset (blue) for both non-African (non-AFR) and African (AFR) individuals (n = 3,202). The first and third quartiles (Q1 and Q3, respectively; black boxes), median (white dots), and minima and maxima (black lines) are shown. **e**, Estimated k-mer-based QV for 60 haplotypes from the 1kGP-HC-phased set (GRCh38 based), HGSVC-phased genotypes using PanGenie, SHAPEIT5 (PG-SHAPEIT, T2T-CHM13 based) and all HGSVC genome assemblies. 'Syntenic' refers to regions of T2T-CHM13 also present in GRCh38. Baseline QV estimated by randomizing samples (red dashed line), first and third quartiles (black boxes), median (orange line), outliers (white dots) and whiskers (quantile 1 − 1.5(quantile 3−quantile 1) and quantile 3 + 1.5(quantile 3−quantile 1)) are shown. **f**, Haplotype availability, Locityper genotyping accuracy and trio concordance across 347 polymorphic loci in terms of variant-based QV. Availability and accuracy are calculated for 61 HGSVC individuals, whereas trio concordance is calculated for 602 trios. Full, HPRC + HGSVC; HPRC, HPRC only; HPRC + HGSVC*, HPRC + HGSVC leave-one-out.

362 rare alleles. For African individuals, we detected 1,490 rare SVs per genome, whereas there were 382 previously for the HPRC and 477 for the 1kGP-HC set.

Personal genome reconstruction

Next, we asked to what extent our improved genotyping abilities allow us to reconstruct the full haplotypic sequences of genomes sequenced with short reads. To this end, we combined our filtered PanGenie genotypes with rare SNP and indel calls obtained from Illumina reads for all 3,202 1kGP individuals (Methods) and phased this combined set using SHAPEIT5 (Supplementary Fig. 31, step 3, and Supplementary Figs. 40 and 41; see Methods).

We produced consensus haplotype sequences for all 3,202 individuals (6,404 haplotypes) by implanting the phased variants into T2T-CHM13

(only chromosomes 1–22 and X chromosome) and compared with consensus haplotypes produced from the GRCh38-based phased 1kGP-HC panel[29]. While the median *k*-mer-based quality value of the long-read assemblies was 53, we observed a median *k*-mer-based quality value of 45 for the consensus haplotypes computed from our short-read-based phased genotypes (Fig. 2e and Supplementary Fig. 42). To enable a fair comparison with the GRCh38-based 1kGP-HC consensus haplotypes, we additionally computed our *k*-mer-based quality value estimates restricted to regions shared between T2T-CHM13 and GRCh38 ('CHM13-syntenic'). For these regions, we observed a median quality value of 48, whereas the quality value for the 1kGP-HC set was lower (median of 43; Fig. 2e and Supplementary Fig. 42). In addition, we observed higher *k*-mer completeness values (median of 97.4%) than for the 1kGP-HC-phased set (median of 97.1%; Extended Data Fig. 4a and Supplementary Fig. 42). Because *k*-mer-based quality value estimates do not fully capture structural sequence correctness, we additionally used PAV to compute variant-calling-based quality value estimates for each 1-Mb genomic window (Methods). This expectedly resulted in lower quality value estimates (median quality value for 1kGP-HC of 26.7; median quality value for PanGenie of 34.2), but confirms the gain of PanGenie over standard short-read pipelines (Supplementary Figs. 43–45). Of note, PanGenie enables an accurate genome reconstruction of quality value > 30 routinely (78% of all 1-Mb windows), whereas that is rarely achieved for the 1kGP-HC callset (24% of all 1-Mb windows).

Targeted genotyping of complex loci

Although PanGenie performed well in this genome-wide setting, its use of *k*-mer information could make it difficult to genotype complex, repeat-rich loci with few unique *k*-mers. We therefore used the targeted method Locityper[30] to genotype the 1kGP cohort across 347 polymorphic targets covering 18.2 Mb and 494 protein-coding genes (Methods), including 268 challenging medically relevant genes[31]. For this challenging set of regions, the 1kGP-HC callset reaches a variant-based quality value of 30 for only 34.5% and a variant-based quality value of 40 for only 12.8% of predictions[30].

The performance of Locityper is constrained by the haplotypes available in the reference set. Therefore, we first evaluated haplotype availability by comparing sequences of the unrelated assembled haplotypes. Across all target loci, 51.5% of our assembled haplotypes were similar (variant-based quality value ≥ 40) to some other haplotype from the full reference panel described above, compared with only 39.6% of haplotypes when restricting to an HPRC-only reference panel[1] (Fig. 2f).

The increased haplotype availability translates into improved genotyping of polymorphic loci and we observed 80.0% haplotypes to be predicted with variant-based quality value ≥ 30 using a leave-one-out experiment compared with 74.6% haplotypes for the HPRC-only panel (Methods). These global improvements are mirrored by improvements of individual genes (Extended Data Fig. 4b), including *HLA-DRB5*, *HLA-DPA1* and *HLA-B* (Extended Data Fig. 4c). Finally, we asked what performance could potentially be achieved for growing reference panels and therefore used the full reference panel, including samples to be genotyped. Here Locityper predicts haplotypes with average quality value of 45.8, suggesting that sequence resolution of more reference haplotypes will aid future re-genotyping of challenging medically relevant genes, with applications to disease cohorts.

Major histocompatibility complex

Given the disease relevance and complexity of the 5-Mb MHC region[32–34] (Fig. 3a), we annotated 27–33 human leukocyte antigen (HLA) genes and 140–146 non-HLA genes or pseudogenes along with the associated repeat content of the 130 complete or near-complete MHC haplotypes (Supplementary Table 47). While 99.2% (357 of 360) of the HLA alleles agree with classical typing results[35] (Supplementary Tables 48 and 49), we resolved a total of 826 incomplete HLA allele annotations in the IPD-IMGT/HLA reference database[36] (Supplementary Table 50), including 112 sequences from the HLA-DRB loci, important for vaccine response and autoimmune disease[37,38]. We detected 170 SVs absent from reported reference haplotypes[39,40] (Supplementary Table 51), including a deletion of *HLA-DPA2* (HG03807, haplotype 1).

The observed MHC class II haplotypes reflect the established DR group system (Fig. 3b and Supplementary Table 52) and comprise representatives of DR5, DR8 and DR9, which have not previously been analysed in detail[39,40]. In this system, the functional DRB3, DRB4 and DRB5 genes differentially associate across the DR groups, with DR1 and DR8 groups uniquely lacking either of them. Repeat element analyses (Supplementary Figs. 46–48; see Methods) suggest that DR8 arose from an intrachromosomal deletion mediated by 150 bp of sequence homology between *HLA-DRB1* and *HLA-DRB3* on the DR3/5/6 haplotype, as previously reported[41] (Fig. 3c). DR1 is most likely derived by recombination between DR2 and DR4/7/9 (Fig. 3c and Supplementary Figs. 46 and 49). Finally, our catalogue of solitary *HLA-DRB* exon sequences[42] includes refined copy number estimates (for example, three solitary *HLA-DRB* exon 1 sequences instead of one in the *HLA-DRB9* region of DR1), as well as identification of a polymorphic, solitary exon 10 kb 3′ of *HLA-DRB1* (Fig. 3b; see Methods).

Similarly, we characterized the RCCX (*STK19* (R), *C4* (C), *CYP21* (C) and *TNX* (X)) multi-allelic cluster (Fig. 3d, Supplementary Table 53 and Supplementary Fig. 50), in which phasing and variant classification has been challenging due to extensive sequence homology[43]. Tandem duplications (aka RCCX bi-modules) are the most abundant (74.6% or $n = 97$), with mono-modules and tri-modules comparable in frequency (13.1% ($n = 17$) and 12.3% ($n = 16$), respectively; Supplementary Fig. 50). Resolved haplotypes also facilitate the detection of interlocus gene conversion events critical for RCCX evolution[44], such as two haplotypes with a tri-modular RCCX with two functional *CYP21A2* copies, one mono-modular and one bi-modular haplotype with no functional *CYP21A2* genes; and one tri-modular haplotype with a unique configuration where *C4B* precedes *C4A* and carries two *CYP21A2* copies, one of which being non-functional (Fig. 3d). We suggest that the latter haplotype was generated by introduction of a nonsense mutation and two gene conversion events, converting *CYP21A1P* into *CYP21A2* and *C4A* into a *C4B* that now unusually encodes the Rodgers blood group epitope. We also identified seven novel *C4* amino acid variants (Supplementary Figs. 51 and 52).

Next, we evaluated the performance of Locityper across 19 MHC protein-coding genes and 14 pseudogenes. Across all 33 loci, Locityper correctly predicted gene alleles in 81.0% cases when restricting to a limited HPRC-only reference panel (45 individuals)[1]. Inclusion of our assemblies ($n = 107$ individuals or 214 phased haplotypes) increased accuracy to 86.3% (leave-one-out experiment) and 97.1% (full panel leveraging all 214 phased haplotypes; Extended Data Fig. 4c), underscoring the value of accurate phased assemblies for the interpretation of short-read data.

Finally, we tested whether the established HLA class II DR group nomenclature could be recapitulated using unbiased, sequence-based analysis. Applying a pangenomic multiscale approach, PGR-TK[45] (Fig. 3e), to a subset of our genomes ($n = 55$) as well as T2T-CHM13 (ref. 4), we identified 63 conserved blocks greater than 6 kb. Multiscale hierarchical clustering of the haplotypes perfectly reconstituted the traditional DR group system in the region around *HLA-DRB1* (Fig. 3e). However, we also observed additional diversified subgroups indicating the possibility for a more fine-grained future classification of HLA-DR haplotypes or utility in the context of GWAS, especially when coupled with the improved targeted genotyping ability (Extended Data Fig. 4c).

Complex structural polymorphisms

Long-read-assembled genomes significantly enhance the detection and characterization of complex structural variants (CSVs) defined

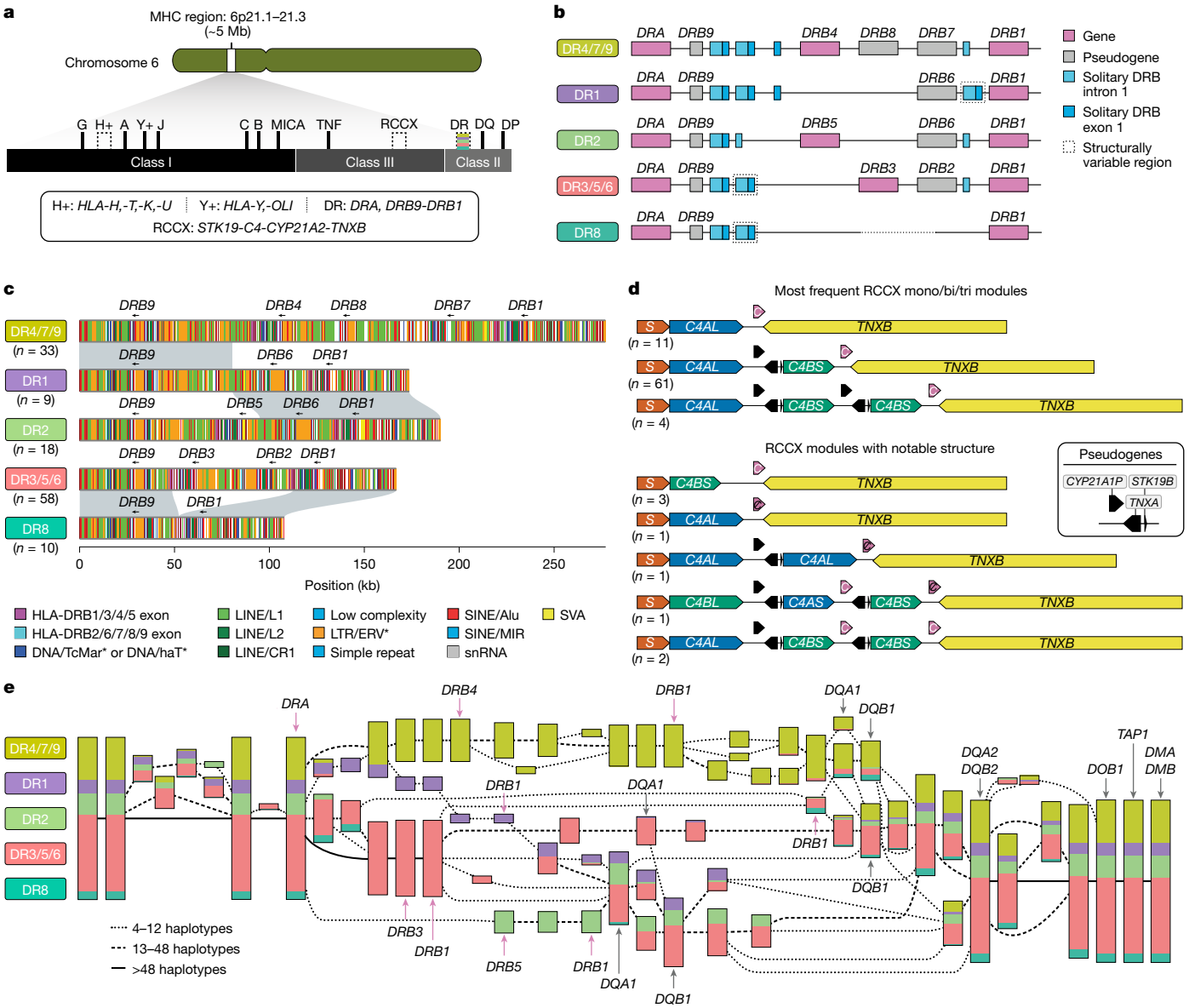


Fig. 3 | Structurally variable regions of the MHC locus. a, Overview of the organization of the MHC locus into class I, class II and class III regions and the genes contained therein. Structurally variable regions are indicated by dashed lines. The coloured stripes show the approximate location of the regions analysed in **b**–**d**. **b**, Gene content and locations of solitary *HLA-DRB* exon 1 and intron 1 sequences in the HLA-DR region of the MHC locus by the DR group, an established system for classifying haplotypes in the HLA-DR region according to their gene or pseudogene structure and their *HLA-DRB1* allele. **c**, High-resolution repeat maps and locations of gene or pseudogene exons for different DR group haplotypes in the HLA-DR region, highlighting sequence homology between the DR1 and DR4/7/9 and DR2, and between the DR8 and DR3/5/6, haplotype groups, respectively. Also shown is the number of analysed MHC haplotypes per DR group. CR1, chicken repeat 1; ERV, endogenous retrovirus; MIR, mammalian interspersed repeat; snRNA, small nuclear RNA. **d**, Visualization of common and notable RCCX haplotype structures observed in the HGSVC MHC haplotypes, showing variation in gene and pseudogene content as well as the modular structure of RCCX (*STK19* (S), non-functional *CYP21A2* (black C), functional *CYP21A2* (white C) and *C4L*/*S* (long ((HERV-K insertion)/short(no HERV-K insertion))). **e**, Visualization of a PGR-TK analysis of 55 MHC loci and T2T-CHM13 for 111 haplotypes in total. The colours indicate the relative proportion of distinct DR group haplotypes flowing through the visualized elements.

here as a single event composed of simple SVs spanning more than one repair junction. Because CSV breakpoints are often located in repetitive sequences, including SDs and MEIs[46–49], we recently updated PAV[8] to identify CSVs embedded in large complex repeats such as SDs (Methods). Using this method against the T2T-CHM13 reference genome, we found on average 72 CSVs per genome[50] (range of 51–91; Supplementary Table 54; see Data availability). Across all genomes, we identified 1,247 CSVs with 128 distinct complex reference signatures[50], consistent with known CSVs derived from diverse individuals[51]. We found that 27% of CSVs have locally duplicated sequences, and 38% have local inversions. Many of the complex structures that we identified are mediated by SDs, such as INVDUP-INV-DEL (174 CSVs and 92% SDs), DEL-INV-DEL (34 CSVs and 21% SDs) and INVDUP-INV-INVDUP (8 CSVs and 75% SDs) where DEL is a reference deletion, INV is an inverted sequence that is not duplicated and INVDUP is a duplicated inversion (one copy in each orientation)[50]. As an example, we highlight two CSVs involving *NOTCH2NL* and *NBPF*, genes implicated in the expansion of the human brain during evolution[8], as well as a core duplicon associated with genomic instability[52]. Although the full structures could not be resolved by previous optical mapping or sequencing experiments, we can distinguish three distinct haplotype structures, including a reference haplotype (13.7% allele frequency), a 930-kb CSV (DEL-INV-DEL) inverting *NBPF8* and deleting

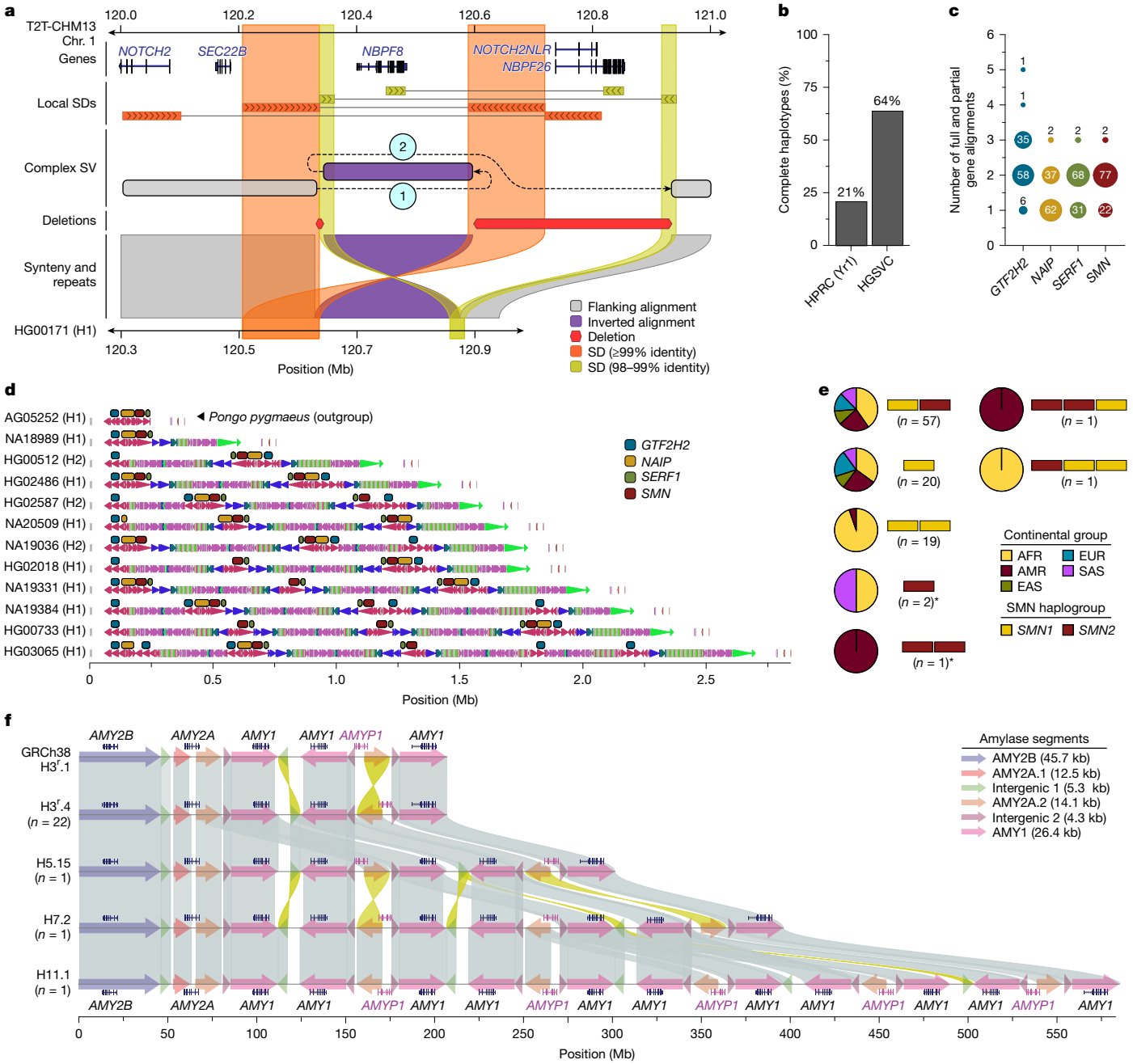


Fig. 4 | Complex SVs in human populations. a, An SD-mediated CSV inverts *NBPF8* and deletes *NOTCH2NLR* and *NBPF26*. Inverted SD pairs (orange and yellow bands) each mediate a template switch (dashed lines '1' and '2'). PAV refines alignment artefacts in large repeats surrounding CSVs to obtain a more accurate representation of these structures. The allele shown is HG00171 haplotype 1. **b**, Fraction of all assemblies having complete and accurate sequence over the SMN region, stratified by study (HPRC-Yr1 and HGSVC). **c**, Copy number (full and partial gene alignments) of each multi-copy gene (*SMN1/2* in red, *SERF1A/B* in green, *NAIP* in gold and *GTF2H2/C* in blue) across assembled haplotypes ($n = 101$). **d**, SMN duplications from 11 diverse human haplotypes assembled from this study, the HPRC (HG02486) and one *Pongo pygmaeus* haplotype (top) used as an outgroup. **e**, Summary of *SMN1* (yellow) and *SMN2* (red) gene copies genotyped across human haplotypes ($n = 101$). The yellow and red bars show a unique copy number of *SMN1* and *SMN2*, whereas the pie charts show their relative proportions in continental groups. The asterisks show haplotypes with only *SMN2* gene copies. **f**, The structure of the human amylase locus shows amylase genes (coloured arrows) and alignments between haplotypes (99–100% sequence identity). The H3^r.4 haplotype represents the most common haplotype, H5.15 and H7.2 are haplotypes previously unresolved at the base-pair level, and H11.1 is a previously unknown haplotype. Amylase gene annotations are displayed above each haplotype structure. The structure of each amylase haplotype, composed of amylase segments, is indicated by the coloured arrows. Sequence similarity between haplotypes ranges from 99% to 100%.

NOTCH2NLR and *NBPF26* (35.9% allele frequency; Fig. 4a), and a 513-kb CSV with a distal template switch replacing *NBPF8* with *NBPF9* (50.8% allele frequency; Supplementary Fig. 53).

As a second example, the structurally complex region containing *SMN1* and *SMN2* gene copies is associated with spinal muscular atrophy, and successful ASO-mediated gene therapies involve *SMN2* (refs. 53,54). The genes are embedded in a large SD region (approximately 1.5 Mb) that has been almost impossible to fully sequence resolve despite the advances of the past two decades[1,2,8] (Supplementary Fig. 54). We successfully assembled, validated and characterized 101 haplotypes to fully

resolve the structure and copy number of *SMN1/2*, *SERF1A/B*, *NAIP* and *GTF2H2/C* (Methods). We found that 48% (n = 48) of haplotypes carry exactly two copies of *SMN1/2*, *SERF1A/B* and *GTF2H2/C*, whereas *NAIP* is present mostly in a single copy. We highlight 11 human haplotypes showing increasing complexity (Fig. 4b–d). We specifically distinguished functional *SMN1* and *SMN2* copies based on our assemblies (Supplementary Fig. 55) and compared them with the short-read-based genotyping methods Parascopy and SMNCopyNumberCaller (Methods). For individuals with two fully assembled haplotypes (n = 31), predicted *SMN1/2* copy numbers matched perfectly among the three methods (Supplementary Fig. 56). Our analysis shows that 98 haplotypes carry the ancestral *SMN1* copy but three do not and are potentially disease-risk loci that may have arisen as a result of interlocus gene conversion (Fig. 4e and Supplementary Fig. 57).

Finally, we analysed the complex amylase locus spanning 212.5 kb on chromosome 1 (GRCh38; chr. 1: 103554220–103766732) and containing genes *AMY2B*, *AMY2A*, *AMY1A*, *AMY1B* and *AMY1C*[55] (Fig. 4f). From 65 sequence-resolved genomes, we identified 39 distinct amylase haplotypes, capturing approximately 83% of the haplotypes in the population (Supplementary Table 55 and Supplementary Figs. 58 and 59), 35 of which were supported by both Verkko and optical genome mapping de novo assemblies. The length of these amylase haplotypes ranges from 111 kb ($H1^a$.1 and $H1^a$.2) to 582 kb (H11.1; Fig. 4f), including those that are structurally identical to the GRCh38 ($H3^r$.1) and T2T-CHM13 (H7.3) assemblies. Among these, four are common: $H1^a$.1 (n = 14), $H3^r$.1 (n = 13), $H3^r$.2 (n = 19) and $H3^r$.4 (n = 22; constituting 57% of all genomes), whereas 23 are singletons. We identified nine haplotypes previously supported only by optical genome mapping data and fully sequence resolved the largest haplotype (H11.1; 11 *AMY1* (8.8 kb) copies)[55–57] (Fig. 4f).

Centromeres

Human centromeres are among the most mutable genomic regions and are composed of tandemly repeating α-satellite DNA organized into higher-order repeats (HORs) spanning up to several megabases on each chromosome[58]. It has been estimated that approximately 22% of centromeres vary by over 1.5-fold in length, and approximately 30% of them vary in their structure[59]. To understand the genetic and epigenetic centromeric variation in these 65 individuals, we first assessed contiguity and accuracy using two assembly algorithms (Methods). We identified 822 Verkko centromeres and 777 hifiasm centromeres that were completely and accurately assembled. Only 28.3% were correctly assembled by both assemblers, with Verkko and hifiasm uniquely resolving a similar subset (37.7% and 34.1%, respectively). We combined these two datasets into a non-redundant set of 1,246 completely and accurately assembled centromeres (approximately 52 centromeres per chromosome and approximately 19.5 centromeres per genome, on average; Extended Data Fig. 5a and Supplementary Tables 56 and 57).

We first measured the variation in the length of the centromeric α-satellite HOR array (or arrays) on each chromosome. Although active centromeric α-satellite HOR arrays are, on average, 2.3 Mb in length, there is considerable variation, including outliers (Fig. 5a, Supplementary Table 57 and Supplementary Figs. 60 and 61). For example, the active α-satellite HOR arrays from chromosomes 3, 4, 10, 13–16, 21 and the Y chromosome are consistently smaller, whereas those on chromosomes 1, 11 and 18 are larger than average (Supplementary Fig. 61). Among the 1,246 centromeres, we identified 4,153 new α-satellite HOR variants and novel active α-satellite HOR array organizations (Fig. 5b and Supplementary Figs. 62 and 63). On chromosome 1, for example, we identified an insertion of monomeric α-satellite into the *D1Z7* α-satellite HOR array, effectively splitting the α-satellite into two distinct HOR arrays (Fig. 5b). A similar bifurcation event also occurred on the centromeres of chromosomes 12 and 19, generating two α-satellite HOR arrays where there typically is only one (Fig. 5b,c). In addition, we found novel α-satellite HOR array organizations for chromosomes 6 and 10 that differ from the CHM1 and CHM13 arrays on those chromosomes[59] (Fig. 5b and Supplementary Fig. 62b,c). These array organizations, which are the most common in our dataset, are primarily composed of either 18-monomer α-satellite HORs (chromosome 6) or 6-monomer α-satellite HORs (chromosome 10).

To determine how variation in centromeric sequence and structure affects their epigenetic landscape, we assessed the CpG methylation pattern along each centromere using native ONT data. We found that all centromeres contain at least one region of hypomethylation (termed the 'centromere dip region' (CDR))[58,60], which is thought to mark the site of the kinetochore. However, in many cases, such as on chromosomes 6, 15 and 19, there were at least two CDRs more than 80 kb apart (Fig. 5b, Extended Data Fig. 5b–d and Supplementary Fig. 64). This suggests the presence of a 'di-kinetochore', which may form a dicentric chromosome on approximately 7% of chromosomes, but additional analyses that assess the location of the centromeric histone H3 variant, CENP-A, will need to be performed to confirm these putative kinetochore sites. We generated sequence identity heatmaps of each centromere and found that the CDR often resides within the most highly identical regions of the α-satellite HOR arrays (Fig. 5c and Extended Data Fig. 5d). Even when the α-satellite HOR array is split into two arrays, such as on chromosome 19, the CDR associates with the array containing some of the most highly identical α-satellite HORs (Extended Data Fig. 5d). This suggests that the kinetochore may track with actively homogenizing α-satellite HOR sequences in response to a co-evolution between centromeric DNA and proteins[61].

MEI investigation in many of the α-satellite HOR arrays (Methods) revealed that approximately 30% contained at least one MEI. In total, we identified 89 unique polymorphic insertions with varying allele frequencies (Supplementary Table 58), with L1HS being the most prevalent (58%), followed by *Alu* elements (41%) and SVAs (1%). The *D2Z1* α-satellite HOR array on chromosome 2 was highly enriched with MEIs (Fig. 5d), with at least one L1HS and/or *Alu* insertion in 80% of haplotypes (Supplementary Fig. 65). Although L1HS insertions or duplications were the most common, occurring on average three times per array, three unique *Alu* insertions (two *Alu*Yb8 and one *Alu*Ya5) were also present, albeit with low allele frequency. Nearly all insertions, as well as their duplications, were located outside of the CDRs and typically towards the periphery. However, one *Alu*Yb8 insertion (NA20509 (H1)) was located between two CDRs and appeared to 'break' a single CDR into two, whereas a pair of L1HSs were found on either side of a CDR in two haplotypes (NA19331 (H1) and NA19650 (H1)), possibly acting as boundaries that restrict CDR and CENP-A chromatin movement, as previously suggested[62].

Discussion

LRS and assembly have enabled both the full resolution of a human genome sequence[4] and fundamentally deepened our understanding of human genetic diversity[1,8,13,63]. The development of a human pangenome reference[1,64] requires ideally completely phased and assembled diverse genomes. Although hundreds of genomes are being assembled as part of international efforts[65], practically, few are yet truly T2T. Meanwhile, pangenome augmentation methods based on shallow long-read data have been used to capture variants with lower allele frequencies[66]. Nevertheless, algorithms and technology have advanced significantly, and we have demonstrated that more than 99% of the human genome can be accurately phased and assembled by focusing on 65 diverse humans (130 haplotypes). We characterized regions previously excluded or collapsed[1,2], including centromeres, biomedically complex regions such as *SMN1/SMN2*, the MHC and thousands of more complex SV patterns.

Combining our assemblies with previous HPRC assemblies to create a reference set, we were able to reconstruct a genome from short

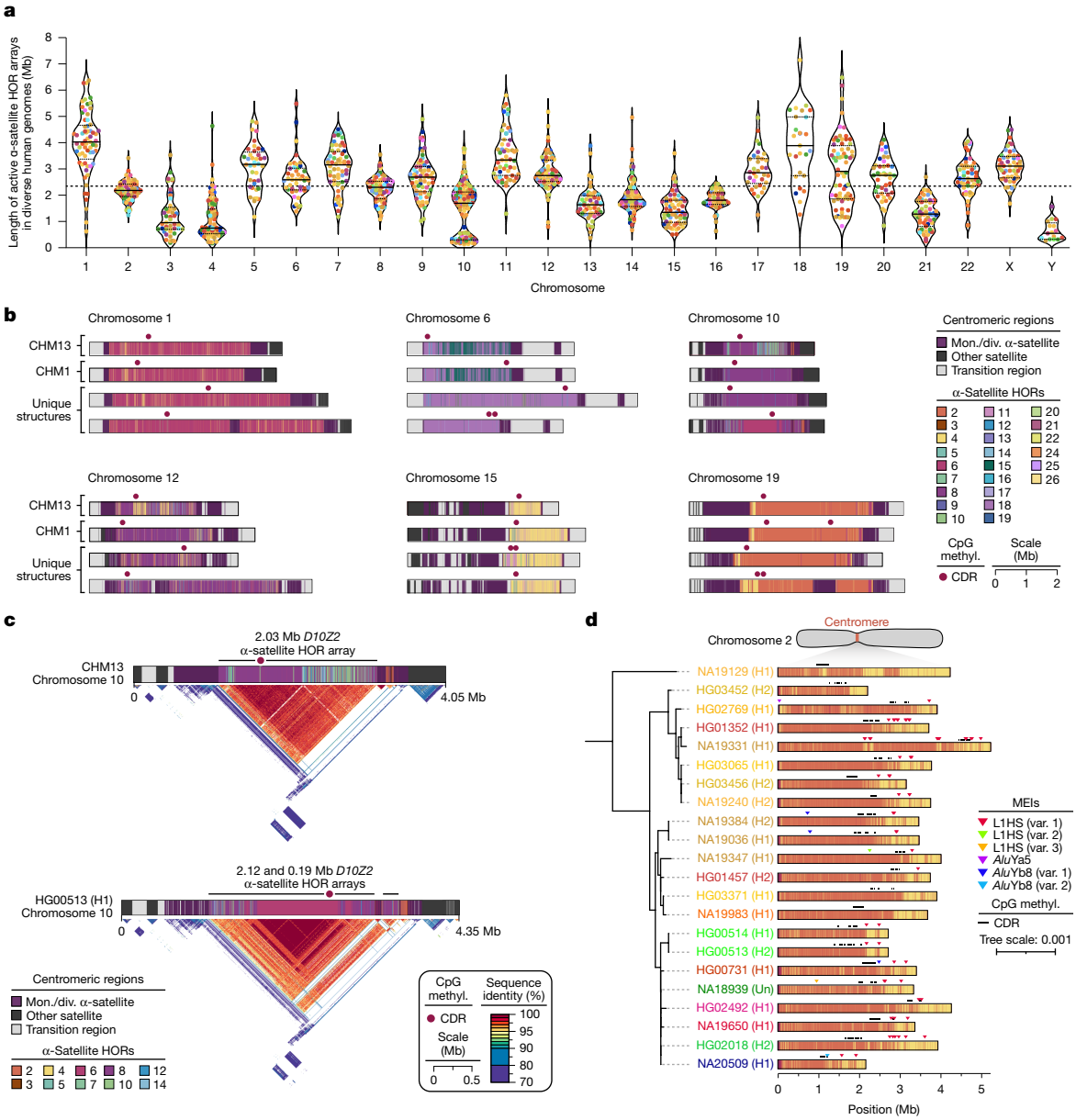


Fig. 5 | Variation in the sequence, structure and methylation pattern among 1,246 human centromeres. a, Length of the active α-satellite HOR array (arrays) for each complete and accurately assembled centromere from each genome. Each data point indicates an active α-satellite HOR array and is coloured by population group. The median length of all α-satellite HOR arrays is shown as a dashed line. For each chromosome, the median (solid line) and first and third quartiles (dashed lines) are shown. **b**, Sequence, structure and methylation (methyl.) map of centromeres from CHM13, CHM1 and a subset of 65 diverse human genomes. The α-satellite HORs are coloured by the number of α-satellite monomers within them, and the site of the putative kinetochore, indicated by the CDR, is shown. Mon., monomeric; div., divergent. **c**, Differences in the α-satellite HOR array organization and methylation patterns between the CHM13 and HG00513 (H1) chromosome 10 centromeres. The CDRs are located on highly identical sequences in both centromeres, despite their differing locations. **d**, MEIs in the chromosome 2 centromeric α-satellite HOR array. Most MEIs are consistent with duplications of the same element rather than distinct insertions, and all of them reside outside of the CDR. Var., variant.

reads to an average base error of about 0.00158% (quality value of 48). This process detects 26,115 SVs per genome on average from short-read sequence data and notably now recovers more rare SVs (allele frequency < 1%) than direct variant discovery from short reads. This advance was made possible by improvements in assembly quality, the larger sample size, improved versions of the Minigraph-Cactus and PanGenie applications, and the switch to the more complete T2T-CHM13 reference genome. As the number of HPRC genomes increases to several hundreds and they reach T2T status[65], genotyping accuracy will probably improve further. This, in turn, will make disease-association studies from short reads considerably more powerful for complex variation.

Using our assembly method, we fully assembled 1,246 centromeres – 42% of all possible centromeres in these individuals. As expected, we observed considerable variation in the content and length of the α-satellite HOR array (up to 37-fold for chromosome 10) consistent with its higher mutation rate and more rapid evolutionary turnover[2,59]. We have also documented recent *Alu*, L1 and SVA retrotransposition into the α-satellite HORs and showed that these may be used to tag HOR expansions on particular human haplotypes. Using the CDR[58,60] as a marker of kinetochore attachment, we have shown considerable variation in the location across human centromeres and remarkably that 7% of human chromosomes show evidence of two or more putative kinetochores (that is, di-kinetochores)

in lymphoblastoid cell lines. The significance of both MEIs and di-kinetochore on chromosome segregation or missegregation will need to be experimentally assessed, and these phased genomes (and their corresponding cell lines) provide the foundation for such future work.

Finally, from a technical perspective, application of two independent assembly algorithms, hifiasm (ultra-long) and Verkko, nearly doubled the number of sequence-resolved centromeres. Although the two methods were strongly complementary for centromeres, Verkko was clearly superior for the Y chromosome (Supplementary Fig. 26c). As the performance of both Verkko and hifiasm has been shown to be very similar for large portions of the euchromatin[10], there is benefit in applying both assembly algorithms to resolve the most structurally complex regions of the genome until a tool combining the strengths of both methods becomes available.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-025-09140-6.

1. Liao, W.-W. et al. A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
2. Porubsky, D. et al. Gaps and complex structurally variant loci in phased genome assemblies. *Genome Res.* **33**, 496–510 (2023).
3. Ebler, J. et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.* **54**, 518–525 (2022).
4. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
5. Garg, S. et al. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-020-0711-0 (2020).
6. Porubsky, D. et al. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-020-0719-5 (2020).
7. Koren, S. et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **36**, 1174–1182 (2018).
8. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
9. Rautiainen, M. et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* **41**, 1474–1482 (2023).
10. Cheng, H., Asri, M., Lucas, J., Koren, S. & Li, H. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. *Nat. Methods* **21**, 967–970 (2024).
11. 1000 Genomes Project Consortiumet al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
12. Henglin, M. et al. Graphasing: phasing diploid genome assembly graphs with single-cell strand sequencing. *Genome Biol.* **25**, 265 (2024).
13. Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
14. Aganezov, S. et al. A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022).
15. Kazazian, H. H. Jr et al. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**, 164–166 (1988).
16. Porubsky, D. et al. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell* **185**, 1986–2005.e26 (2022).
17. Cooper, G. M. et al. A copy number variation morbidity map of developmental delay. *Nat. Genet.* **43**, 838–846 (2011).
18. Sudmant, P. H. et al. Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).
19. Sudmant, P. H. et al. Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**, aab3761 (2015).
20. Vollger, M. R. et al. Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).
21. Jeong, H. et al. Structural polymorphism and diversity of human segmental duplications. *Nat. Genet.* **57**, 390–401 (2025).
22. Hallast, P., Agdzhoyan, A., Balanovsky, O., Xue, Y. & Tyler-Smith, C. A Southeast Asian origin for present-day non-African human Y chromosomes. *Hum. Genet.***140**, 299–307 (2021).
23. Hallast, P. et al. Assembly of 43 human Y chromosomes reveals extensive complexity and variation. *Nature* **621**, 355–364 (2023).
24. Rhie, A. et al. The complete sequence of a human Y chromosome. *Nature* **621**, 344–354 (2023).
25. Porubsky, D. et al. Human de novo mutation rates from a four-generation pedigree reference. *Nature* **643**, 427–436 (2025).
26. Ruderfer, D. M. et al. Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nat. Genet.* **48**, 1107–1111 (2016).
27. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
28. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
29. Byrska-Bishop, M. et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440.e19 (2022).
30. Prodanov, T. et al. Locityper: targeted genotyping of complex polymorphic genes. Preprint at *bioRxiv* https://doi.org/10.1101/2024.05.03.592358 (2024).
31. Wagner, J. et al. Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat. Biotechnol.* **40**, 672–680 (2022).
32. Horton, R. et al. Gene map of the extended human MHC. *Nat. Rev. Genet.* **5**, 889–899 (2004).
33. Norman, P. J. et al. Sequences of 95 human MHC haplotypes reveal extreme coding variation in genes other than highly polymorphic HLA class I and II. *Genome Res.* **27**, 813–823 (2017).
34. Trowsdale, J. & Knight, J. C. Major histocompatibility complex genomics and human disease. *Annu. Rev. Genomics Hum. Genet.* **14**, 301–323 (2013).
35. Abi-Rached, L. et al. Immune diversity sheds light on missing variation in worldwide genetic diversity panels. *PLoS ONE* **13**, e0206512 (2018).
36. Barker, D. J. et al. The IPD-IMGT/HLA Database. *Nucleic Acids Res.* **51**, D1053–D1060 (2023).
37. Mentzer, A. J. et al. High-resolution African HLA resource uncovers HLA-DRB1 expression effects underlying vaccine response. *Nat. Med.* **30**, 1384–1394 (2024).
38. Liu, B., Shao, Y. & Fu, R. Current research status of HLA in immune-related diseases. *Immun. Inflamm. Dis.* **9**, 340–350 (2021).
39. Horton, R. et al. Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. *Immunogenetics* **60**, 1–18 (2008).
40. Houwaart, T. et al. Complete sequences of six major histocompatibility complex haplotypes, including all the major MHC class II structures. *Hladnikia* **102**, 28–43 (2023).
41. Gorski, J. The HLA-DRw8 lineage was generated by a deletion in the DR B region followed by first domain diversification. *J. Immunol.* **142**, 4041–4045 (1989).
42. Gongora, R. Presence of solitary exon 1 sequences in the HLA-DR region. *Hereditas* **127**, 47–49 (1997).
43. Chung, E. K. et al. Genetic sophistication of human complement components C4A and C4B and RP-C4-CYP21-TNX (RCCX) modules in the major histocompatibility complex. *Am. J. Hum. Genet.* **71**, 823–837 (2002).
44. Bánlaki, Z. et al. Intraspecific evolution of human RCCX copy number variation traced by haplotypes of the CYP21A2 gene. *Genome Biol. Evol.* **5**, 98–112 (2013).
45. Chin, C.-S. et al. Multiscale analysis of pangenomes enables improved representation of genomic diversity for repetitive and clinically relevant genes. *Nat. Methods* **20**, 1213–1221 (2023).
46. Gu, S. et al. Alu-mediated diverse and complex pathogenic copy-number variants within human chromosome 17 at p13.3. *Hum. Mol. Genet.* **24**, 4061–4077 (2015).
47. Balachandran, P. et al. Transposable element-mediated rearrangements are prevalent in human genomes. *Nat. Commun.* **13**, 7115 (2022).
48. Beck, C. R. et al. Megabase length hypermutation accompanies human structural variation at 17p11.2. *Cell* **176**, 1310–1324.e10 (2019).
49. Collins, R. L. et al. A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
50. Audano, P. A., Paisie, C., The Human Genome Structural Variation Consortium & Beck, C. R. Large complex structural rearrangements in human genomes harbor cryptic structures. Preprint at *bioRxiv* https://doi.org/10.1101/2024.12.19.629504 (2024).
51. Collins, R. L. et al. Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome Biol.* **18**, 36 (2017).
52. Marques-Bonet, T. & Eichler, E. E. The evolution of human segmental duplications and the core duplicon hypothesis. *Cold Spring Harb. Symp. Quant. Biol.* **74**, 355–362 (2009).
53. Winkelsas, A. M. et al. Targeting the 5′ untranslated region of SMN2 as a therapeutic strategy for spinal muscular atrophy. *Mol. Ther. Nucleic Acids* **23**, 731–742 (2021).
54. Sivanesan, S., Howell, M. D., Didonato, C. J. & Singh, R. N. Antisense oligonucleotide mediated therapy of spinal muscular atrophy. *Transl. Neurosci.* https://doi.org/10.2478/s13380-013-0109-2 (2013).
55. Bolognini, D. et al. Recurrent evolution and selection shape structural diversity at the amylase locus. *Nature* https://doi.org/10.1038/s41586-024-07911-1 (2024).
56. Yilmaz, F. et al. Reconstruction of the human amylase locus reveals ancient duplications seeding modern-day variation. *Science* **386**, eadn0609 (2024).
57. Usher, C. L. et al. Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. *Nat. Genet.* **47**, 921–925 (2015).
58. Altemose, N. et al. Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).
59. Logsdon, G. A. et al. The variation and evolution of complete human centromeres. *Nature* **629**, 136–145 (2024).
60. Gershman, A. et al. Epigenetic patterns in a complete human genome. *Science* **376**, eabj5089 (2022).
61. Shepelev, V. A., Alexandrov, A. A., Yurov, Y. B. & Alexandrov, I. A. The evolutionary origin of man can be traced in the layers of defunct ancestral alpha satellites flanking the active centromeres of human chromosomes. *PLOS Genet.* **5**, e1000641 (2009).
62. O'Neill, R. J., O'Neill, M. J. & Graves, J. A. Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature* **393**, 68–72 (1998).
63. Chaisson, M. J. P. et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
64. Gao, Y. et al. A pangenome reference of 36 Chinese populations. *Nature* **619**, 112–121 (2023).
65. Wang, T. et al. The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).

66. Schloissnig, S. et al. Structural variation in 1,019 diverse humans based on long-read sequencing *Nature* https://doi.org/10.1038/s41586-025-09290-7 (2024).

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [2]Department of Genetics, Epigenetics Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [3]Core Unit Bioinformatics, Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University, Düsseldorf, Germany. [4]Center for Digital Medicine, Heinrich Heine University, Düsseldorf, Germany. [5]The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA. [6]Department of Genetics and Biochemistry, Clemson University, Clemson, SC, USA. [7]Center for Human Genetics, Clemson University, Greenwood, SC, USA. [8]Institute for Medical Biometry and Bioinformatics, Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University, Düsseldorf, Germany. [9]Program in Medical and Population Genetics and Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [10]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [11]Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [12]Medical University of South Carolina, College of Graduate Studies, Charleston, SC, USA. [13]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA. [14]Pathos AI Inc., Chicago, IL, USA. [15]Institute of Medical Microbiology and Hospital Hygiene, Medical Faculty, Heinrich Heine University, Düsseldorf, Germany. [16]European Molecular Biology Laboratory, Wellcome Genome Campus, European Bioinformatics Institute, Cambridge, UK. [17]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. [18]Oncode Institute, Utrecht, The Netherlands. [19]Division of Computational Genomics and Systems Genetics, German Cancer Research Center, Heidelberg, Germany. [20]Department of Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT, USA. [21]Department of Biomedical Informatics and Data Science, Heersink School of Medicine, University of Alabama, Birmingham, AL, USA. [22]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA. [23]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. [24]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA. [25]Department of Structural Biology, School of Medicine, Stanford University, Stanford, CA, USA. [26]Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. [27]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, CA, USA. [28]Genome Informatics Section, Center for Genomics and Data Science Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [29]Department of Computer and Information Sciences, College of Science and Technology, Temple University, Philadelphia, PA, USA. [30]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA. [31]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA. [32]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [33]Department of Biomedical Informatics, University of Colorado School of Medicine, Aurora, CO, USA. [34]Department of Immunology and Microbiology, University of Colorado School of Medicine, Aurora, CO, USA. [35]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. [36]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [37]Stanford Health Care, Palo Alto, CA, USA. [38]New York Genome Center, New York, NY, USA. [39]The University of Connecticut Health Center, Farmington, CT, USA. [40]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [41]Present address: The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA. [42]These authors contributed equally: Glennis A. Logsdon, Peter Ebert, Peter A. Audano, Mark Loftus. ✉e-mail: mkonkel@clemson.edu; jan.korbel@embl.org; Charles.Lee@jax.org; Christine.Beck@jax.org; ee3@uw.edu; tobias.marschall@hhu.de

Methods

Sample selection

A total of 65 diverse humans were included in the current study. The majority of the individuals (63 of 65) originated from the 1kGP sample set[11], one (NA21487) from the International HapMap Project[67] and one (NA24385, also called HG002) commonly used for benchmarking by the Genome in a Bottle (GIAB) Consortium[68] was included in all analyses with publicly available data from other efforts (Supplementary Tables 1–4, 6 and 7). Individuals were selected to maximize genetic diversity and Y chromosome lineages (Supplementary Methods).

Data production

In addition to data generated through previous efforts[8,23], sequencing libraries were prepared from high-molecular-weight DNA or RNA extracted from lymphoblastoid cell lines (Coriell Institute). PacBio HiFi sequencing data were generated on the Sequel II or Revio platforms using 30-h movie times. UL ONT libraries were generated using a modified fragmentase protocol and sequenced on R9.4.1 flow cells on a PromethION instrument for 96 h. Bionano Genomics optical mapping data using DLE-1 tagging were collected on Saphyr 2nd generation instruments. Strand-seq data were produced using BrdU incorporation and second-strand DNA removal during PCR-based library construction to generate single-nucleus barcoded libraries sequenced on an Illumina NextSeq 500 platform[69,70]. Hi-C data were collected using Proximo Hi-C kits (v.4.0; Phase Genomics) and sequenced on an Illumina NovaSeq 6000. RNA-seq libraries were generated using KAPA RNA Hyperprep with RiboErase (Roche) and sequenced on an Illumina NovaSeq 6000 platform. Iso-Seq full-length cDNA libraries were created with the Iso-Seq Express protocol and sequenced on a PacBio Sequel II system. Detailed descriptions of materials and methods are available (Supplementary Methods).

Assembly

We produced fully phased hybrid assemblies using Verkko (v.1.4.1)[9] as our primary assembler (Supplementary Methods). We additionally created hifiasm (ultra-long; v.0.19.6)[10] assemblies (Supplementary Methods), which were used to complement our analysis of the most challenging regions (centromeres and Yq12). The phasing signal for all assemblies was generated using the Graphasing pipeline[12] (v.0.3.1-alpha). All assemblies were scanned for contamination with NCBI's Foreign Contamination Screening workflow (v.0.4.0)[71] and annotated for potential assembly errors using Flagger (v.0.3.3)[1], Merqury (v.1.0)[72], NucFreq[73] (commit #bd080aa) and Inspector (v.1.2)[74] (Supplementary Methods). Assembly quality was assessed by computing quality value estimates with Merqury and DeepVariant (v.1.6)[75] as previously described[8]. Gene completeness of the assemblies was evaluated using compleasm (v.0.2.5)[76] and the primate set of known single-copy genes of OrthoDB (v.10)[77]. The T2T status of the assembled chromosomes and the closing status of previously reported gaps[2] were determined relative to the T2T-CHM13 reference genome[4] by factoring in the above quality control information in the evaluation of the contig-to-reference alignment produced with minimap2 (v.2.26)[78,79] and mashmap (v.3.1.3)[80] (Supplementary Methods). The parental support for the assembled child haplotypes in the three family trios was computed by evaluating the CIGAR operations in the minimap2 contig-to-contig alignments between the parents and child.

Variant calling

Genome reference. Callsets were constructed against two references: GRCh38 (GRCh38-NoALT) and T2T-CHM13 (T2T-CHM13v2.0)[4].

Variant discovery and merging. For assembly-based callsets, we ran PAV (v.2.4.1)[8] with minimap2 (v.2.26)[78] and LRA (v.1.3.7.2)[81] alignments, DipCall (v.0.3)[82] and SVIM-asm (v.1.0.3)[83]. SVIM-asm used PAV alignments before PAV applied any alignment trimming, and DipCall produced minimap2 alignments for DipCall variants (Supplementary Methods).

For PacBio HiFi callsets, we ran PBSV (https://github.com/PacificBiosciences/pbsv; v.2.9.0), Sniffles (v.2.0.7)[84], Delly (v.1.1.6)[85], cuteSV (v.2.0.3)[86], DeBreak (v.1.0.2)[87], SVIM (v.2.0.0)[88], DeepVariant (v.1.5.0)[75] and Clair3 (v.1.0.4)[89]. The same callers and versions were run for ONT except for PBSV, and DeepVariant was executed through PEPPER-Margin-DeepVariant (r0.8)[90]. The callset process was the same for both references (Supplementary Methods).

SV-Pop[8] was used to merge PAV calls from minimap2 alignments and generate per-sample support information from all other callers. Calls in T2T-CHM13 were filtered if they intersected the UCSC 'CenSat' track for T2T-CHM13 (UCSC hs1) with monomeric ('mon') records excluded or if they were in telomere repeats. GRCh38 variants intersecting modelled centromeres were removed (Supplementary Methods).

MEIs

MEIs were identified within the 130 haplotype assemblies using two separate pipelines and human references (T2T-CHM13 and GRCh38). One detection pipeline, L1ME-AID (v.1.0.0-beta; L1 Mediated Annotation and Insertion Detector; see Code availability), leverages a local RepeatMasker (v.4.1.6)[91] installation with the Dfam (v.3.8) database[92] to annotate the freeze4 PAV-merged SV insertion callsets (T2T-CHM13 and GRCh38). The second pipeline called MEIs directly from the alignment of contigs to a reference genome with PALMER2 (Code availability). Putative MEIs from both callers were then merged using MEI coordinates, element family (*Alu*, L1, SVA, HERV-K or snRNA) and sequence composition (Supplementary Methods). Next, MEIs were curated to distinguish MEIs from deletions (T2T-CHM13 or GRCh38), duplications or potential artefacts (for example, possible genome assembly errors; Supplementary Methods). All MEIs called by a single pipeline that passed quality control were manually curated. Finally, both callsets were compared against an orthogonal MEI callset produced by MELT-LRA (Supplementary Methods; see Code availability). To determine intact ORFs across LINE-1 elements, we followed a previously described method[8] to detect intact ORF1p and ORF2p from full-length (more than 5,900 bp) LINE-1 insertions.

Separately, MEIs within centromere HOR arrays were identified with RepeatMasker (v.4.1.6)[91] and the Dfam library (v.3.8)[92], annotation of complete and accurately assembled centromeres (see 'Centromeres' in the Methods). The sequences of *Alu* elements, L1s and SVAs identified by RepeatMasker within the centromere HOR array boundaries were retrieved using SAMtools (v.1.15.1)[93]. Element sequences were then scrutinized with L1ME-AID (v.1.0.0-beta) utilizing the same cut-offs applied to the freeze4 PAV-merged SV insertion callset to distinguish young MEIs from older mobile element fragments. Sequence of all putative MEIs that passed filtering were re-retrieved along with a flanking sequence (±100 bp) using SAMtools (v.1.15.1)[94], and then aligned against one another using MUSCLE (v.3.38.31)[95] to distinguish unique MEIs from duplicated insertions of MEIs residing in centromere regions (Supplementary Table 58).

Inversions

We performed validation of the T2T-CHM13-based and GRCh38-based PAV inversion callsets, individually, using Strand-seq-based re-genotyping of the inversion calls. Before genotyping, we performed Strand-seq cell selection using ASHLEYS[96]. The good-quality Strand-seq cells were used as input to perform genotyping by ArbiGent[16] (Supplementary Methods).

We evaluated the PAV inversion callset for one candidate carrier per region using manual dotplot analysis with NAHRwhals[97]. NAHRwhals was applied to detect the false discovery rate and classify all candidate inversion regions larger than 5 kb into distinct inversion classes.

We compared the PAV inversion callset reported with respect to T2T-CHM13 to a previously published callset[98] based on a subset of

genomes reported in this study. Using the 25% reciprocal overlap criterion, we defined inversions detected in both callsets as well as inversions that are new to the current study. We evaluated all novel inversion candidates manually using dotplot analysis of each putative novel inversion.

SD and copy number polymorphic genes

Identification of SDs. SD annotation was performed using SEDEF (v.1.1)[99] after masking repeats (TRF (v.4.1.0)[100], RepeatMasker (v.4.1.5)[101] and Windowmasker (v.2.2.22)[102]; Supplementary Methods). SDs with a sequence identity of more than 90%, length of more than 1 kb, satellite content of less than 70% and free of putative erroneous regions (see Code availability) were retained. In addition, the highly confident SD callset was further validated by fastCN[103]. Comparative analysis of SDs was conducted in T2T-CHM13 space. Positions of the SDs in T2T-CHM13 were mapped as follows: (1) linking SDs within 10-kb distance, (2) identifying those SD chains that are located in alignment block of at least 100 kb in size, and (3) projecting the chained SDs onto putative homologous SD loci containing at least one 10-kb unique flank. In addition, syntenic SDs were further assessed for whether they share sequence content by aligning SDs with minimap2 (v.2.26)[78]; the following SDs were quantified: (1) SDs unobserved by T2T-CHM13, (2) having changed sequence content (less than 80% of the sequence conserved), and (3) expanded size (at least twofold).

Duplicated genes. Protein-coding transcripts from GENCODE v.44 (Liftoff to T2T-CHM13) were aligned to the genome assemblies (excluding NA19650, NA19434 and NA21487) using minimap2 ('-cx asm20 -f 5000 -k15 -w10 -p 0.05 -N 200 -m200 -s200 -z10000 --secondary=yes --eqx'). The mapped genes were further filtered to exclude alignments due to nested repeats, keeping minimum length of 2 kb, percent identity of more than 90% and coverage of more than 80%. Multi-copy genes were determined by maximum gene counts greater than one. Variable copy number genes were defined by assessing the copy number across the population (at least one of the genome assemblies with different copy number; Supplementary Methods).

Y chromosome variation

Construction and dating of Y phylogeny. The construction and dating of Y-chromosomal phylogeny combining the 30 males from the current study plus two males (HG01106 and HG01952 from the HPRC year 1 dataset for which contiguous Yq12 assemblies were used from[23]) were done as previously described[23]. Detailed descriptions of methods are available (Supplementary Methods). Please note that the male individual HG03456 appears to have a XYY karyotype as previously reported[29].

Identification of sex-chromosome contigs. Contigs containing Y-chromosomal sequences from the whole-genome assemblies were identified and extracted for the 30 males as previously described[23]. Y assemblies for the two HPRC individuals, HG01106 and HG01952, were used from ref. 23.

Y chromosome annotation and analysis. The annotation of Y-chromosomal subregions was performed as previously described using both the GRCh38 and T2T-CHM13 Y reference sequences[23]. The centromeric α-satellite repeats for the purpose of Y subregion annotation were identified using RepeatMasker (v.4.1.2-p1)[91]. The Yq12 repeat annotations were generated using HMMER (v.3.3.2dev)[104], and identification of *Alu* insertions was performed as previously described[23]. To maximize the number of contiguously assembled Yq12 subregions, hifiasm assemblies of this subregion were analysed from four individuals (NA19239, HG03065, NA19347 and HG00358) following manual inspection of repeat unit orientation and distance from each other in the assembled sequences (Supplementary Table 40).

Dotplots to compare Y-chromosomal sequences were generated using Gepard (v.2.0)[105]. Although we also assembled the T2T (NA24385/HG002) Y as a single contig (Supplementary Table 40), all analyses conducted here used the existing published T2T assembly[24].

Visualization of eight completely assembled Y chromosomes (Supplementary Fig. 27) was based on pairwise alignments generated using minimap2 (v.2.26)[78,79] with the following options: '-x asm20 -c -p 0.95 --cap-kalloc = 1 g -K4g -I8g -L --MD --eqx'. For visualization, alignments of less than 10 kb in length were filtered out. In addition, alignments were broken at SVs of more than 50 bp or more in size and then binned in 50-kb bins.

SVs affecting genes

We annotated the potential effect of long-read SVs on genes using the coding transcripts and exons defined in GENCODE (v.45)[106], as per Ensembl VEP (v.111)[107]. Long-read deletions or insertions are classified as coding overlapping events if at least one breakpoint falls within the coding exons of a gene. We considered genes that have a LOEUF score under 0.35 as intolerant to loss-of-function variants[27]. To specifically analyse the potential effect of MEIs on genes, the merged GRCh38 MEI callset was intersected with the findings from Ensembl[108] (release 111) VEP[107] (see transcriptional effect of SVs below). The MEIs were categorized by insertion location (for example, protein-coding exons, untranslated regions of protein-coding transcripts and non-coding exons), and within each category, the number of MEIs present, genes disrupted and transcripts affected were quantified. The Ensembl VEP nonsense-mediated decay (NMD) plugin (https://github.com/Ensembl/VEP_plugins/blob/release/112/NMD.pm) was utilized to predict which protein-coding transcripts with MEI-induced premature stop codons would escape NMD. Transcripts were further scrutinized by manually comparing the MEI location within the transcript sequence using the UCSC Genome Browser[109]. To ensure that the premature stop codon met one of the four requirements for NMD escape according to the exon-junction complex model[110]. Allele frequencies were then calculated (children of trios excluded) for the exon-disrupting MEIs. In the event of a '.' (indicating misassembly) in the genotyping information, the haplotype was excluded from the calculation.

Functional effect of SVs

Effects on exons and isoform. We used the Ensembl[108] (release 111) Variant Effect Predictor[107] with the NMD plugin (https://github.com/Ensembl/VEP_plugins/blob/release/112/NMD.pm) to screen the PAV freeze 4 callset for SVs that disrupt gene loci in the merged GRCh38 annotation (NCBI RefSeq GCF_000001405.40-RS_2023_03, Ensembl 111, GENCODE v.45). Protein-coding genes impacted by putative exon disruptions were evaluated for evidence of Iso-Seq expression (in more than 1 individual) across the 12 individuals. Isoforms associated with these SV-containing genes were screened for the presence of unreported splice variants using SQANTI3 (v.5.1.2)[111]. All isoforms of these candidate genes were aligned to GRCh38p14 using pbmm2 (https://github.com/PacificBiosciences/pbmm2; v.1.5.0) and visualized with IGV[112] to identify variant-specific patterns. We compared all isoforms phased to variant haplotypes to known transcripts represented in RefSeq[113], CHESS[114] and GENCODE[106] gene annotation databases to identify novel splice products and isoforms. MUSCLE (v.3.8.425)[95] and Aliview[115] were used to perform a multiple sequence alignment and visualize the multiple sequence alignment, respectively, between wild-type and variant haplotype assemblies to identify SV breakpoints.

Effects on gene expression. We next assessed SVs for enrichment near genes with altered expression in the 12 individuals with Iso-Seq data. Using gene expression quantifications from short-read RNA-seq data, we performed differential expression analysis using DESeq2 (v.1.38.3)[116] between individuals who carried and did not carry each SV, supplemented with outlier expression analysis for singleton SVs

(Supplementary Methods). We assessed SV overlap with multiple GENCODE v.45-derived genomic elements, such as protein-coding and pseudogene classes[106], and ENCODE-derived candidate *cis*-regulatory elements[117], using permutation tests to find enrichment or depletion of SVs for each annotation (Supplementary Methods).

Effects on chromatin structure and colocalization with GWAS hits. 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 first filtered out SVs with missing genotypes in 6 or more out of 12 individuals. For each remaining SV, we extracted the 50 kb upstream and downstream of the annotated transcription start site position for each paired gene with corresponding insulation scores under 10-kb resolution (Supplementary Methods). For those insulated regions intersecting with more than one SV, we applied a local multi-test correction. A false discovery rate < 0.05 from the two-sided Wilcoxon rank-sum test was considered significant. We investigated the association between variants and human phenotypes or traits by intersecting SNVs, indels and SVs with SNPs identified in GWAS (GWAS summary statistics; gwas_catalog_v1.0.2-associations_e111_r2024-04-16.tsv)[118]. We used Plink (v.1.90b6.10)[119] to examine the linkage disequilibrium between SNVs, indels and SVs with GWAS SNPs within 1-Mb window size.

Genome-wide genotyping with PanGenie

We built a pangenome graph containing 214 haplotypes with Minigraph-Cactus (v.2.7.2)[120] from the haplotype-resolved assemblies of 65 HGSVC individuals and 42 individuals from the HPRC[1] and produced a CHM13-based VCF representation of the top-level bubbles of the graph that can be used as input for genotyping with PanGenie (Supplementary Methods). This was done by converted genotypes of male sex chromosomes to a homozygous representation, filtering out records for which at least 20% of haplotypes carry a missing allele ('.') and running our previously developed decomposition approach to detect and annotate variant alleles nested inside of graph bubbles (Supplementary Methods). We genotyped all 30,490,169 bubbles (representing 28,343,728 SNPs, 10,421,787 indels and 547,663 SVs) across all 3,202 1kGP individuals based on short reads[29] using PanGenie (v.3.1.0)[3] with additional parameter -a 108. We filtered the resulting genotypes based on a support vector regression approach[1,8], resulting in 25,695,951 SNPs, 5,774,201 indels and 478,587 SVs that are reliably genotypable (Supplementary Methods).

Personal genome reconstruction

Reference panel and personal genome construction. We used our filtered genotypes across all 3,202 individuals and added 70,174,243 additional rare SNPs and indels from an external short-read-based callset for the same 3,202 1kGP individuals (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/; Supplementary Methods). We filtered out variants reported with a genotype quality below 10 and ran SHAPEIT5 (v.5.1.1)[121] phase_common to phase this joint callset. We used the resulting reference panel to reconstruct personal genomes for all 3,202 individuals by implanting phased variants into the CHM13 reference genome with BCFtools[93] to create the 6,404 consensus haplotype sequences of all 1kGP individuals (Supplementary Methods).

Evaluation. For the evaluation of the consensus haplotypes produced from 1kGP and PG-SHAPEIT phased genotypes, PAV was run with one of the consensus haplotypes as a reference and the other one as a query sequence, together with the respective haplotype assemblies for the same individual. We analysed the resulting variant calls to determine all variant positions with conflicting genotypes between the consensus and assembly haplotypes. For such erroneous variant positions, we then counted the number of base-pair changes in both consensus haplotypes within windows of 1 Mb in length along the reference haplotype and computed a quality value estimate as: $-10 \times \log_{10}(\text{bp\_changes}/(2 \times \text{window\_size}))$. In addition, we also counted the number of erroneous variants more than 20 bp in each window. We then plotted the distributions of these two metrics and computed the median (Supplementary Figs. 43–45b,c). We evaluated consensus sequences for a second individual (HG01114) to verify consistency of results across individuals. For each individual, we ran the experiment twice, using either haplotype as a reference sequence. In addition to evaluating the consensus haplotypes, we repeated the same experiment for HG002, using the Q100 assemblies (https://github.com/marbl/HG002) as reference sequences to align to, and our HGSVC3 assemblies as queries (Supplementary Figs. 44 and 45d). To get a baseline estimate, we also ran the experiment using CHM13 as a reference and two copies of GRCh38 as well as another copy of CHM13 as query sequences (Supplementary Fig. 45a).

Targeted genotyping of complex polymorphic loci

Targeted genotyping was performed using Locityper (v.0.15.1)[30] across 347 complex polymorphic target loci (Supplementary Methods). On the basis of the input short-read whole-genome sequencing data, at each of the targets, Locityper aims to identify two haplotypes from the reference panel that are most similar to the input data. Three reference panels were used: HPRC haplotypes (90 haplotypes); HPRC + HGSVC3 haplotypes (216 haplotypes); and leave-one-out HPRC + HGSVC3 panel (leave-one-out evaluation; 214 haplotypes), where two assemblies corresponding to the input dataset were removed. To evaluate prediction accuracy, we constructed sequence alignments between actual and predicted haplotypes and estimated variant-based quality values (Supplementary Methods). Locityper accuracy is limited by haplotypes present in the reference panel; consequently, we evaluated haplotype availability quality value as the highest Phred-scaled sequence divergence between actual assembled haplotypes and any haplotype from the reference panel (Supplementary Methods).

MHC

Gene annotation. Immuannot-based HLA types were compared in two-field resolution to the HLA typing published earlier and obtained with PolyPheMe[35] (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HLA_types/20181129_HLA_types_full_1000_Genomes_Project_panel.txt). Of the 130 haplotypes, 58 were not in the PolyPheMe dataset and were excluded. In addition to Immuannot (MHC reference version: IPD-IMGT/HLA-v.3.55.0)[36], haplotypes were annotated using MHC-annotation v.0.1 (see Code availability). Cases of overlapping genes were resolved after inspection by removing superfluous annotations. Reported gene counts for HLA genes and C4 annotation were based on Immuannot.

SV detection. To search for structural variation in the DRB gene region, HGSVC MHC haplotypes were cut from (start of DRA) to (end of DRB1 + 20 kb). The coordinates were obtained using MHC-annotation v.0.1. On the basis of their DRB1 allele as determined by Immuannot (see above), the sequences were grouped into DR groups. Within each group, every sequence was aligned with nucmer[122] (v.3.1; -nosimplify -maxmatch) to the same sequence (arbitrarily selected as the sequence with the alphanumerically smallest ID) and plotted with a custom gnuplot script based on mummerplots output. Sequences were annotated as follows: (1) repeat elements were masked with RepeatMasker (v.4.1.2)[91]; (2) full DRB genes and pseudogenes were searched for with minimap (v.2.26; '--secondary=no -c -x --asm10 -s100') by aligning the sequence from against all DRB alleles from IMGT and the larger *DRB9* sequence Z80362.1 (results were highlighted and masked for the next step); (3) DRB exons were searched for with BLASTN (v.2.14.1)[123] by aligning all DRB exons from IMGT to the sequence and filtering for highest

matches (results were highlighted and masked for the next step); and (4) as for step 3 but with introns.

For each HGSVC MHC haplotype, SVs were called with PAV[8] against eight completely resolved MHC reference haplotypes[39,40]. To determine which SVs in the HGSVC haplotypes were not present in any of the eight reference haplotypes, for each HGSVC haplotype, the 'query' coordinates (that is, the coordinates of the calls relative to the analysed HGSVC haplotype) of the PAV calls were padded with 50 bp on each side and the intersection of SV calls (based on the padded query coordinates, across the eight MHC reference sequences) was computed. Only variants longer than 50 bp were included for further analysis, and the smallest variant relative to any of the eight MHC references was reported. The sequences of the calls so-defined were annotated with RepeatMasker (v.4.1.2)[91]. Variants were grouped by starting position on their closest MHC reference sequence and, in the case of insertions, repeat content was averaged.

Annotation. We applied Immuannot (see above) to all retrieved MHC loci for the identification and annotation of protein-coding HLA-DRB genes (*HLA-DRB1*, *HLA-DRB3*, *HLA-DRB4* and *HLA-DRB5*). Subsequently, a custom RepeatMasker (v.4.1.2)[91] library was constructed containing the exonic sequences of HLA-DRB pseudogenes (*HLA-DRB2*: ENSG00000227442.1, *HLA-DRB6*: ENSG00000229391.8, *HLA-DRB7*: ENSG00000227099.1, *HLA-DRB8*: ENSG00000233697.2, and *HLA-DRB9*: ENSG00000196301.3) and RCCX genes and pseudogenes (*C4*: ENSG00000244731.10, *CYP21A2*: ENSG00000231852.9, *CYP21A1P*: ENSG00000204338.9, *STK19*/*RP1*: ENSG00000204344.16, *STK19B*/*STK19P*/*RP2*: ENSG00000250535.1, *TNXB*: ENSG00000168477.21 and *TNXA*: ENSG00000248290.1). Canonical exonic sequences were sourced from the Ensembl genome browser[108] (release 111). The exons of HLA-DRB or RCCX genes and pseudogenes within individual haplotype MHC regions were annotated using this custom library. Repetitive elements were identified using RepeatMasker (v.4.1.2) with the Dfam library (v.3.4)[92]. We utilized SAMtools (v.1.15.1)[93] and MUSCLE (v.3.8.31)[95] for sequence retrieval and alignment, respectively, followed by manual annotation to analyse recombination events associated with DR subregion haplotypes and within the RCCX modules (Supplementary Methods). Novel *C4*-coding variants were identified through comparison with Ensembl *C4A* and *C4B* protein variant tables, as well as an additional database of variants obtained from 95 human MHC haplotypes[33].

Complex structural polymorphisms

CSV detection. CSVs were identified with a development version of PAV (methods are available[124]). In brief, the method identifies candidate variant anchors and scores variants between them. A directed acyclic graph is constructed with alignment records as nodes and variants connecting them as edges, which is solved in $O(N + E)$ time with the Bellman–Ford algorithm[125]. Variants on the optimal path were accepted into the callset. CSVs intersecting centromeric repeats were eliminated. CSVs were merged into a non-redundant callset with SV-Pop by 50% reciprocal overlap and 80% sequence identity (SV-Pop merge parameter 'nr::ro(0.5):match(0.8)').

SMN analysis. We evaluated complexity and copy number of SMN genes by extracting with FASTA the desired region (chr. 5: 70300000–72100000) from assemblies reported in this study along with previously published assemblies[1,126] (Supplementary Methods). Among these, we identified 101 fully assembled haplotypes. We followed this by aligning exon sequences for multicopy genes (*SMN1/2*, *SERF1A/B*, *NAIP* and *GTF2H2/C*) to each assembled haplotype. To assign a specific SMN copy to each haplotype, we extracted FASTA sequence from SMN exon regions for each haplotype and concatenated them into a single sequence (Supplementary Methods). We then constructed a multiple sequence alignment and calculated the distance among all haplotypes. We set the orangutan sequence as an outgroup and split all human haplotypes into two groups representing *SMN1* and *SMN2* gene copies where the SMN1 copy is the one closer to the outgroup. We utilized Illumina short-read data from the 1kGP for the same individuals, and processed it with Parascopy (v.1.16.0)[127] and SMNCopyNumberCaller (v.1.1.2)[128] to independently obtain *SMN1/2* copy numbers. Illumina-based and assembly-based copy number predictions matched perfectly across all 31 examined individuals.

Centromeres

Centromere identification and annotation. To identify the centromeric regions within each Verkko and hifiasm (ultra-long) genome assembly, we first aligned the whole-genome assemblies to the T2T-CHM13 (v.2.0) reference genome[4] using minimap2 (v.2.24)[78] with the following parameters: -ax asm20 --secondary=no -s 25000 -K 15 G --eqx --cs. We filtered the alignments to only those contigs that traversed each human centromere, from the p to the q arm, using BEDtools (v.2.29.0)[129] intersect. Then, we ran dna-brnn (v.0.1)[130] on each centromeric contig to identify regions containing α-satellite sequences, as indicated by a '2'. Once we identified the regions containing α-satellite sequences, we ran RepeatMasker (v.4.1.0)[91] to identify all repeat elements and their organization within the centromeric region. We also ran HumAS-HMMER (https://github.com/fedorrik/HumAS-HMMER_for_AnVIL) with the AS-HORs-hmmer3.0-170921.hmm model, which was trained on GRCh38 as previously described[58], to determine the α-satellite HOR sequence composition and organization. We used the resulting RepeatMasker and HumAS-AMMER stv_row.bed files to visualize the organization of the α-satellite HOR arrays with R (v.1.1.383)[131] and the ggplot2 package[132].

Validation of centromeric regions. We validated the assembly of each centromeric region by first aligning native PacBio HiFi and ONT data from the same genome to each relevant whole-genome assembly using pbmm2 (v.1.1.0; for PacBio HiFi data; https://github.com/PacificBiosciences/pbmm2) or minimap2 (v.2.28)[78] (for ONT data). We then assessed the alignments for uniform read depth across the centromeric regions via IGV[112] and NucFreq[73]. Centromeres that were found to have a collapse in sequence, false duplication of sequence and/or misjoin were flagged and removed from our analyses.

Estimation of α-satellite HOR array length. To estimate the length of the α-satellite HOR arrays for each human centromere, we first ran HumAS-HMMER (https://github.com/fedorrik/HumAS-HMMER_for_AnVIL) on the centromeric regions using the hmmer-run.sh script and the AS-HORs-hmmer3.0-170921.hmm hidden Markov model. Then, we used the stv_row.bed file to calculate the length of the α-satellite HOR arrays by taking the minimum and maximum coordinate of the 'live' α-satellite HOR arrays, marked by an 'L', and plotting their lengths with GraphPad Prism (v9). We note that live or 'active' α-satellite HOR arrays are those that belong to an array that associates with the kinetochore in several individuals[58,133]. By contrast, 'dead' or 'inactive' α-satellite HORs (denoted with a 'd' in the HumAS-HMMER BED file) are those that have not been found to be associated with the kinetochore and are usually more divergent in sequence than the live or active arrays.

Pairwise sequence identity heatmaps. To generate pairwise sequence identity heatmaps of each centromeric region, we ran StainedGlass (v.6.7.0)[134] with the following parameters: window = 5,000, mm_f = 30,000 and mm_s = 1,000. We normalized the colour scale across the StainedGlass plots by binning the percent sequence identities equally and recolouring the data points according to the binning.

CpG methylation analysis. To determine the CpG methylation status of each centromere, we aligned ONT reads of more than 30 kb in length from the same source genome to the relevant whole-genome assembly via minimap2 (v.2.28)[78] and then assessed the CpG methylation status

of the centromeric regions with Epi2me modbam2bed (https://github.com/epi2me-labs/modbam2bed; v.0.10.0) and the following parameters: -e -m 5mC --cpg. We converted the resulting BED file to a bigWig using the bedGraphToBigWig tool (https://www.encodeproject.org/software/bedgraphtobigwig/) and then visualized the file in IGV. To determine the length of the hypomethylated region (termed CDR[58,60]) in each centromere, we used CDR-Finder[135]. This tool first binned the assembly into 5-kb windows, computed the median CpG methylation frequency within windows containing α-satellite (as determined by RepeatMasker (v.4.1.0)[91]), selected bins that have a lower CpG methylation frequency than the median frequency in the region, merged consecutive bins into a larger bin, filtered for merged bins that are more than 50 kb and reported the location of these bins.

Reporting summary

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

Data availability

All data produced by the HGSVC and analysed as part of this study are available under the following accessions (see Supplementary Tables 2–4, 6–8 and 23 for details): PRJEB58376, PRJEB75216, PRJEB77558, PRJEB75190, PRJNA698480, PRJEB75739, PRJEB36100, PRJNA988114, PRJNA339722, PRJEB41778 and ERP159775 for PacBio HiFi and ONT LRS data; PRJEB39750 and PRJEB12849 for Strand-seq; PRJNA339722, PRJEB41077, PRJEB58376 and PRJEB77842 for Bionano Genomics; PRJEB39684, PRJEB75193 and PRJEB58376 for Hi-C; PRJEB75191 for PacBio Iso-Seq; PRJEB75192 and PRJEB58376 for RNA-seq; PRJEB76276 for phased genome assemblies generated by Verkko; and PRJEB83624 for phased genome assemblies generated by hifiasm. Released resources, including simple and complex variant calls, genome graphs, genotyping results (genome-wide and targeted), and annotations for centromeres, MEIs and SDs can be found in the IGSR release directory hosted publicly via HTTP and/or FTP (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/release) and on the Globus end point 'EMBL-EBI Public Data' in the directory '/1000g/ftp/data_collections/HGSVC3/release'.

Code availability

All software, scripts and workflows used in this project that have not been formally published are publicly available via a central GitHub repository (https://github.com/hgsvc/phase3-main-pub; section 'Software') and on Zenodo[136] (https://doi.org/10.5281/zenodo.14546729).

67. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
68. Zook, J. M. et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 160025 (2016).
69. Sanders, A. D., Falconer, E., Hills, M., Spierings, D. C. J. & Lansdorp, P. M. Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nat. Protoc.* **12**, 1151–1176 (2017).
70. Falconer, E. et al. DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat. Methods* **9**, 1107–1112 (2012).
71. Astashyn, A. et al. Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol.* **25**, 60 (2024).
72. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
73. Vollger, M. R. et al. Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019).
74. Chen, Y., Zhang, Y., Wang, A. Y., Gao, M. & Chong, Z. Accurate long-read de novo assembly evaluation with Inspector. *Genome Biol.* **22**, 312 (2021).
75. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
76. Huang, N. & Li, H. compleasm: A faster and more accurate reimplementation of BUSCO. *Bioinformatics* **39**, btad595 (2023).
77. Kriventseva, E. V. et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **47**, D807–D811 (2019).
78. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
79. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
80. Jain, C., Koren, S., Dilthey, A., Phillippy, A. M. & Aluru, S. A fast adaptive algorithm for computing whole-genome homology maps. *Bioinformatics* **34**, i748–i756 (2018).
81. Ren, J. & Chaisson, M. J. P. lra: a long read aligner for sequences and contigs. *PLoS Comput Biol.* **17**, e1009078 (2021).
82. Li, H. et al. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat. Methods* **15**, 595–597 (2018).
83. Heller, D. & Vingron, M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* **36**, 5519–5521 (2021).
84. Smolka, M. et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat. Biotechnol.* **42**, 1571–1580 (2024).
85. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
86. Jiang, T. et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189 (2020).
87. Chen, Y. et al. Deciphering the exact breakpoints of structural variations using long sequencing reads with DeBreak. *Nat. Commun.* **14**, 283 (2023).
88. Heller, D. & Vingron, M. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35**, 2907–2915 (2019).
89. Zheng, Z. et al. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nat. Comput. Sci.* **2**, 797–803 (2022).
90. Shafin, K. et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat. Methods* **18**, 1322–1332 (2021).
91. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0 (Institute for Systems Biology, 2013).
92. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* **12**, 2 (2021).
93. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
94. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
95. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
96. Gros, C., Sanders, A. D., Korbel, J. O., Marschall, T. & Ebert, P. ASHLEYS: automated quality control for single-cell Strand-seq data. *Bioinformatics* **37**, 3356–3357 (2021).
97. Höps, W. et al. Impact and characterization of serial structural variations across humans and great apes. *Nat. Commun.* **15**, 8007 (2024).
98. Porubsky, D. et al. Inversion polymorphism in a complete human genome assembly. *Genome Biol.* **24**, 100 (2023).
99. Numanagic, I. et al. Fast characterization of segmental duplications in genome assemblies. *Bioinformatics* **34**, i706–i714 (2018).
100. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
101. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* https://doi.org/10.1002/0471250953.bi0410s25 (2009).
102. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics* **22**, 134–141 (2006).
103. Pendleton, A. L. et al. Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication. *BMC Biol.* **16**, 64 (2018).
104. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121 (2013).
105. Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026–1028 (2007).
106. Frankish, A. et al. GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res.* **51**, D942–D949 (2023).
107. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
108. Martin, F. J. et al. Ensembl 2023. *Nucleic Acids Res.* **51**, D933–D941 (2023).
109. Lee, B. T. et al. The UCSC Genome Browser database: 2022 update. *Nucleic Acids Res.* **50**, D1115–D1122 (2022).
110. Lindeboom, R. G. H., Supek, F. & Lehner, B. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat. Genet.* **48**, 1112–1118 (2016).
111. Pardo-Palacios, F. J. et al. SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nat. Methods* **21**, 793–797 (2024).
112. Robinson, J. T. et al. Integrative Genomics Viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
113. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
114. Pertea, M. et al. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.* **19**, 208 (2018).
115. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278 (2014).
116. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
117. ENCODE Project Consortiumet al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
118. Sollis, E. et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
119. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
120. Hickey, G. et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat. Biotechnol.* **42**, 663–673 (2024).

121. Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S. & Delaneau, O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nat. Genet.* **55**, 1243–1249 (2023).
122. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
123. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
124. Audano, P., Christine, B. & Human Genome Structural Variation Consortium. A method for calling complex SVs. *Zenodo* https://doi.org/10.5281/zenodo.13800981 (2024).
125. Bellman, R. On a routing problem. *Quart. Appl. Math.* **16**, 87–90 (1958).
126. Yoo, D. et al. Complete sequencing of ape genomes. *Nature* **641**, 401–418 (2025).
127. Prodanov, T. & Bansal, V. Robust and accurate estimation of paralog-specific copy number for duplicated genes using whole-genome sequencing. *Nat. Commun.* **13**, 3221 (2022).
128. Chen, X. et al. Spinal muscular atrophy diagnosis and carrier screening from genome sequencing data. *Genet. Med.* **22**, 945–953 (2020).
129. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
130. Li, H. Identifying centromeric satellites with dna-brnn. *Bioinformatics* **35**, 4408–4410 (2019).
131. R Core Team. *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/ (R Foundation for Statistical Computing, 2020).
132. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis* (Springer, 2009).
133. McNulty, S. M. & Sullivan, B. A. Alpha satellite DNA biology: finding function in the recesses of the genome. *Chromosome Res.* **26**, 115–138 (2018).
134. Vollger, M. R., Kerpedjiev, P., Phillippy, A. M. & Eichler, E. E. StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps. *Bioinformatics* **38**, 2049–2051 (2022).
135. Mastrorosa, F. K. et al. Identification and annotation of centromeric hypomethylated regions with CDR-Finder. *Bioinformatics* **40**, btae733 (2024).
136. Ebert, P. hgsvc/phase3-main-pub: v1.1 HGSVC phase 3 revision stage/ZENODO (v1.1). *Zenodo* https://doi.org/10.5281/zenodo.14546729 (2024).

Acknowledgements Funding was provided by the US National Institutes of Health (NIH) grants U24HG007497 (to S.E.H., O.A.-O., L.S., E.E.E., S.E.D., B.G., T.M. and C. Lee), R00GM147352 (to G.A.L.), R01HG002385 and R01HG010169 (to E.E.E.), R01HG011649 (to M.J.P.C. and B.G.), K99HG012798 (to H.C.), U01HG013748 (to B.P., H.L. and T.M.); NIH National Institute of General Medical Sciences R35GM133600 (to P.A.A., P.B. and C.R.B.), 1P20GM139769 (to M.K.K. and M.L.), 1R35GM138212 (to Z.C.); NIH National Institute of Allergy and Infectious Disease U01AI090905 (to A.T.D., T.P., L.A.G. and P.J.N.); NIH National Cancer Institute R01CA261934 and R21CA259309 (to J.C. and S.E.D.), and P30CA034196 (to P.A.A. and C.R.B.); National Science Foundation CAREER 2046753 (to M.J.P.C. and K.R.); the Ministry of Culture and Science of North Rhine-Westphalia (MODS, 'Profilbildung 2020' (grant no. PROFILNRW-2020–107-A); to A. Söylev and T.M.); and the German Research Foundation grant 496874193 (to T.M.). This work was also supported, in part, by the Intramural Research Program of the National Human Genome Research Institute, NIH (to A.M.P. and S.K.), the Jürgen Manchot Foundation (to S.S. and A.T.D.) and the Düsseldorf School of Oncology (grant SPATIAL to T.M.). E.E.E. is an investigator of the Howard Hughes Medical Institute. We thank the Centre for Information and Media Technology and the Research IT Department of the Medical Faculty at Heinrich Heine University Düsseldorf for providing computational infrastructure and support; the staff at Clemson University for their allotment of compute time on the Palmetto HPC; HPC resources at Temple University supported in part by the National Science Foundation through major research instrumentation grant number 1625061 and by the US Army Research Laboratory under contract number W911NF-16-2-0189; the staff at the Scientific Services at The Jackson Laboratory, including the Genome Technologies Service for their assistance with the work described herein; the members of the HPRC (https://humanpangenome.org) for making their data publicly available; and the people who contributed samples as part of the 1kGP.

Author contributions P. Hallast, K.M.M., T.R., A. Sulovari, C. Lee and E.E.E. selected samples. K.M.M., P. Hallast, P. Hasenfeld, K.H., Q.Z. and S.E.D. produced the data. P.E., P.A.A., S.E.H., P. Hallast, F.Y., K.M.M., Y.K., O.A.-O. and L.S. managed data. P.E., W.T.H., M.H., Z.C., M.R., S.K., Y.K., H.C., A.M.P., Y.S., E.E.E. and T.M. produced the assembly and undertook quality control. P.A.A., C.A.P. and C.R.B. discovered variants. M.L., W.Z., P.B., R.E.M., J.C., S.E.D., C.R.B. and M.K.K. contributed to the mobile elements. H.A., V.T., D.P., T.R., J.O.K. and T.M. contributed to inversions. D.Y., K.R., M.J.P.C. and E.E.E. contributed to SDs. B.G. and M.J.P.C. provided STR and VNTR annotation. P. Hallast, P.E., M.L., M.K.K. and C. Lee contibuted to Y chromosome analysis. G.V.M., M.L. and M.K.K. conducted Iso-Seq phasing. X.Z., G.V.M., M.L., M.E.T. and M.K.K. assessed the SV effect on genes. G.V.M., M.L., M.J., Y.J., J.L., M.G. and M.K.K. analysed the transcriptional effects of SVs. C. Li, M.J.B. and X.S. analysed Hi-C and additional function. J.E., T.P., G.H., B.P. and T.M. conducted genotyping. J.E., T.R., M.C.Z. and T.M. contributed to the integrated reference panel. M.L., S.S., C.-S.C., Y.Z., N.R.P., P.J.N., L.A.G., P.A.A., P.E., A. Söylev, T.P., C.R.B., H.L., T.M., M.K.K. and A.T.D. contributed to the MHC. P.A.A., D.P., F.Y., M.L., M.K.K., C.R.B., C. Lee and E.E.E. analysed the complex structural polymorphisms. G.A.L., K.K.O., M.L., M.K.K. and E.E.E. contributed to analysis of the centromeres. G.A.L., P.E., P.A.A., M.L., D.P., J.E., F.Y., P. Hallast, T.P., D.Y., X.Z., G.V.M., C.-S.C., H.A., M.J., C. Li, X.S., M.E.T., M.J.P.C., A.T.D., M.K.K., J.O.K., C. Lee, C.R.B., E.E.E. and T.M. wrote the manuscript. All authors read and approved the final manuscript. J.O.K., C. Lee, E.E.E. and T.M. are HGSVC co-chairs.

Funding Open access funding provided by Heinrich-Heine-Universität Düsseldorf.

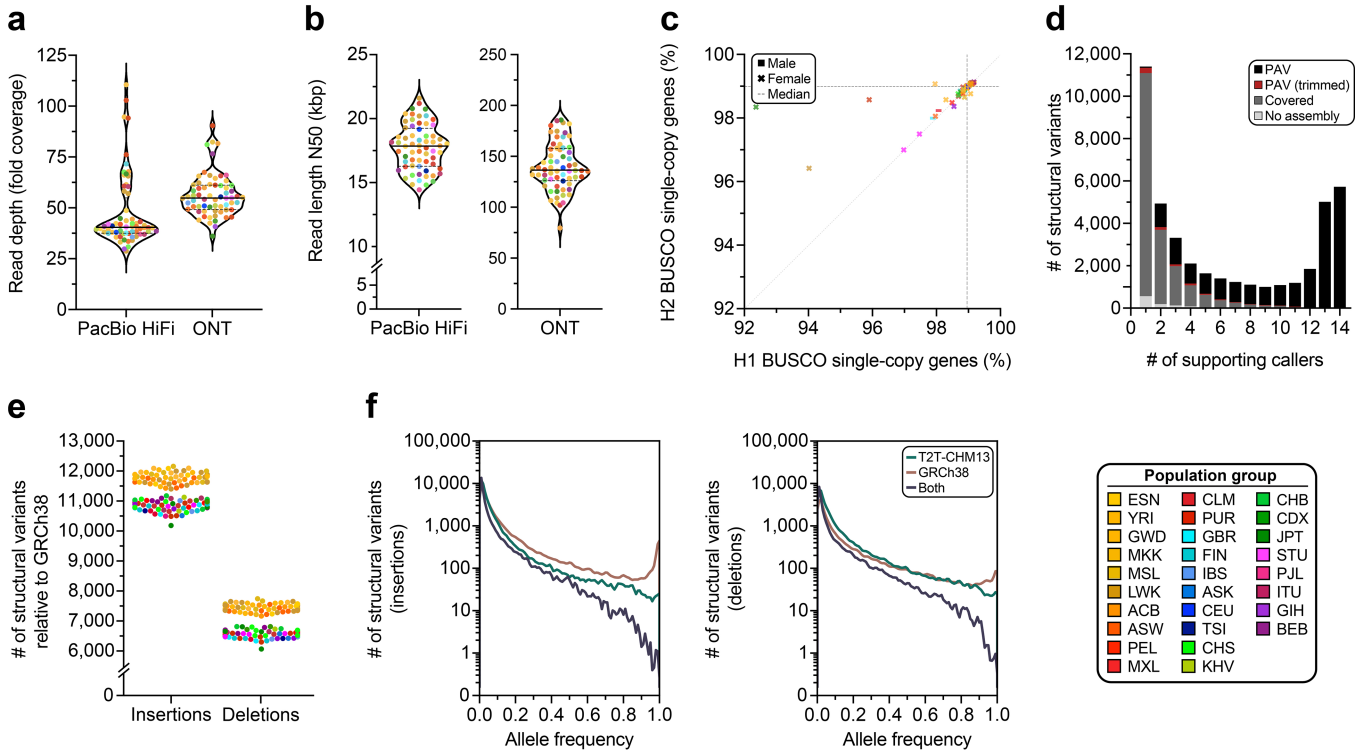
Competing interests E.E.E. is a scientific advisory board member of Variant Bio. C. Lee is a scientific advisory board member of Nabsys. S.K. has received travel funds to speak at events hosted by ONT. J.O.K., T.M. and D.P. have previously disclosed a patent application (no. EP19169090) relevant to Strand-seq. The other authors declare no competing interests.

Additional information

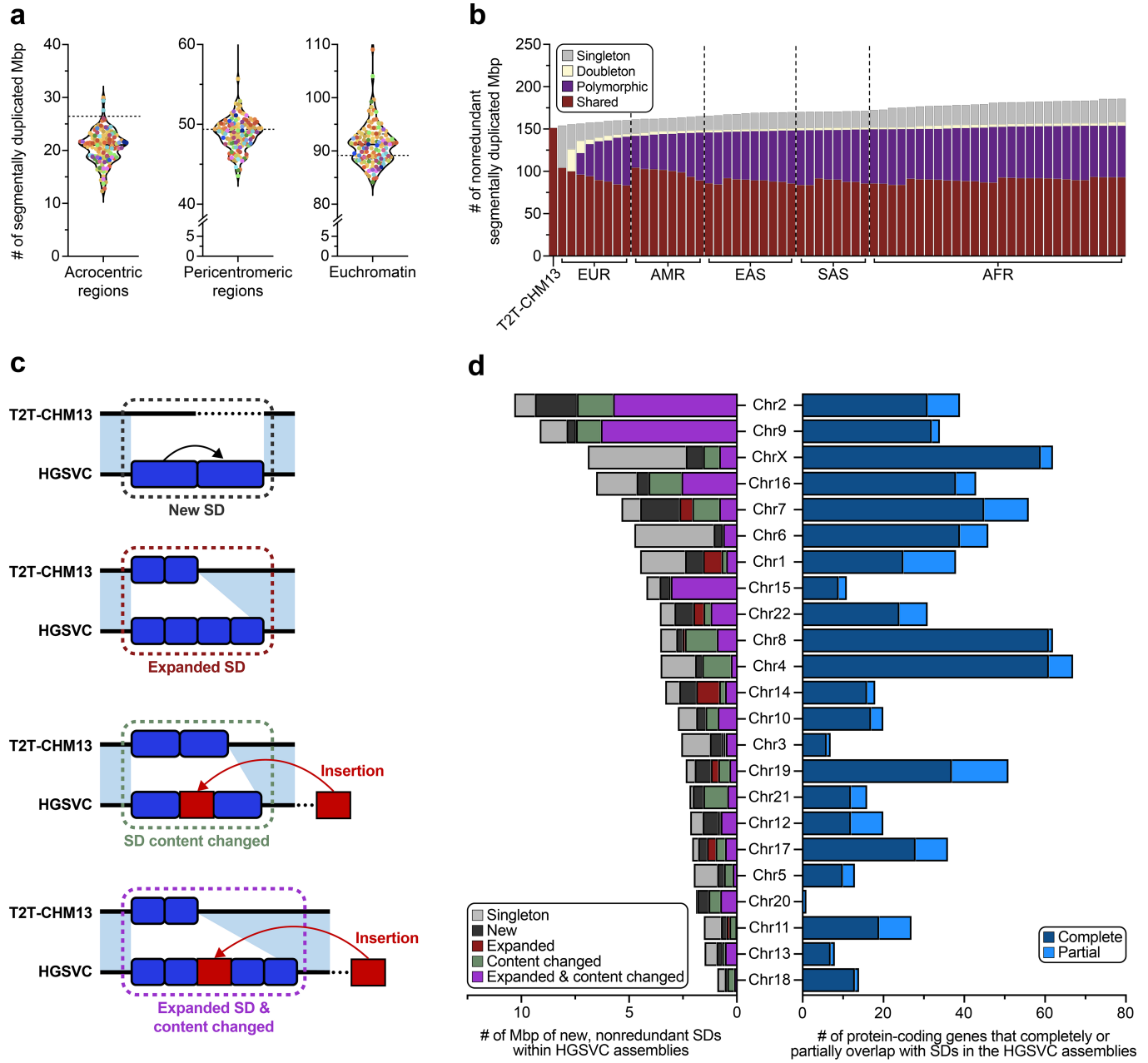
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-025-09140-6.

Correspondence and requests for materials should be addressed to Miriam K. Konkel, Jan O. Korbel, Charles Lee, Christine R. Beck, Evan E. Eichler or Tobias Marschall.

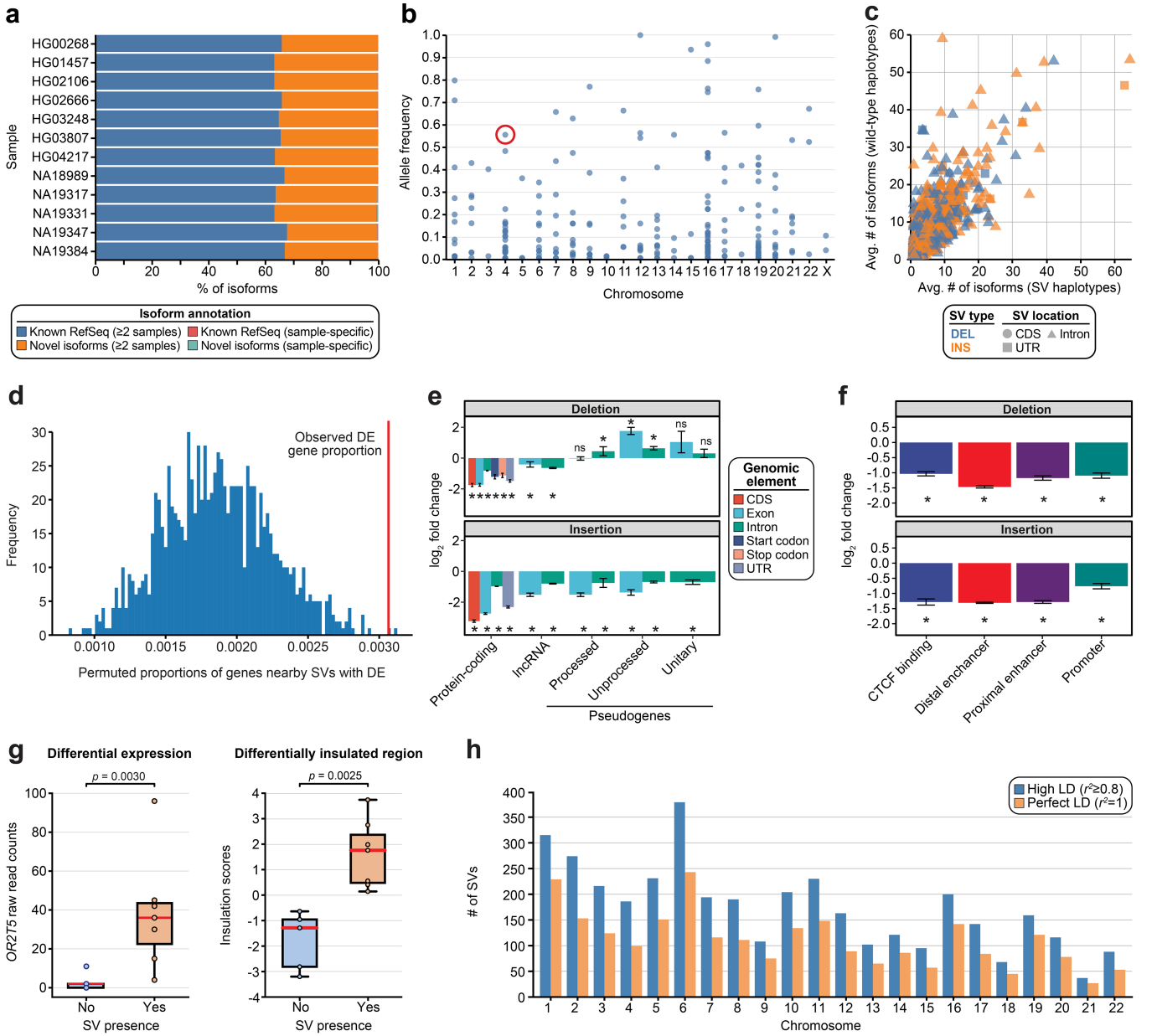
Peer review information *Nature* thanks Kai Ye and the other, anonymous, reviewer(s) for their contribution to the peer review of thiswork.



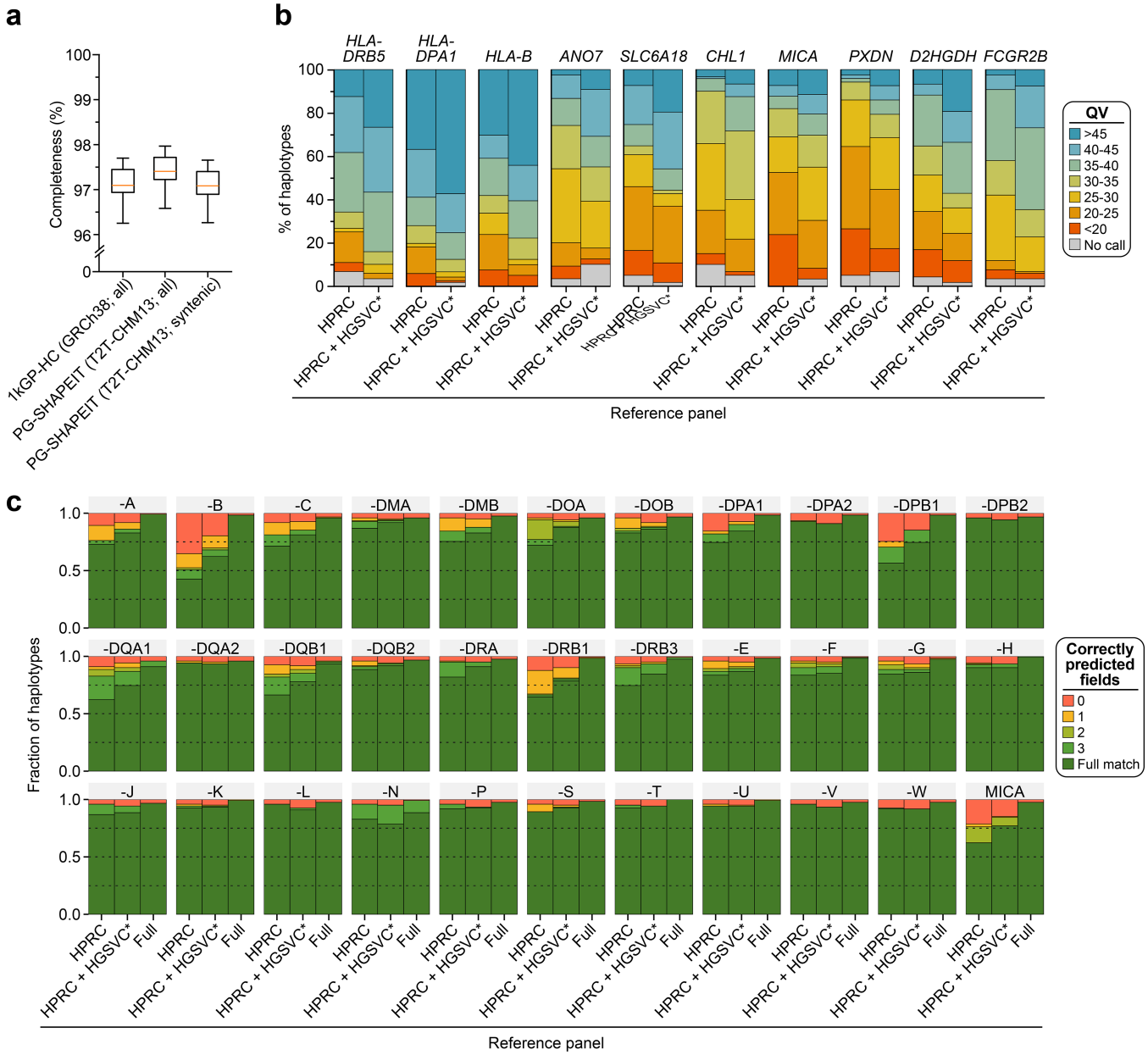
Extended Data Fig. 1 | Statistics of long-read sequencing data and genome assemblies generated in this study as well as variant calls for 65 diverse human genomes. a) Fold coverage of the Pacific Biosciences (PacBio) high-fidelity (HiFi) and Oxford Nanopore Technologies (ONT) long-read sequencing data generated for each genome in this study. The median (solid line) and first and third quartiles (dotted lines) are shown. **b**) Read length N50 of the PacBio HiFi and ONT data generated for each genome in this study. The median (solid line) and first and third quartiles (dotted lines) are shown. **c**) Gene completeness as a percentage of BUSCO single-copy orthologs detected in each haplotype from each genome assembly (Methods). **d**) The number of SVs identified in one individual by 14 different SV callers, including PAV (Methods). Each bar is divided into four categories as follows: PAV, SVs identified by PAV (black); PAV (trimmed), false SVs from other callers in redundantly aligned sequences that PAV removes (red); Covered, SVs not called by PAV but within callable loci spanned by assembly alignments (dark gray); No assembly, SVs identified in locations not callable by PAV (light gray). Before applying caller-based QC, 99.75% of PAV calls are supported by at least one other call source. The individual evaluated is HG00171. **e**) Number of SVs called for each haplotype relative to the GRCh38 reference genome, colored by population. Insertions and deletions are imbalanced when called against the GRCh38 reference genome but balanced when called against the T2T-CHM13 reference genome (Fig. 1g). **f**) Number of SV insertions (left) and deletions (right) called against T2T-CHM13, GRCh38, or both reference genomes relative to their allele frequency. SVs called against both references tend to be rarer because they are less likely to appear in a reference genome. A sharp peak for high allele frequency (~1.0) for insertions is detected relative to the GRCh38 reference genome but not the T2T-CHM13 reference genome.



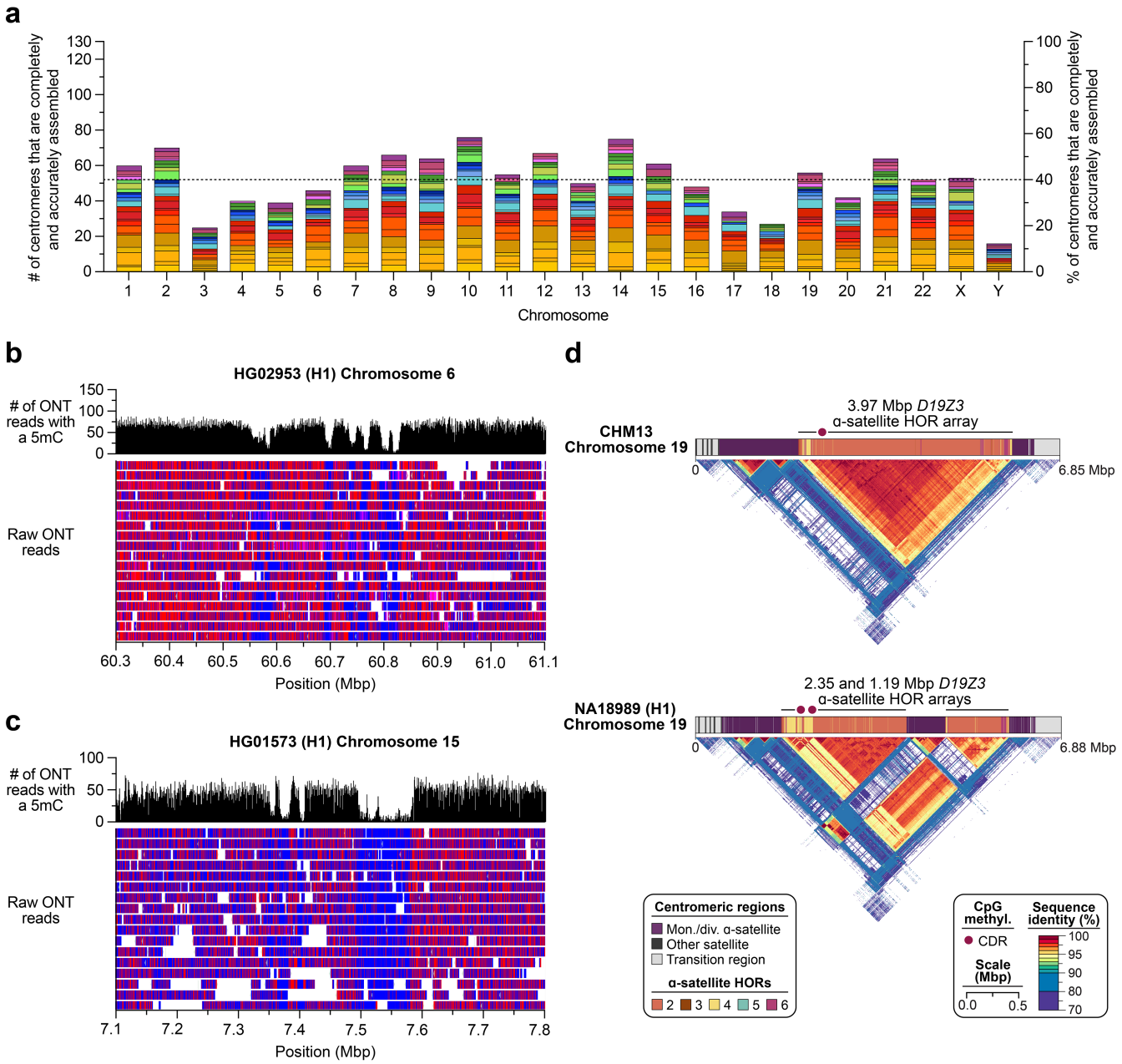
Extended Data Fig. 2 | Classification and distribution of changes in SD content in the 65 genomes. a) Number of segmentally duplicated bases assembled in different regions of the genome for each individual in this study, excluding sex chromosomes. The dashed line indicates the number of segmentally duplicated bases in the T2T-CHM13 genome. **b**) Segmental duplication (SD) accumulation curve. Starting with T2T-CHM13, the SDs (excluding those located in acrocentric regions and chrY) of 63 individuals (excluding NA19650 and NA19434) were projected onto T2T-CHM13 genome space in the continental group order of: EUR, AMR, EAS, SAS and AFR. For each bar, the SDs that are singleton, doubleton, polymorphic (>2) and shared (>90%) are indicated. The first bar is classified as "shared", as the assembly is only being compared to itself. **c**) Schematic depicting the four categories of non-reference SDs: 1) new (i.e., unique in the reference), 2) expanded copy number, 3) content or composition changed, and 4) expanded and content changed SDs with respect to the SDs in the reference genome, T2T-CHM13. **d**) Quantification in terms of Mbp and predicted protein-coding genes across the four categories of new SDs compared to T2T-CHM13. The left panel shows the Mbp by category, while flagging those that are singleton (i.e., duplicated in T2T-CHM13 but not in other genomes). The right panel quantifies the number of complete (100% coverage) and partial overlaps (>50% coverage) with protein-coding genes for the respective chromosomes.



Extended Data Fig. 3 | Effects of SVs on gene expression, chromosome conformation, and complex traits. a) The percentage of Iso-Seq isoforms identified for each individual classified as previously identified in RefSeq (present in at least two individuals; blue), novel (present in at least two individuals; orange), individual-specific previously identified isoforms (red), or individual-specific novel (teal). **b**) Manhattan plot of the allele frequencies for 256 SVs disrupting protein-coding exons of 136 genes with expression present in Iso-Seq. Circled in red is the 6,142 bp polymorphic deletion in *ZNF718*. **c**) Comparison of the average unique isoforms in Iso-Seq phased to wild-type and variant haplotypes for 1,471 single SV-containing protein-coding genes. The color represents the type of SV [deletion (DEL): blue, insertion (INS): orange] and the shape indicates where the SV occurs in relation to the canonical transcript [circle: coding sequence (CDS), square: untranslated region (UTR), triangle: intron]. **d**) Proportion of genes located within 50 kbp of SV regions that show differential expression (DE; RNA-seq) among individuals who carry the SVs (red line), compared with the distribution of DE gene proportions nearby simulated SV regions (1,000 permutations). **e**) Enrichments and depletions of SVs within GENCODE v45 protein-coding, long noncoding RNA (lncRNA), and pseudogene elements, subdivided into various biotypes. *empirical $p < 0.05$ from 1,000 permutations with Benjamini-Hochberg correction. ns, nonsignificant. Error bars indicate ±1 s.d. centered on the mean. p-values are listed in Supplementary Table 43. **f**) Enrichments and depletions of SVs within classes of ENCODE candidate cis-regulatory elements (cCREs). *empirical $p < 0.05$ from 1,000 permutations with Benjamini-Hochberg correction. ns, nonsignificant. Error bars indicate ±1 s.d. centered on the mean. p-values are listed in Supplementary Table 59. **g**) A differentially insulated region in individuals with chr1-248444872-INS-63 SV, located nearby the DE gene *OR2T5*, suggests an SV-mediated novel chromatin domain could lead to increased gene expression. n = 7 individuals with the SV and 5 without the SV. Box plots indicate median and first and third quartiles, with whiskers extending to 1.5 times the interquartile range. Two-sided Wilcoxon rank-sum test with Benjamini-Hochberg correction. **h**) Number of SVs per chromosome that are in high ($r^2 > 0.8$) or perfect ($r^2 = 1$) linkage disequilibrium (LD) with GWAS SNPs significantly associated with diseases and human traits.



Extended Data Fig. 4 | Genotyping from short-read sequencing data. **a**) Completeness statistics for haplotypes produced from the 1kGP-HC phased set (GRCh38-based) and by genome inference with Pangenie followed by phasing (T2T-CHM13–based). To allow for comparison between the GRCh38- and T2T-CHM13-based callsets, we additionally restricted our analysis to "syntenic" regions of T2T-CHM13, i.e., excluding regions unique to T2T-CHM13. For both phased sets, completeness was computed on a subset of n = 30 individuals. The median is marked in yellow, and the lower and upper limits of each box represent lower and upper quartiles (Q1 and Q3). Lower and upper whiskers are defined as Q1 − 1.5(Q3−Q1) and Q3 + 1.5(Q3−Q1). **b**) Locityper genotyping accuracy for 10 target loci with the highest average variant-based QV improvement. **c**) Locityper genotyping results for HLA genes on 61 Illumina short-read HGSVC datasets using three reference panels: HPRC (90 haplotypes), leave-one-out HPRC + HGSVC (HPRC + HGSVC*, 214 haplotypes), and HPRC + HGSVC (full, 216 haplotypes). Accuracy is evaluated as the number of correctly identified allele fields in the corresponding gene nomenclature.



Extended Data Fig. 5 | Assembly of 1,246 human centromeres across 65 diverse human genomes show genetic and epigenetic variation. a) Number (left y-axis) and percentage (right y-axis) of centromeres that are completely and accurately assembled among 65 diverse human genomes, colored by population group. Mean, dashed line. **b,c**) Examples of di-kinetochores, defined as two CDRs located >80 kbp apart from each other, on the **b**) HG02953 chromosome 6 centromere and **c**) HG01573 chromosome 15 centromere. UL ONT reads span both CDRs in each case, indicating that the CDRs occur on the same chromosome in the cell population. **d**) Differences in the α-satellite HOR array organization and methylation patterns between the CHM13 and NA18989 (H1) chromosome 19 centromeres. The NA18989 (H1) chromosome 19 centromere has two CDRs, indicating the potential presence of a di-kinetochore.

nature portfolio

Corresponding author(s): Tobias Marschall, Evan E. Eichler, Christine R. Beck, Charles Lee, Jan O. Korbel, Miriam K. Konkel

Last updated by author(s): Apr 18, 2025

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection

Pacific Biosciences (PacBio) high-fidelity (HiFi) long-read sequencing data was collected using SMRT Link v10.1, v12.0, and v13.0 (software version 10.1.0.119549), and Oxford Nanopore Technologies (ONT) long-read sequencing data was collected using PromethION software (v21.02.17 - 23.04.5). BioNano Genomics optical mapping data was collected via Saphyr 2nd generation instruments (Part #60325) using Instrument Control Software (ICS) v4.9.19316.1, and PacBio isoform-sequencing (Iso-Seq) data was collected with SMRT Link v12.0 and v13.0.

Data analysis

Custom software developed for this study include L1ME-AID (https://github.com/Markloftus/L1ME-AID) and MELT-LRA (github.com/Scott-Devine/MELT-LRA).

Custom scripts and pipelines developed for this study include those for sample selection (github.com/tobiasrausch/kmerdbg and github.com/asulovar/HGSVC3_sample_selection); Verkko genome assembly [github.com/core-unit-bioinformatics/workflow-smk-genome-hybrid-assembly (prototype branch)]; assembly evaluation [github.com/core-unit-bioinformatics/workflow-smk-assembly-evaluation (prototype branch)]; project-specific code for assembly-related evaluations, supplementary tables, and plots (github.com/core-unit-bioinformatics/project-run-hgsvc-assemblies); PanGenie genotyping and reference panel construction (github.com/eblerjana/hgsvc3); MEI, MHC, Iso-Seq, and SMN analysis (github.com/Markloftus/HGSVC3 and github.com/Markloftus/L1ME-AID); MHC annotation (github.com/DiltheyLab/MHC-annotation); and segmental duplication analysis (SDA2: https://github.com/ChaissonLab/SegDupAnnotation2).

All other software used in this study are publicly available and include Verkko (v1.4.1), hifiasm (v0.19.6), Graphasing (v0.3.1-alpha), MBG (v1.0.15 and v1.0.16), GraphAligner (v1.0.17 and v1.0.18), and MashMap (v3.0.6 and v3.1.3), Foreign Contamination Screening (FCS) (v0.4.0), minimap2 (v2.24, v2.26, and v2.28), SAMtools (v1.15.1 and v1.17), ISOGG (v15.73), CDR-Finder, modbam2bed (v0.10.0), bedGraphToBigWig, IGV, HMMER (v.3.3.2dev), TRF (v4.1.0), NAHRwhals, ArbiGent, ASHLEYS, MosaiCatcher(v2), Immuannot (MHC reference version: IPD-IMGT/HLA-V3.55.0), L1ME-AID (v1.0.0-beta), SNPrelate R package157 (v1.26.0), Factoextra (v1.0.7), NucFreq (NucFreq version "bd080aa" (from fork

NucFreqTwo / branch "split-two-phases"), Flagger (v0.3.3), Meryl (v1.0), Winnowmap2 (v2.03), DeepVariant (v1.6.0), bcftools (v1.17), Merqury (v1.0), compleasm (v0.2.5), OrthoDB (v10), mashmap (v3.1.3), PAV (v2.4.1), pbmm2 (v1.1.0, v1.5.0, and v1.12.0), LRA (v1.3.7.2), DipCall (v0.3), SVIM-asm (v1.0.3), PBSV (v2.9.0), Sniffles (v2.0.7), Delly (v1.1.6), cuteSV (v2.0.3), DeBreak (v1.0.2), SVIM (v2.0.0), DeepVariant (v1.5.0), DeepVariant executed through PEPPER-Margin-DeepVariant (vr0.8), Clair3 (v1.0.4), SV-Pop (v3.4.4), BCFtools (v1.16 and v1.17), BEDtools (v2.29.0 , v2.30.0, and v2.31.1), SciPy (v1.11.4), RepeatMasker (v4.1.0, v4.1.2, and v4.1.6), Biopython (v1.82), SEDEF (v1.1), Windowmasker (v2.2.22), seqtk (v1.3), vamos (v1.3.2), VCFtools (v.0.1.16), BEAST (v.1.10.4), RAxML (v.8.2.10), Tree-Annotator (v.1.10.4), FigTree software (v.1.4.4), Trimmomatic (v0.39), STAR (v2.7.10b), Cufflinks (v2.2.1), Lima (v2.1.0), isoseq3 (v3.8.2), SQANTI3 (v5.1.2), MUSCLE (v3.8.425), DESeq2 (v1.38.3), FAN-C (v0.9.26b2), Minigraph-Cactus (v2.7.2), PanGenie (v3.1.0), BLASTN (v2.14.1), rustybam (v0.1.33, 10.5281/zenodo.8106233), R (v1.1.383), R package ape (v5.7-1), R package phangorn (v2.11.1), Parascopy (v1.16.0), SMNCopyNumberCaller (v1.1.2), pgr-tk (v0.5.1), dna-brnn (v0.1), Graphpad Prism (v9), StainedGlass (v6.7.0), and Snakemake (v7.19.1).

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data produced by the HGSVC and analyzed as part of this study are available under the following accessions: PacBio HiFi and ONT long reads: PRJEB58376, PRJEB75216, PRJEB77558, PRJEB75190, PRJNA698480, RJEB75739, PRJEB36100, PRJNA988114, PRJNA339722, PRJEB41778, ERP159775; Strand-seq: PRJEB39750, PRJEB12849; Bionano Genomics: PRJNA339722, PRJEB41077, PRJEB58376, PRJEB77842; HiC: PRJEB39684, PRJEB75193, PRJEB58376; PacBio Iso-Seq: PRJEB75191; RNA-seq: PRJEB75192, PRJEB58376. Released resources including simple and complex variant calls, graph genomes, genotyping results (genome-wide and targeted), and annotations for centromeres, mobile element insertions, and segmental duplications can be found in the IGSR release directory hosted publicly via HTTP and FTP (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/release) and on the Globus endpoint "EMBL-EBI Public Data" in directory "/1000g/ftp/data_collections/HGSVC3/working".

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We sequenced 65 human samples in this study, including 30 males (46,XY) and 35 females (46,XX).
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	The samples included in the study represent 26 human populations, as defined by the 1000 Genomes Project.
Recruitment	Samples included in this study were of either African (n=30), Admixed American (n=9), European (n=8), East Asian (n=10), or South Asian (n=8) descent.
Ethics oversight	The lymphoblastoid cell lines and genomic DNA for each sample are available from the Coriell Institute for Medical Research (https://www.coriell.org/) for research purposes and are covered by the appropriate ethics approvals by the Coriell Institute.

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

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	A total of 65 human samples were included in this study.
Data exclusions	No data was excluded.
Replication	N/A. All computational analyses can be replicated using the provided codes and pipelines.

Randomization	N/A. Samples were not assigned to groups.
Blinding	N/A. All experiments were done computationally and do not involve a human experimenter.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

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

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	All cell lines were obtained from the Coriell Institute for Medical Research (https://www.coriell.org/) and used to generate sequencing data, including: HG00096, HG00171, HG00268, HG00358, HG00512, HG00513, HG00514, HG00731, HG00732, HG00733, HG00864, HG01114, HG01352, HG01457, HG01505, HG01573, HG01596, HG01890, HG02011, HG02018, HG02059, HG02106, HG02282, HG02492, HG02554, HG02587, HG02666, HG02769, HG02818, HG02953, HG03009, HG03065, HG03248, HG03371, HG03452, HG03456, HG03520, HG03683, HG03732, HG03807, HG04036, HG04217, NA12329, NA18534, NA18939, NA18989, NA19036, NA19129, NA19238, NA19239, NA19240, NA19317, NA19331, NA19347, NA19384, NA19434, NA19650, NA19705, NA19836, NA19983, NA20355, NA20509, NA20847, NA21487, and NA24385.
Authentication	We did not authenticate the cell lines.
Mycoplasma contamination	According to information provided by the Coriell Institute for Medical Research, all cell lines are free of bacterial, fungal or mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used.

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A