Gaps and complex structurally variant loci in phased genome assemblies

David Porubsky,[1] Mitchell R. Vollger,[1] William T. Harvey,[1] Allison N. Rozanski,[1] Peter Ebert,[2,3] Glenn Hickey,[4] Patrick Hasenfeld,[5] Ashley D. Sanders,[6,7,8] Catherine Stober,[5] Human Pangenome Reference Consortium,[11] Jan O. Korbel,[5,9] Benedict Paten,[4] Tobias Marschall,[2,3] and Evan E. Eichler[1,10]

[1]*Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA;* [2]*Institute for Medical Biometry and Bioinformatics, Medical Faculty, Heinrich Heine University, 40225 Düsseldorf, Germany;* [3]*Center for Digital Medicine, Heinrich Heine University, 40225 Düsseldorf, Germany;* [4]*UC Santa Cruz Genomics Institute, University of California Santa Cruz, Santa Cruz, California 95064, USA;* [5]*European Molecular Biology Laboratory (EMBL), Genome Biology Unit, 69117 Heidelberg, Germany;* [6]*Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, 10115 Berlin, Germany;* [7]*Berlin Institute of Health (BIH), 10178 Berlin, Germany;* [8]*Charité-Universitätsmedizin, 10117 Berlin, Germany;* [9]*European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom;* [10]*Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA*

There has been tremendous progress in phased genome assembly production by combining long-read data with parental information or linked-read data. Nevertheless, a typical phased genome assembly generated by trio-hifiasm still generates more than 140 gaps. We perform a detailed analysis of gaps, assembly breaks, and misorientations from 182 haploid assemblies obtained from a diversity panel of 77 unique human samples. Although trio-based approaches using HiFi are the current gold standard, chromosome-wide phasing accuracy is comparable when using Strand-seq instead of parental data. Importantly, the majority of assembly gaps cluster near the largest and most identical repeats (including segmental duplications [35.4%], satellite DNA [22.3%], or regions enriched in GA/AT-rich DNA [27.4%]). Consequently, 1513 protein-coding genes overlap assembly gaps in at least one haplotype, and 231 are recurrently disrupted or missing from five or more haplotypes. Furthermore, we estimate that 6–7 Mbp of DNA are misorientated per haplotype irrespective of whether trio-free or trio-based approaches are used. Of these misorientations, 81% correspond to bona fide large inversion polymorphisms in the human species, most of which are flanked by large segmental duplications. We also identify large-scale alignment discontinuities consistent with 11.9 Mbp of deletions and 161.4 Mbp of insertions per haploid genome. Although 99% of this variation corresponds to satellite DNA, we identify 230 regions of euchromatic DNA with frequent expansions and contractions, nearly half of which overlap with 197 protein-coding genes. Such variable and incompletely assembled regions are important targets for future algorithmic development and pangenome representation.

[Supplemental material is available for this article.]

The past two years have witnessed tremendous progress with respect to advances in sequencing technology (Lu et al. 2016; Vollger et al. 2019; Wenger et al. 2019), as well as numerous assembly strategies that now make it possible to phase and assemble >95% of the content of a diploid genome (Logsdon et al. 2021; Jarvis et al. 2022). Because of these developments, genome assemblies have changed in two significant ways. We no longer consider collapsed 3-Gbp genome assemblies as state of the art (i.e., one representation of an individual where both haplotypes are merged) but instead consider two genomes for every diploid genome assembled (i.e., 6 Gbp vs. 3 Gbp) where parental haplotypes are phased and fully resolved. Second and in part because of the first, the number of gaps being produced has been reduced from thousands to only a few hundred. As a result, there have been a series of efforts to generate more complete and phased human genome assemblies using long-read sequencing platforms, including the Human Genome Structural Variation Consortium (HGSVC) and the Human Pangenome Reference Consortium (HPRC) (Ebert et al. 2021; Liao et al. 2023). Efforts such as these have generated data with different sequence technologies and applied different algorithms and strategies to generate multiple phased human genomes, including some that now rival the contiguity and accuracy of the current human genome reference (GRCh38).

In particular, the development of Pacific Biosciences (PacBio) high-fidelity (HiFi) reads, based on circular consensus sequencing (CCS) technology, provides ~20-kbp sequencing reads that compete with short reads with respect to their accuracy (QV > 30), whereas the Oxford Nanopore Technologies (ONT) platform now can routinely generate sequencing reads in excess of 100 kbp (so-called ultralong [UL] sequencing reads) (Nurk et al. 2020; Shafin et al. 2020; Logsdon et al. 2021). The use of parent–child

[11]A complete list of contributing Consortium members appears at the end of this paper.
Corresponding author: eee@gs.washington.edu
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.277334.122. Freely available online through the *Genome Research* Open Access option.

trio (trio-hifiasm) Illumina whole-genome sequencing (WGS) data in conjunction with CCS data provides the greatest power to phase a genome into its constituent paternal and maternal haplotypes. In the absence of parental data, however, methods have been developed (PGAS and HiC-hifiasm) using linked-read data, such as Strand-seq (Porubsky et al. 2021) or Hi-C (Garg et al. 2021; Cheng et al. 2022), that can phase genomes at the local and chromosomal level.

The challenge that remains is routine telomere-to-telomere (T2T) assembly of human genomes such that the full genetic diversity of species can be understood. Assembly gaps are, unfortunately, still an integral feature of every de novo diploid genome assembly. This status quo will remain until the sequencing technology and assembly algorithms evolve so that each homologous chromosome of any genome can be routinely assembled T2T in an automated fashion. Key to this aspirational goal is understanding why gaps persist, which in turn requires a detailed analysis of gap size, frequency, genomic location, and the sequence properties that define these regions. With the completion and annotation of the first T2T genome (Nurk et al. 2022), we are in a position to characterize the properties of the gaps that remain when diploid human genomes are routinely sequenced. We focus on a detailed characterization of these remaining gaps in an effort to understand their origin, biology, and the relative importance of getting these through the last impasses to T2T assembly. We focus on human diploid genomes because resolution of the gaps will improve discovery of both disease-related variation as well as genetic changes important for the evolution and adaptation of our species.

Results

We investigated the gaps and contig breaks in a total of 182 haploid assemblies obtained from a diversity panel of 77 unique human samples sequenced with long-read technology. The underlying long-read data and assemblies were generated by two consortia over the past two years, HGSVC (88 assemblies) and HPRC (94 assemblies), using different long-read sequencing platforms as well as assembly strategies. The HGSVC used two different long-read sequencing technologies, continuous long-read (CLR; 60 assemblies) sequencing (Ebert et al. 2021) and CCS (or HiFi sequencing, 28 assemblies) with an additional eight samples shared between HGSCV and HPRC used only for validation purposes. CCS and CLR data from HGSVC were assembled using a trio-free assembly pipeline, called PGAS (Ebert et al. 2021; Porubsky et al. 2021; Ebler et al. 2022) using both the Peregrine (Chin and Khalak 2019) (PGASv12) and the hifiasm (Cheng et al. 2021) (PGASv13) assemblers for CCS and the Flye assembler (Kolmogorov et al. 2019) for CLR data. The HPRC effort, which began more than a year later, focused exclusively on CCS data (n = 94) generated from diploid samples assembled using trio-based hifiasm (Cheng et al. 2021). Here, parent–child data were directly used to aid assembly phasing of all HPRC samples (Wang et al. 2022; Liao et al. 2023), allowing for both platform and methodology comparisons (Supplemental Table S1; Supplemental Fig. S1A).

Evaluation metrics and gap definitions

In this study, we set out to evaluate assembly quality and completeness using four metrics (Methods). We start with defining regions between subsequent contigs mapped to the T2T-CHM13 human genome reference. These are defined based on reliable "contig end alignments" (≥50 kbp at the contig edges) mapped in agreement with an expected contig length. Contig end alignments were used to localize regions (assembly gaps) in between subsequent contigs (Fig. 1A, i). Second, we define "simple contig ends" as terminal contig positions with respect to the reference genome. Simple contig ends were used for enrichment analysis of various genomic features near terminal contig alignment positions (Fig. 1A, ii). To evaluate structural differences between assemblies, we set to document all regions that break contig alignments, referred to here as "contig alignments discontinuities." We focus on discontinuities that create internal gaps within contig alignments <1 Mbp in length to document regions of putative structural differences that cannot be readily aligned to a single reference (Fig. 1A, iii). Lastly, we turn our attention to regions with a higher coverage than expected in a haploid genome (multicoverage regions) caused by two or more overlapping contig alignments. Such regions point to positions of either true structural differences or genome assembly artifacts (Fig. 1A, iv).

Platform and assembly method comparisons

We initially compared assembly statistics between different sequencing technologies and assembly algorithms to determine what combination provides the most continuous and complete assembly. The most fragmented assemblies were obtained using a combination of the trio-free PGAS pipeline and the Peregrine assembler with a median contig count of 7900 per assembly (Ebert et al. 2021). Improved contiguity was achieved by combination of the PGAS pipeline and CLR data assembled by Flye (median contigs, 2170) and CCS data assembled by hifiasm (median contigs, 1647) (Ebert et al. 2021; Ebler et al. 2022). The most continuous assemblies were obtained using the trio-based hifiasm assembly, resulting in an order of magnitude fewer gaps (e.g., 399 median contigs per assembly) (Fig. 1B). The least complete assemblies resulted from a combination of PGAS and CLR data (median size, 2.85 Gbp). This is expected because higher error rates of CLR in comparison to CCS data prevent them from assembling highly identical segmental duplications (SDs) in the human genome. Assemblies using CCS data provide comparable assembly completeness (median size, ~3.05 Gbp) with a slightly higher median assembly size for the trio-free PGAS pipeline combined with hifiasm (median size, 3.14 Gbp) (Fig. 1B). Lastly, the assembly contiguity was evaluated as a function of contig N50, and again, we conclude that trio-based assembly (N50, 40.83 Mbp) outperforms those assembled in trio-free settings (Fig. 1C). Because of suboptimal performance, we excluded Peregrine assemblies from subsequent analyses.

Consistent with a recent study (Jarvis et al. 2022), trio-based assemblies contain the least number of gaps between contig alignment ends (median, 141) followed by PGAS-hifiasm with about double that amount (median, 320) and PGAS-Flye (median, 392) (Supplemental Fig. S1B). Based on projections to the T2T-CHM13 reference, the number of missing base pairs follows a similar trend, with trio-based assemblies having the least number of bases within gaps between defined contig alignment ends (median, 78.4 Mbp) followed by PGAS-hifiasm (median, 126.7 Mbp) and PGAS-Flye (median, 244.8 Mbp) (Supplemental Fig. S1C), although there are outliers (Supplemental Fig. S2). CCS-based assemblies are generally superior to those produced from CLR because highly identical SDs, including disease-relevant regions such as Prader–Willi, are largely absent from CLR-based assemblies (Fig. 1D, white gaps). As a result, ~59.9 Mbp is missing in CLR assemblies in contrast to only ~690 kbp in CCS-based assemblies,

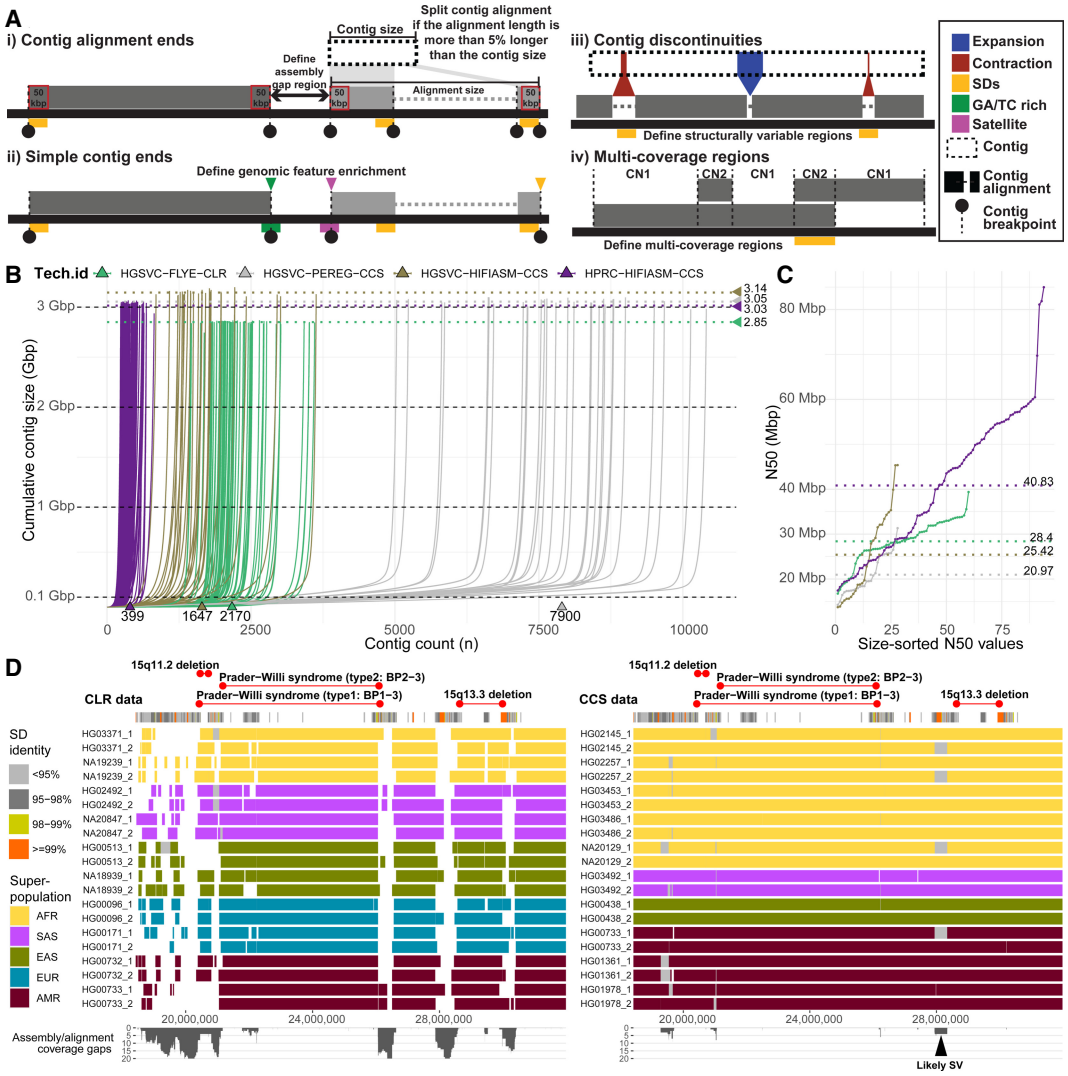

Figure 1. Comparison and evaluation of phased assemblies. (*A*) Assembly metrics evaluated in this study. (*i*) Contig alignment ends are defined as terminal contig alignments such that the total alignment size does not exceed the actual contig size by >5%. When this requirement is not met, multiple contig end alignments will be reported. (*ii*) Simple contig ends are defined as the first and last alignments of each contig to the reference (T2T-CHM13 v1.1) with at least 25 kbp aligned. (*iii*) Contig discontinuities are defined as alignment gaps between subsequent pieces of a single contig <1 Mbp. (*iv*) Detection of regions with coverage more than 1n as is expected for a haploid genome. (*B*) A cumulative contig size distribution colored by assembly technology. Each line represents a single haploid assembly (HGSVC-FLYE-CLR, n = 60; HGSVC-PEREG-CCS, n = 28; HGSVC-HIFIASM-CCS, n = 28; HPRC-HIFIASM-CCS, n = 94). Median total assembly length per assembly technology is highlighted as horizontal dotted lines. (*C*) Contig N50 values colored by assembly technology as in *B*. Each dot represents a single haploid assembly. Median N50 value per assembly technology is highlighted as horizontal dotted lines. (*D*) Track definition from *top* to *bottom*: Regions corresponding to known genomic disorders between 15q11.2–15q13.3. *Below* is the annotation of SDs in this region colored by sequence identity. Main track shows the visualization of contig alignments for 10 random samples from trio-free CLR assemblies (*left*) in comparison to trio-based HPRC assemblies (*right*). Contig alignments are colored by sample superpopulation (AFR, African; SAS, Southeast Asian; EAS, East Asian; EUR, European; AMR, American). White spaces between contig alignments represent boundaries between subsequent contig. Spaces filled with gray color represent unaligned portions of a single contig with respect to the reference (T2T-CHM13) and likely represent a structural variation (black arrowhead). The last track summarizes the extent of assembly gaps (between contigs; white space) and contig gaps (within contigs; gray rectangles) as coverage plot.

allowing us to begin to assess SD-associated copy number variation and structural variation (Fig. 1D, gray gaps). Given these observations, we exclude CLR-based assemblies from subsequent analysis and focus exclusively on CCS-hifiasm assemblies.

Parent–child trio-based versus trio-free assemblies

We compared in more detail eight human genomes for which both long-range linked reads (Strand-seq) and parental data (Illumina WGS) were available from the same individuals. Using the same underlying long-read input data (CCS), we specifically performed a head-to-head comparison of trio-based (TRIO; using parental Illumina WGS for phasing) and trio-free (PGAS; using Strand-seq for phasing) assemblies. We find that assemblies generated in the absence of parental data (trio-free) have about twice as many contigs and a decreased contig N50 by ~10 Mbp (Supplemental Fig. S3), likely because the underlying assembly algorithm reuses paths as opposed to generating a primary and alternate in the absence of parental data. We next evaluated phasing accuracy of trio-free assemblies using the genomes phased by parental data as the

truth set (Methods). For the metacentric and submetacentric chromosomes, we observe a high accuracy of phased 1-Mbp segments, achieving 98% concordance with trio-based phasing. With acrocentric chromosomes, this accuracy drops to 94% (Fig. 2A; Supplemental Fig. S4). The majority of incorrectly assigned 1-Mbp segments (>75%) map within centromeric satellite repeats, most likely owing to the lower density of uniquely mapped single-nucleotide variants (SNVs) (Supplemental Fig. S5). There was only one sample (HG01891) with large-scale switch errors on a short arm of Chromosome 9 (~42 Mbp) and one at the very end of Chromosome 9 (~1 Mbp) (Fig. 2B). The data show that trio-free assemblies provide comparable phasing accuracy and completeness and are a viable option for phased genome assembly for samples in which parental data are not easily available or are cost prohibitive.

Strand-seq also preserves directionality of single-stranded DNA and thus is also able to unambiguously define misoriented regions of the genome. Such misorientations will appear as unresolved homozygous inversions based on Strand-seq reads mapping from the original genome sample (Methods). We detected comparable numbers of unresolved homozygous inverted regions in trio-based (n = 23) and trio-free (n = 15) assemblies, respectively (Supplemental Table S2), resulting in 6.8 Mbp (0.23%) and 7.3 Mbp (0.25%) of misoriented base pairs per assembly (Fig. 2C). The majority (31/38, >81%) of these misorientations overlap with previously defined true inversion polymorphisms in the human genome (Porubsky et al. 2022), six of which were unresolved in both trio-based and trio-free assemblies (Supplemental Fig. S6A) and, as expected, are flanked by large tracts of SDs (Supplemental Fig. S6B). Some of these span genomic disorder critical regions where recurrent de novo copy number variants (CNVs) associate with neurodevelopmental delay, such as the 16p11.2–p12.2 microdeletion and microduplication (Supplemental Fig. S7).

We more systematically evaluated the potential of both assembly approaches to resolve known large (≥100 kbp, n = 20) inversions considering both heterozygous as well as homozygous sites (Methods). Trio-based assemblies resolve 78% of inversion polymorphisms, whereas trio-free assemblies resolve 68% (Fig. 2D). Trio-based approaches generally more accurately represent more inverted base pairs (64%) compared with the trio-free

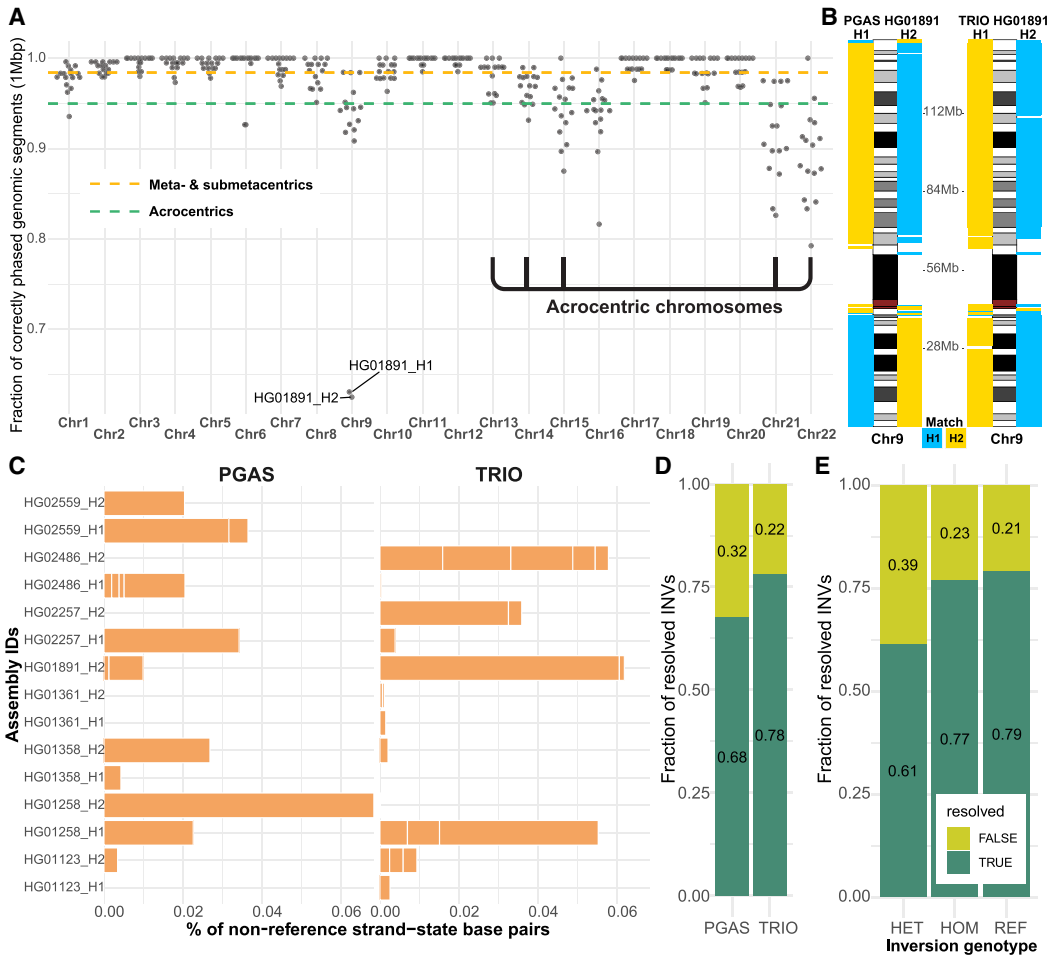

Figure 2. Phasing accuracy and inversion analysis of trio-based and trio-free assemblies. (*A*) Phasing accuracy of PGAS (trio-free) assemblies with respect to trio-based phasing. (*B*) Haplotype assignment of 1-Mbp-sized blocks (*left* from ideogram, H1; *right* from ideogram, H2) to either haplotype 1 or 2 (blue, H1; yellow, H2) using single-nucleotide polymorphisms phased using trio information (1000 Genomes Project panel) with respect to the reference (GRCh38). (*C*) A barplot reporting the percentage of base pairs in an opposite (reverse) orientation in contrast to the expected (direct) orientation based on Strand-seq analysis of assembly directionality, shown separately for trio-free (PGAS, n = 15; *left*) and trio-based (TRIO, n = 23; *right*) assemblies. (*D*) Fraction of tested inversion sites that are fully informative (TRUE; dark green). (*E*) Fraction of tested inversion sites that are fully informative (TRUE; dark green) as a function of inversion genotype. (HET) Heterozygous, (HOM) homozygous inverted, (REF) homozygous reference.

approach (48%) by virtue of the fact they often assemble one end of an inversion polymorphism (Supplemental Fig. S8A,B). It is notable that nearly a quarter of all large inversion polymorphisms are not accurately represented in existing trio-based genome assemblies, with heterozygous inversions being the most difficult to fully resolve (Fig. 2E; Supplemental Fig. S8C). All sites (n = 14) that are unresolved two or more times in trio-based and trio-free assemblies are flanked by large (>40 kbp; median, 228.2 kbp) highly identical SDs (median, 99.4%). The availability of Strand-seq data provides a valuable orthogonal method for detection of such errors in the assembly, which in turn can guide targeted reassemblies of such regions using UL ONT reads.

Sequence properties of the gaps

Because the HPRC-phased genome assemblies represent the current state of the art in terms of both accuracy, phasing, and contiguity (Figs. 1, 2), we focused on a more in-depth analysis of sequence content of gap regions by mapping all sequence contigs to the complete human reference (T2T-CHM13, v1.1) (Nurk et al. 2022). Among the 94 HPRC haplotype assemblies, we identified a total of 68,515 simple contig ends for an average of 729 per haplotype (median, 700) (Fig. 1C; Supplemental Table S3). Of these contig breaks, about two-thirds correspond to SDs (35.4%; [11,702 + 12,550]/68,515) or satellite DNA (22.3%; [2896 + 12,363]/68,515) (Fig. 3A). Because long tracts of GA repeats have been predicted to reduce the coverage of CCS data (Nurk et al. 2020), it is important to note that 27.4% ([6212 + 12,550]/68,515) of the gaps, including recurring gaps, within the assemblies correspond to regions where high GA/TC tracts are observed (1-kbp window with >80% GA/TC within 10 kbp). These GA/TC tracts show the most substantial (29.36-fold) (Fig. 3B) enrichment for gaps and, along with high AT content, account for ∼40% of the assembly breaks not associated with large repetitive sequences ([6212 + 5494]/[68,515 – 2896 – 12,363 – 11,702 – 12,550]). Controlling for sequence coverage, we estimate that nearly two-thirds of the GA/TC gaps can be remedied by simply increasing sequence coverage from approximately 30- to 50-fold (Fig. 3C). However, we also find long tracts of GA/TC repeats nonrandomly associated with regions of SDs (Fig. 3A). In such regions, increasing coverage has little effect on reducing the number of gaps and perhaps has even the opposite effect (Fig. 3D). We considered both the length and sequence identity of SDs and found that the longer and more identical an SD is, the more likely it was associated with a gap. Thus, the longest and most identical SDs are preferentially associated with gaps in the majority of analyzed assemblies (Fig. 3E; Supplemental Fig. S9).

Despite the differences in contig end definition, we found a high level of agreement between simple contig ends and assembly gap regions, with >85% of simple contig ends falling into assembly gaps and >99% of assembly gaps overlapping with simple contig ends (Fig. 3F; Supplemental Fig. S10). Assembly gaps are regions that are not completely assembled across HPRC assemblies. This is especially problematic when assessing human diversity among protein-coding genes. The whole set of assembly gaps (n = 14,662) from all HPRC assemblies overlaps a total of 1513 protein-coding genes (Supplemental Fig. S11) that fall within 894 nonredundant gap regions. There are 231 protein-coding genes that fall within regions broken in five or more HPRC assemblies (Supplemental Fig. S12; Supplemental Table S4), and 31 of these lie within regions of recurrent microdeletion and microduplication syndromes (Cooper et al. 2011; Coe et al. 2014). Among these, there are a number of biomedically relevant genes, such as *PAK2* affected by 3q29 microdeletion, *CTNND2* affected in Cri-du-Chat syndrome, or *MAPT* affected by 17q21.31 microdeletion (Fig. 3F, inset).

Overall, we define 592 nonredundant regions, outside of satellite DNA, with an assembly gap in five or more of the HPRC assemblies (Supplemental Fig. S13; Supplemental Table S5). Among the most recurrent gaps, there are 44 euchromatic regions that fail to resolve in half or more of the HPRC assemblies. Although a third of these are associated with SDs, 28 of these are dropouts associated with the presence of low-complexity DNA (Supplemental Table S6). In these 28 regions, we observe continuous tracts of dinucleotides (AT or GA/TC) ranging from ∼300–6500 kbp in the T2T-CHM13 reference (Supplemental Fig. S14); however, we noticed several such low-complexity tracts in regions associated with SDs (n = 16) (Supplemental Fig. S15). We further explored the extent of the variability in size of low-complexity regions between humans and nonhuman primates in assemblies that managed to span these regions (Methods). We catalog 27/44 regions with observable differences in size of dinucleotide tracts, with humans carrying longer dinucleotide tracts in all but one instance (Fig. 4A). Our analysis suggests that many of these regions appear to have expanded specifically in the human lineage, where they continue to show variability in size (Fig. 4B,C).

Discontinuous alignments and large structural variants

One of the advantages of the new assemblies of the human genome is that they are not guided by existing human references. Such de novo assemblies have the potential to identify large discontinuities corresponding to potential larger forms of genetic variation, including partially sequence-resolved CNVs. We searched specifically for contig alignment discontinuities (<1 Mbp) as identified by alignment to the complete human reference genome (T2T-CHM13, v1.1; Methods) (Fig. 1A). Across all 94 human haplotypes, we report a median 6.6% and 0.06% of unaligned bases per assembly within and outside of centromeric satellite DNA, respectively (Supplemental Fig. S16). Per haploid genome, we define a median number of 165 contractions and 262 expansions, which corresponds to ∼11.9 Mbp and 161.4 Mbp, respectively (Supplemental Fig. S17A,B). The vast majority of these bases (contractions, 10.9 Mbp; expansions, 159.8 Mbp) belong to centromeric satellite DNA, which are known to vary extensively in size and composition among human haplotypes and are often incompletely assembled (Supplemental Fig. S17C). Nevertheless, within euchromatic regions, we identified 230 regions that showed evidence of contraction (n = 120) or expansion (n = 110) in multiple human haplotypes (five or more) compared with the T2T-CHM13 reference (Fig. 5A; Supplemental Table S7). A large number of these regions overlap with SDs (∼40%, 93/230) and include biomedically relevant loci that are known to be structurally variable, such as 8p23.1, HLA, *SMN1/SMN2*, and *TBC1D3* (Fig. 5B; Supplemental Fig. S18; Vollger et al. 2022). Based on the read-depth analysis of Illumina WGS data, we confirm 41 of these regions: the majority of which correspond to copy number losses in their respective genomes (Methods) (Supplemental Fig. S19). We highlight a region on Chromosome 11 (Chr 11: 55,535,304–55,628,574, 11q12.1) where the contracted region (∼93 kbp) is associated with short inversion (∼4 kbp) that flips *OR4C6* into a direct orientation with respect to *OR4C11*, which likely promotes a microdeletion via nonallelic homologous recombination (NAHR) as this deletion is observed in association with an inverted haplotype (Supplemental Fig. S20).

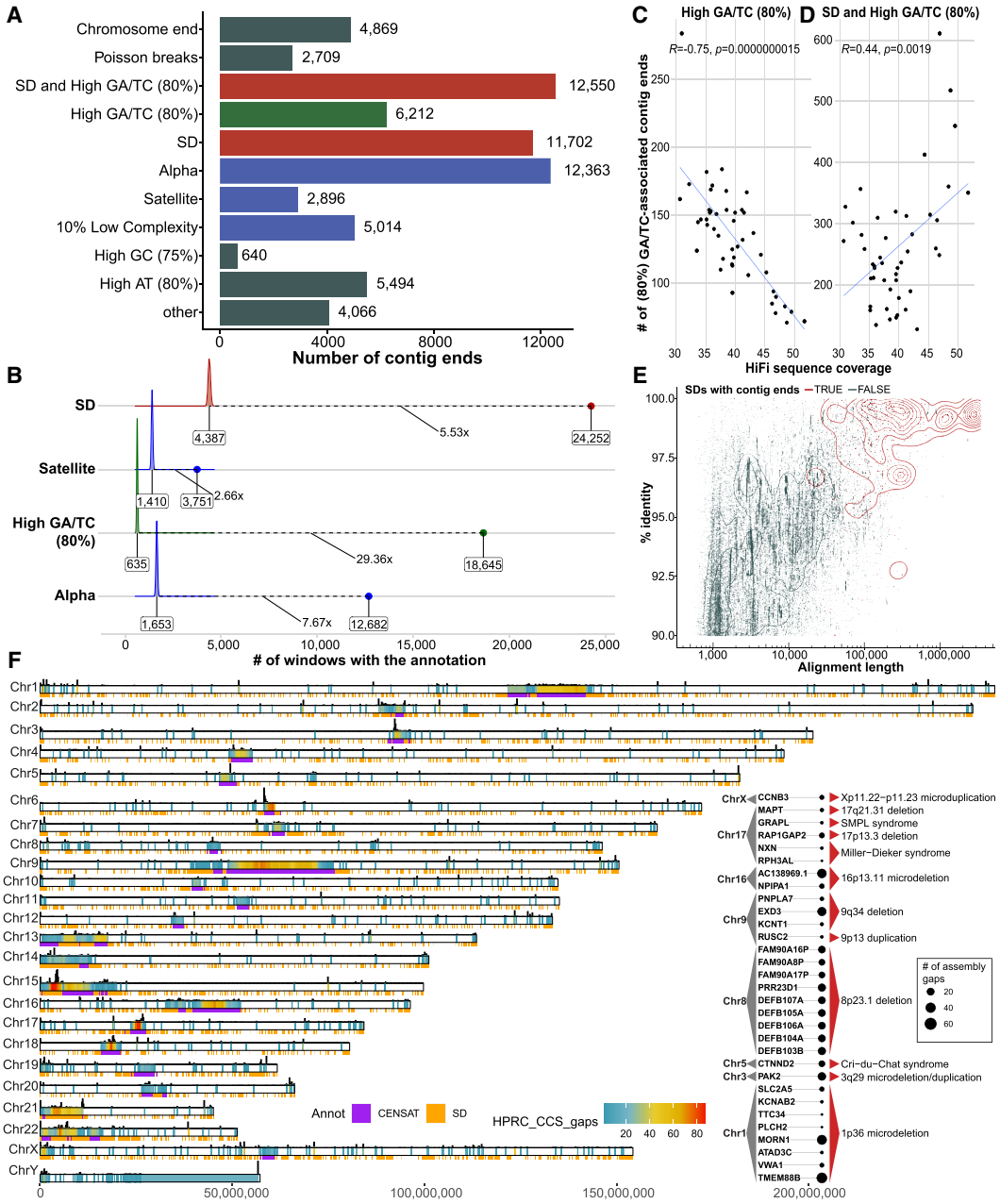

Figure 3. Sequence properties at defined contig ends. (*A*) The number of simple contig ends that are within or near (at most 10 kbp) a particular sequence annotation. Annotations are nonredundant and are prioritized in the order shown; for example, if a contig end is near the end of a chromosome and in an SD, it will only be annotated as a chromosome end. Note that chromosome ends are contig ends within the last 100 kbp of contigs. Poisson ends are contig ends that happen in only one haplotype (nonrecurrent and therefore likely to be random). SD and high GA/TC mean that the end is within 10 kbp of an SD and within 10 kbp of a 1-kbp window with at least 80% GA/TC content. (*B*) The fold enrichment in the number of contigs ends within 10 kbp of a sequence annotation compared with a distribution of randomly placed contig end simulations (10,000 permutations). Shown in text is the median of the random distribution (*left*), the fold enrichment (*middle*), and the observed value (*right*). In this analysis contig ends may exist in multiple categories; for example, if a contig end is near both an SD and a satellite sequence, it will appear in both simulations. (*C*) The effect of HiFi coverage on number of GA/TC breaks is negatively correlated when considered independently; however, when combined with SDs, the trend is inverted, as shown in *D*. (*E*) All SDs in T2T-CHM13 displayed by their length and percentage of identity (blue) versus the SDs that intersect contig ends (red). (*F*) Genome-wide distribution of gaps defined in between contig alignment ends (Methods) across all HPRC assemblies (n = 94). Color range reflects the number of assembly gaps overlapping each other in any given genomic region. On the *top* of each chromosomal bar, there is a density of simple contig ends. The height of each bar reflects the number of simple contig ends counted in 200-kbp-long genomic bins. *Inset*: List of protein-coding genes (n = 31) overlapping assembly breaks and reported microdeletion and microduplication syndromes.

In addition to the assembled sequence that does not readily map to the reference, we also cataloged regions where there are multiple contig mappings (more than one) instead of the expected haploid single copy (Fig. 1A, iv). Per haplotype, we observe ~15.4 Mbp of euchromatic sequence with multiple contig mappings with respect to the reference (T2T-CHM13, v1.1). Although such

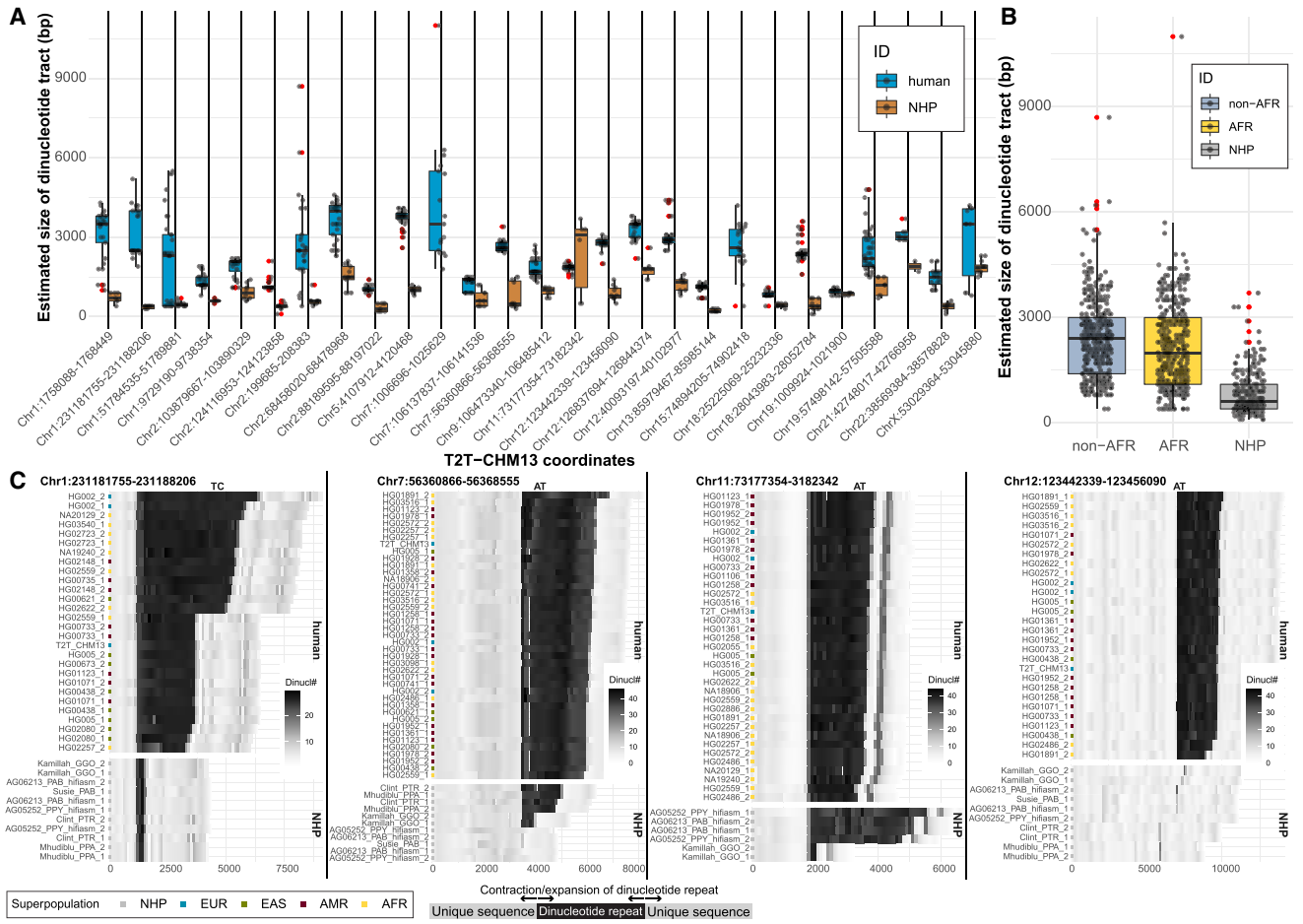

Figure 4. Sequence variation in low-complexity regions. (*A*) Size distribution comparison of dinucleotide tracts (*y*-axis) between human (blue) and non-human primates (NHPs; brown) for 27 selected regions (Methods). Outliers are highlighted as red dots. (*B*) A summary of size distribution of dinucleotide tracts (*y*-axis) between human samples of African (AFR; yellow) and non-African (non-AFR; light blue) origin and NHPs (gray) across all complete assemblies from 27 selected regions. (*C*) Difference in dinucleotide frequency (TC, AT) between humans and NHP in four genomic regions. Shades of gray color reflect the number of detected dinucleotides (defined at the *top* of each plot) in 100-bp-long DNA sequence chunks. Assembly names (*y*-axis) from NHP contain sample IDs and species-specific ID: (PTR) *Pan troglodytes*, (GGO) *Gorilla gorilla*, (PPA) *Pan paniscus*, (MMU) *Macaca mulatta*, (PAB) *Pongo abelii*, (PPY) *Pongo pygmaeus*. Numbers 1 and 2 represent parental homolog IDs of given sample assembly.

multimapping regions likely represent CNV regions arising from SD, they may also result from ambiguous contig mappings or artifacts of the assembly process. Improved mapping and assembly algorithms will be required to understand the biological significance of these regions. To enrich for true CNVs, we searched for CNV regions that were also supported by read-depth analysis of short-read data (Methods). Indeed, we identified ~3.2 Mbp predicted to be CNV (2–10 copies) and supported by short-read sequence data. An even greater fraction (~10.1 Mbp) of multimapping regions show greater CNVs (more than 10 copies) based on short-read depth, although the true copy number is more difficult to determine as the majority (>95%) of these regions overlap with SDs by >90%.

Nevertheless, we identify ~1.6 Mbp per haplotype of multimapping regions where we find no obvious CNV in short-read data (Fig. 5C). We note that a subset of these are large (≥500 kbp) and often (85/118) represent sequence contigs that are completely embedded within another larger contig in a single haplotype. We investigated eight of the longest such contigs in more detail (Methods). Comparison of heterozygous SNV patterns across these regions based on CCS data (DeepVariant calls) and phased assemblies (dipcalls) reveals conspicuous stretches of loss of heterozygosity over the region where the multimapping contigs overlap (Supplemental Fig. S21). Closer inspection reveals that the sequence variation between parental haplotypes is, however, not lost but rather present only in one contig, whereas the other contig is nearly identical to the other parental haplotype (Fig. 5D). Although the origin of such assembly artifacts is unclear, such overlapping contigs will likely pose challenges for SNV calling depending on which, if any, sequence contig is chosen.

We focused specifically on euchromatic regions where both long- and short-read data were in agreement regarding increased copy number variation (a fewer than 10 copy number increase with respect to the reference). We identified 255 nonredundant CNV regions that encompass 44.9 Mbp of the genome (Supplemental Table S8). Of these CNV regions, 87% (39.1 Mbp) correspond to SDs that are known to be copy number variable because of their propensity to undergo NAHR (Supplemental Fig. S22; Sharp et al. 2006; Sudmant et al. 2010, 2015). We find that genomes of African ancestry carry more CNV bases (~3.5 Mbp)

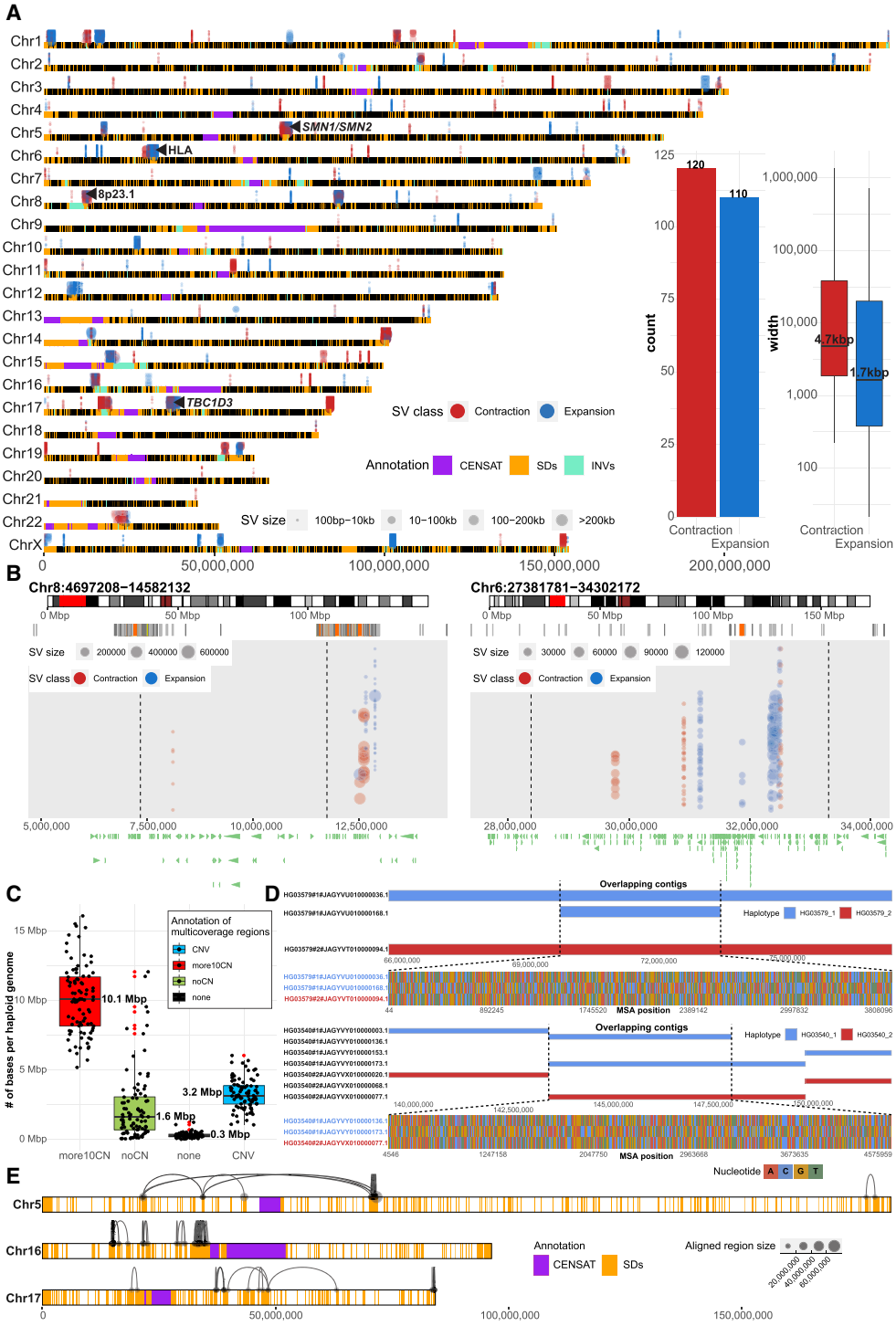

Figure 5. Tracking contig alignment discontinuities and multicoverage regions. (*A*) Genome-wide distribution of frequent (n = 230) contig alignment discontinuities (1 kbp to 1 Mbp in size). Each gap is represented in each separate assembly (HPRC, 94; HGSVC, 28) by a colored dot (blue, expansion [INS]; red, contraction [DEL]), and the size of each dot represents the size of the event in contig coordinates. A region is defined as an INS (blue) if there is a gap in a contig alignment (in reference T2T-CHM13, v1.1 coordinates) that is smaller than the sequence within a contig itself delineated by the *left* and *right* alignments flanking the gap. In contrast, a DEL (red) is defined as a gap in a contig alignment (in reference T2T-CHM13, v1.1 coordinates) that is larger than the sequence within a contig itself delineated by the *left* and *right* alignments around the gap. Putative expansions and contractions above the horizontal chromosomal lines were detected in HPRC assemblies, and those below the lines in HGSVC assemblies. Centromeric satellite regions are highlighted by gray rectangles and regions of segmental duplications (SDs) as orange rectangles on top of each chromosomal line (black). (*B*) Example regions (*left*, defensin locus, 8p23.1; *right*, HLA locus) with frequent expansions and contractions. Each region is highlighted as a red rectangle on chromosome-specific ideogram (*top* track). *Below*, there is an SD annotation for a given region represented as a set of rectangles colored by sequence identity. Expansions and contractions of each contig alignment with respect to the reference (T2T-CHM13, v1.1) are depicted as blue and red dots, respectively. The size of each dot represents the size of an event. (*C*) Assignment of total number base pairs covered by multiple contig alignments, in each haploid genome (n = 88), into four categories based on agreement with short-read-based CNV profiles (for detailed description of categories, see Methods). (*D*) Example regions in samples HG03579 and HG03540, where overlapping contigs associate with loss of heterozygosity. *Top* track shows contig alignments in a given region separately for haplotype 1 (blue; paternal) and haplotype 2 (red; maternal). Overlapping contig alignments are stacked on *top* of each other. The *bottom* track shows all variable positions detected in a multiple sequence alignment (MSA) over the region where contigs overlap (dashed lines). Here, one of the paternal contigs is nearly identical to a maternal contig at the region where contigs overlap. (*E*) Chromosomes 5, 16, and 17 are depicted as horizontal bars with the locations of SDs and centromeric regions highlighted as orange and purple rectangles, respectively. Contig alignment ends divided into multiple pieces are visualized as links between subsequent pieces of a single contig aligned to the reference (T2T-CHM13 v1.1). The length of the aligned pieces of a contig are defined by the size of each dot.

compared with other non-African populations (Supplemental Fig. S23) consistent with previous reports (Sudmant et al. 2015; Chaisson et al. 2019; Byrska-Bishop et al. 2022). The regions are particularly gene-rich, and we identify 420 protein-coding genes in 165 of them (Supplemental Table S8).

Large-scale CNVs within an assembled contig may also lead to alignment discontinuities in which contig alignment ends map far away from each other, thus exceeding the expected contig length. We identified 1721 contigs whose alignments have exceeded the absolute contig length by >5% (Fig. 1A, i). Although the majority of such contigs were observed in satellite DNA, we identified 391 contigs mapping outside of centromeric satellites, of which ~98% are associated with SDs (Fig. 5E; Supplemental Fig. S24). Although we cannot exclude the possibility that such unusual patterns of homology result from assembly error or the inability of mapping algorithms (such as minimap2) to distinguish between paralogous sequences owing to high sequence identity (e.g., *SMN1/2* region) (Supplemental Fig. S25), complete haplotype sequence and assembly of these regions is likely to provide new insights into patterns of human genetic variation and the mutational processes that shape them (Vollger et al. 2023).

Discussion

The recently released gapless assembly of the first haploid human genome has set the bar for T2T human genome assemblies (Nurk et al. 2022). Extending this to diploid samples requires a detailed analysis of the remaining gaps to guide new developments in both sequencing technology and assembly algorithms. Using multiple metrics, we provide a genome-wide assessment to characterize the nature of these last gaps of the human genome. There are several important conclusions. First, we show that recent improvements in sequencing technology (CLR vs. CCS) and assembly algorithms (Peregrine vs. hifiasm) reduce the number of gaps by approximately threefold. Second, the use of parental Illumina WGS data improves phased genome assembly, but the use of linked-read data such as Strand-seq or newer versions of hifiasm that incorporate Hi-C data (Cheng et al. 2021), which is much more widely available than Strand-seq, can create phased assemblies with comparable low levels of switch error. Nevertheless, both trio-based and trio-free assemblies fail to correctly resolve the orientation of 6–7 Mbp of DNA. This is especially the case for large inversion polymorphisms that are flanked by high-identity SDs, which represent one of the most difficult SV classes to accurately assemble (Chaisson et al. 2019; Porubsky et al. 2022). Such complex regions of the genome often coincide with morbid CNVs, where the critical region toggles from a direct to an inverted configuration as a result of recurrent NAHR events (Porubsky et al. 2022).

The current state-of-the-art human genome assembly is represented by approximately 140 gaps per haploid genome with about double the number when trio-free approaches, such as PGAS (Porubsky et al. 2021), are applied. Predictably, gaps cluster within copy number–variable repeat-rich locations corresponding to the largest and most identical repeats (including satellites and SDs) or within low-complexity regions enriched in GA/AT dinucleotides. The latter results from sequence coverage dropouts particular to the HiFi data type over these low-complexity regions (Nurk et al. 2020). Notably, the degree of dropout shows some dependence on the size of the dinucleotide tracts, with the majority of assembled low-complexity regions <6 kbp (Fig. 4A). Many of these regions appear to have expanded specifically in the human–primate lineage so different regions are anticipated in other nonhuman genomes. Our analysis predicts that increasing sequence coverage from 25- to 50-fold eliminates approximately two-thirds of such gaps. Although it does not totally eliminate HiFi-based errors, it has the net effect of also increasing the final base-pair accuracy. In contrast, increasing sequence coverage seems to have little effect on gaps associated with CNV SDs (Fig. 3). This is likely a consequence of the fact that insert size and sequence coverage are inversely correlated, and as a result, high-coverage samples suffer from smaller inserts that fail to resolve large SDs. In this regard, it is interesting that alternate long-read sequencing platforms, such as ONT, do not show the same inherent coverage biases toward GA/AT low-complexity repeats (Nurk et al. 2022). We estimate that, coupled with their much longer read lengths (>50 kbp), ~64% of the remaining gaps within HiFi assemblies can be traversed by ONT (Methods) (Supplemental Fig. S26). Approaches and assembly algorithms that couple both ONT and HiFi data (e.g., Verkko) (Rautiainen et al. 2023) show considerable promise in closing the remaining gaps necessary to achieve routine T2T assemblies of human genomes. The costs of generating deep long-read sequence coverage from two platforms to generate T2T human genomes are, to date, still prohibitively high (more than $10,000), although recently announced increases in throughput from PacBio may reduce this by more than a factor of four.

One of the largest gains from T2T assemblies will be an improved understanding of human structural genetic diversity. Although still incomplete, our analysis identifies ~6.6% and 0.06% of unaligned bases per haploid assembly localized within and outside of centromeric satellite DNA, respectively. Among such gaps caused by contig alignment discontinuities, we identify 230 regions that occurred in at least five haploid assemblies. Nearly half of these (~40%) map to SDs where variation and incomplete assembly pose particular challenges to alignment as well as interpretation. For example, within euchromatic regions, we identified ~15.4 Mbp of sequence per haplotype with two or more mappings per haplotype. Based on Illumina read-depth analysis, we estimate that 86% of these additional alignments represent bona fide human copy number variation. Nevertheless, ~1.6 Mbp of the reported extra alignments are likely false as there is no support in short-read data. Of note, such alignments are often represented by contigs embedded within other larger contigs where one of the overlapping contig alignments has lost allelic variation and now carries, instead, the allelic pattern of variation of the opposing parental haplotype. Allelic variation is, however, still present but maps to only one of the contigs (mostly the shorter one) generated by trio-hifiasm for a given haplotype. This is important because current variant-calling algorithms, such as dipcall or PAV, tend to pick the longer, more contiguous contig in both haploid assemblies to infer allelic variation. We predict that such artifacts may overestimate the amount of loss-of-heterozygosity regions when the longer contig devoid of SNVs is preferentially used. These artifacts also argue that application of state-of-the art methods still requires careful curation and clean-up before their release as new references. It emphasizes the importance of assembly validation using orthogonal data sets such as short reads, optical mapping technology, or Strand-seq to flag remaining errors.

A major challenge going forward will be not only to fully sequence resolve these regions but also to represent complex SVs in such a way that they can be reliably interpreted and assayed in human genetic studies. One of the main objectives of the HPRC efforts is to project all human genome variation through a graph-based representation in which every human haplotype represents

a path in the graph. Unfortunately, there are regions in current genome assemblies that are still completely missing or incorrectly assembled or that otherwise pose challenges for the construction of such pangenome graphs. A set of regions, termed "brnn" regions, were identified and "trimmed" during the construction of the minigraph-cactus graph (Liao et al. 2023). These regions were excluded at least once but, in some instances, up to 88 times and mapped predictably to satellite DNA (~149.7 Mbp), acrocentric (~28.9 Mbp), and SD (~65.7 Mbp) regions and also contain protein-coding genes (n = 171) as well as common inversion polymorphisms (n = 49) (Supplemental Figs. S27–S30; Supplemental Table S9; Supplemental Notes). Here, the challenge will be not only to finish these regions but also to represent changes in meaningful ways such that ectopic exchange events among acrocentric short arms (Guarracino et al. 2023), interlocus gene conversion among SDs (Vollger et al. 2023), hypermutability, and saltatory amplifications in satellite DNA (Logsdon et al. 2021; Altemose et al. 2022) can be adequately captured. Alternate graph-based approaches, such as PGGB (Garrison et al. 2023), hold tremendous promise in this regard, but true representation of such diversity requires a fundamental understanding of the mutational processes that have shaped these regions. Therefore, teasing apart the inheritance status of complex structural variants at the familial level (Noyes et al. 2022) and a better understanding and characterization of the rate of mutational processes such as interlocus gene conversion, recurrent mutation, and duplicative transpositions based on both pedigree and population-level analyses are key (Porubsky et al. 2022; Vollger et al. 2023). Such an understanding will facilitate the development of mutation-aware alignment tools and pangenome graphs in the future.

Methods

Set of evaluated de novo assemblies

De novo assemblies evaluated in this study were obtained from two different sources as part of two international consortia: HGSVC and HPRC. For HGSVC data, we evaluated a panel of 35 samples of diverse ancestry (AFR, 11; AMR, 5; EUR, 7; EAS, 7; SAS, 5). Of those, there are 30 and 14 samples with PacBio CLR and CSS data, respectively (nine samples, or three trios, have both CLR and CCS data). In the HPRC assembly collection, there are 47 samples of mostly African and American ancestry (AFR, 24; AMR, 16; EUR, 1; EAS, 5; SAS, 1) sequenced using PacBio CCS data only. Of those there are five samples also assembled by HGSVC (HG00733, NA19240, HG02818, HG03486, and NA24385/HG002). This accounts for a total of 77 unique samples (35 from HGSVC and 42 from HPRC). We note that the PGAS assembly pipeline at the final step splits long-read (CLR or CCS) data into two haplotype-specific sets that are then assembled separately into haplotype-resolved assemblies (Porubsky et al. 2021).

Alignment of de novo assemblies to the reference genome

Alignments used for simple contig end evaluation

All de novo assemblies were aligned to the most complete version of the human reference genome T2T-CHM13 (v1.1) using minimap2 (v2.22.0; Li et al. 2018) with the following command:

```
minimap2 -K 8G -t {threads} -ax asm20 \
--secondary=no --eqx -s 25000 \
{input.ref} {input.query} \
| samtools view -F 4 -b - > {output.bam}
```

We note that minimap2 had a known issue in which some inversions were missed if they were part of another alignment. To alleviate this issue, we realigned the assemblies with the same parameters after hard masking the reference and query sequence to remove regions that were already aligned in the first alignment step. A complete pipeline for this reference alignment is available at GitHub (https://github.com/mrvollger/asm-to-reference-alignment).

The T2T-CHM13 (v1.1) reference assembly can be found at the NCBI Genome database (https://www.ncbi.nlm.nih.gov/data-hub/genome/) under accession number GCA_009914755.3.

Alignments used for contig alignment end evaluation

All de novo assemblies were aligned to the most complete version of the human reference genome T2T-CHM13 (v1.1) using a newer minimap2 version (v2.24.0) with the following command:

```
minimap2 -K 8G -t {threads} -x asm20 \
--secondary=no --eqx -s 25000 \
{input.ref} {input.query} \
| samtools view -F 4 -b - > {output.bam}
```

A complete pipeline for this reference alignment is available at GitHub (https://github.com/mrvollger/asm-to-reference-alignment).

Evaluation of simple contig ends

Contig ends are defined at the first and last aligned base for each contig in the HPRC haplotype-phased assemblies. Alignments were performed as described above, and the terminal position of each contig was determined using rustybam liftOver (https://github.com/mrvollger/rustybam). A complete pipeline for identifying contigs ends is included at GitHub (https://github.com/mrvollger/asm-to-reference-alignment).

Reading in minimap2 alignments

All minimap2 alignments reported in PAF format were loaded in a set of genomic ranges using custom R (R Core Team) function "paf2ranges" with following given parameters: min.mapq = 10, min.aln.width = 1000, min.ctg.size = 100,000, report.ctg.ends = TRUE, min.ctg.ends = 50,000. At this step, we kept alignments with mapping quality equal to or more than 10 and of minimal size, 1 kbp. Also, contigs with a total size <100 kbp were filtered out.

Evaluation of contig alignment ends

After loading all minimap2 alignments, we extracted terminal contig alignments of at least 50 kbp. When a total alignment size of a contig to the reference was >5% of an actual contig size, we split such contigs into more than one alignment with its own alignment ends. Such splits occur in situations in which the end of the contig maps to distal SD pairs or maps across the centromere, thus increasing the mapped contig size with respect to real contig size.

Defining genomic regions between contig ends and discontinuities within each contig

With minimap2 alignments loaded in a set of genomic ranges, we set out to determine genomic regions spanning between them. For this, we used a custom R function ("reportGaps") in order to report genomic ranges between subsequent contig end mappings.

Strand-seq data generation and data processing

Strand-seq data for eight human samples (HG01123, HG01258, HG01358, HG01361, HG01891, HG02257, HG02486, and HG0 2559) were generated as follows. EBV-transformed lymphoblastoid cell lines from the 1 KG (1000 Genomes Project Consortium 2015) (Coriell Institute) were cultured in BrdU (100 µM final concentration; Sigma-Aldrich B9285) for 18 or 24 h, and single isolated nuclei (0.1% NP-40 substitute lysis buffer) (Sanders et al. 2017) were sorted into 96-well plates using the BD FACSMelody cell sorter. In each sorted plate, 94 single cells plus one 100-cell positive control and one zero-cell negative control were deposited. Strand-specific DNA sequencing libraries were generated using the previously described Strand-seq protocol (Falconer et al. 2012; Sanders et al. 2017) and automated on the Beckman Coulter Biomek FX P liquid handling robotic system (Sanders et al. 2020). Following 15 rounds of PCR amplification, 288 individually barcoded libraries (amounting to three 96-well plates) were pooled for sequencing on the Illumina NextSeq 500 platform (MID-mode, 75-bp paired-end protocol).

The demultiplexed FASTQ files were aligned to the T2T-CHM13 (v1.1) reference assembly using BWA aligner (v0.7.17-r1188) (Li and Durbin 2010) and SAMtools (v1.10) (Li et al. 2009). Duplicate reads were marked using sambamba (v1.0) (Tarasov et al. 2015). Low-quality libraries were excluded from future analyses if they showed low read counts, uneven coverage, or an excess of "background reads" yielding noisy single-cell data, as previously described (Porubský et al. 2016; Sanders et al. 2017). Aligned BAM files were used for assembly evaluations as described below.

Evaluation of assembly quality using Strand-seq

For a set of eight HPRC samples (HG01123, HG01258, HG01358, HG01361, HG01891, HG02257, HG02486, HG02559) for which corresponding Strand-seq data are available, we evaluated the directional and structural contiguity of such assemblies.

Evaluation of misorientations and unresolved homozygous inversions

To evaluate any changes in orientation, we first processed each selected Strand-seq library using breakpointR with the following parameters: windowsize = 2,000,000, binMethod = "size," pairedEndReads = TRUE, min.mapq = 10, genoT = "binom," background = 0.1, minReads = 100. Next, we created so-called composite files that concatenate directional reads across all libraries using breakpointR function "synchronizeReadDir." We set to detect any changes in directionality by running breakpointR on such composite files with the following parameters: windowsize = 10,000, binMethod = "size," pairedEndReads = FALSE, genoT = "binom," background = 0.1, peakTh = 0.25, minReads = 50. Misorientation and unresolved homozygous inversions are reported as regions with the majority of reads mapped in minus orientation ("ww," Watson–Watson strand state), whereas one would expect all Strand-seq reads to map in plus orientation ("cc," Crick–Crick strand state) if the assembly is correctly oriented throughout each contig.

Evaluation of phasing accuracy for selected PGAS assemblies

We evaluated phasing accuracy for HPRC samples (HG01123, HG01258, HG01358, HG01361, HG01891, HG02257, HG02486, HG02559) for which corresponding Strand-seq data are available, and thus, both HPRC and PGAS assemblies could be produced. In this analysis, we consider trio-based HPRC assemblies as the gold standard for phasing evaluation. We used PAV (v1.1.2) to call SNVs in phased HPRC assemblies as described previously (Ebert et al. 2021). To search for large-scale switch errors, we split phased PGAS assemblies into 1-Mbp-long chunks. Subsequently, we used WhatsHap (v1.0) (Patterson et al. 2015) to assign each 1-Mbp chunk to either haplotype 1 or 2 based on a trio-based set of phased SNVs. For each sample, we evaluated a fraction of wrongly assigned 1-Mbp segments separately for haplotype 1 and 2 across all autosomes. Visually, we detected two large-scale switch errors on Chromosome 9 in sample HG01891. There was one switch error around position 42 Mbp (near the centromere); the other, near the end of Chromosome 9 at position 137.3 Mbp.

Evaluation of inversion resolution for selected PGAS assemblies

To evaluate the performance of trio-based and trio-free assemblies to resolve inversion, we selected a set of large inversions (≥100 kbp) from the previous study (Porubsky et al. 2022). We mapped inversion coordinates from GRCh38 to T2T-CHM13 (v1.1) coordinates using minimap2 (v2.20) using following parameters: --secondary=no --eqx -ax asm20 -r 100,1k -z 10000,50. We selected a set of 20 inverted sites (≥100 kbp) with a clear Strand-seq inversion pattern. For dotplot visualization purposes, we added extra padding on each side of the inversion equal to the size of the inversion or minimum of 2 Mbp. We extracted assembly alignments to the reference T2T-CHM13 (v1.1) corresponding to these regions from each trio-based and trio-free phased assembly using rustybam (v0.1.27) function "liftover." Next, we exported a FASTA file from each assembly based on subsetted region-specific PAF files. We used NUCmer (MUMmer v3.23; Delcher et al. 2002) with the parameters --mum --coords to align each FASTA file to the reference sequence (T2T-CHM13 v1.1). We visualized alignments for each assembly in each inverted region as dotplot. Each dotplot was evaluated manually. Inversion was deemed to be resolved if an inversion can be traced in a single contig in both haplotypes and if the inversion status in both haplotypes matches the reported inversion genotype presented by Porubsky et al. (2022).

Definition of centromeric satellite DNA

In this study, centromeric satellite DNA was defined based on T2T-CHM13 annotation obtained from UCSC Table Browser. Annotation was obtained for T2T-CHM13 (v1.1) reference from the annotation group "centromeres and telomeres" and annotation track "CenSat annotation." We define centromeric satellite DNA as regions annotated as human-satellites (hsat), beta-satellites (bsat), and alpha-satellites HOR array (hor).

Protein-coding gene annotations

Gene annotation used in this study is based on T2T-CHM13 annotation obtained from UCSC Table Browser. Annotation was obtained for T2T-CHM13 (v1.1) reference from the annotation group "genes" and annotation track "CAT genes+LiftOff V4." When reporting gene overlap, we selected only protein-coding genes. Any T2T-CHM13-specific genes were not considered. Lastly, subsequent ranges of the same gene were collapsed.

Evaluation of ONT alignments

Available ONT reads (obtained from the NCBI BioProject database [https://www.ncbi.nlm.nih.gov/bioproject/] under accession number PRJNA731524) for 33 HPRC samples were aligned to the T2T-CHM13 (v1.1) reference assembly using minimap2 (v2.24) and filtered secondary alignments using SAMtools (v1.9). We ran the alignments with the following parameters:

```
minimap2 -a -t {threads} -I 10G -Y -x map-ont {assembly} {fastq} | samtools view -u -F 256 - | samtools sort -o {bam_name} -
```

Obtained alignments were exported as read alignment positions in BED format. Only reads with mapping quality 10 or greater were retained for further analysis. We tested each reported assembly gap region per sample and per haplotype if such a region is spanned by 10 or more ONT reads to assume that such assembly gap could eventually be closed by underlying ONT reads. The download locations for ONT data are also reported in Supplemental Table S10.

Low-complexity regions among frequent assembly breaks

Out of the total 592 defined frequent assembly breaks, we extracted 44 regions where there is an assembly break in half or more of the HPRC assemblies. Next, we extracted the T2T-CHM13 FASTA sequence corresponding to these regions (n = 44). We calculated the total number of three dinucleotides (TA, TC, and GA) in nonoverlapping 100-bp-long sequence chunks (bins). To define dinucleotide-enriched bins, we transformed binned dinucleotide counts into the *Z*-scores and marked bins with *Z*-score ≥1.96 (95% confidence interval) as dinucleotide enriched. The size of dinucleotide tracts was estimated as the number of enriched bins × 100 (bin size).

We also investigated FASTA sequence from the previously defined regions (n = 44) in all HPRC assemblies along with nonhuman primate assemblies (n = 18). We processed only those assemblies that span defined regions in a single contig and map to defined breakpoints in T2T-CHM13 coordinates (±100 bp). Next, we transformed observed dinucleotide counts into *Z*-scores as outlined above. Based on visual inspection, we selected 27/44 regions with observable differences in the size of dinucleotide tracts between human and nonhuman primate assemblies (Supplemental Table S6).

Defining regions of putative structural variation

We examined large contig alignment discontinuities as gaps within a single contig alignment that are <1 Mbp. We classified a contig alignment discontinuity as a "contraction" if the alignment gap (in target sequence coordinates) is larger than the corresponding gap within a contig (in query sequence coordinates) (Fig. 1A, iii). In contrast, we classified a contig alignment discontinuity as an "expansion" if the alignment gap (in target sequence coordinates) is smaller than the corresponding gap within a contig (in query sequence coordinates). The number of unaligned bases is defined as the size of the gap in query sequence coordinates. The predicted size of the contractions and expansions was defined as a difference in size between gap in target and query coordinates. We marked contig alignment discontinuities that are within or close (±1 Mbp) to centromeric satellite DNA (marked as "CENSAT") because contig assemblies and alignments within and near centromeres are complicated by the repetitive nature of centromeric satellites and high degree of SDs in these regions. We summarized predicted sites of contraction and expansion into a set of nonredundant regions constructed from sites where contraction and expansion are observed in at least five assemblies and the predicted event size is ≥100 bp (Supplemental Table S7).

Detection of CNV regions

To define regions that are likely copy number variable in any given sample, we searched for regions where there are overlapping contig alignments with respect to the T2T-CHM13 (v1.1) reference. In this analysis, we considered only autosomes, and we filtered out regions that overlap centromeric satellites. We opt to validate putative CNV regions using short-read-based copy number profiles obtained for 44/47 HPRC samples. Short Illumina reads were computationally parsed into 36-bp segments and aligned to a hard-masked T2T-CHM13 (v1.1) reference using mrsFAST (Hach et al. 2010), allowing an edit distance of two. Read-depth-based copy number estimates were generated using the FastCN (Pendleton et al. 2018) software package, which uses known copy number stable regions to correct for Illumina sequencing GC bias and convert read depth to diploid copy number over windows containing 500 unmasked base pairs.

Because of the mapping of short reads to a single paralogous copy in the genome, we set out to determine sample-specific copy number by establishing reference copy number of paralogous regions in T2T-CHM13 (v1.1). We did this by splitting T2T-CHM13 (v1.1) sequence into the same 36-bp subsequences with a slide of one to cover all *k*-mers in the reference. These *k*-mers were mapped back to the reference using mrsFAST, and copy number was determined via FastCN. We refer to this as the *k*-mer-ized T2T-CHM13 reference copy number.

We defined sample-specific CNV regions as those with a diploid copy number less than 10 and at least one diploid copy number increase compared with the *k*-mer-ized T2T-CHM13 reference copy number. Sample-specific regions with a diploid copy number of two and/or no difference (delta = 0) from the *k*-mer-ized T2T-CHM13 reference copy number were defined as not copy number variable and marked as "noCN." Regions where there is an observable diploid copy number increase yet the overall sample-specific copy number is greater than 10 were marked as "more10CN." Regions that do not fall into any of the above categories were marked as "none."

Analysis of pangenome brnn regions

Genomic regions that were excluded from the T2T-CHM13-based pangenome graph construction were obtained from GitHub (https://github.com/human-pangenomics/hpp_pangenome_resources#masked-sequenc). A detailed description of how these regions were defined is reported in the link above. We next took the file "hprc-v1.0-mc-chm13.clipped-intervals.bed.gz," and for each genomic region, we extracted the FASTA sequence from a corresponding phased assembly. We then aligned these to the T2T-CHM13 (v1.1) reference using minimap2 (v2.24) with the following parameters: --secondary=no --eqx -ax asm20 -r 100,1k -z 10000,50. Finally, we kept only alignments of minimum mapping quality of 10 or more and also excluded any alignments from mitochondrial DNA.

Generation of DeepVariant single-nucleotide polymorphism calls for false loss of heterozygosity detection

Alignments of raw PacBio HiFi reads (from seven samples: HG02486, HG02572, HG02622, HG02886, HG03516, HG03540, HG03579) to T2T-CHM13 (v1.1) were made with pbmm2 (https://github.com/PacificBiosciences/pbmm2) using the "CCS" preset. DeepVariant calls were generated using DeepVariant (v1.4.0) (Poplin et al. 2018) and the "PACBIO" pretrained model. PacBio HiFi reads are available at the NCBI BioProject database under the accession number PRJNA731524.

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJEB54100. The DeepVariant callsets for selected samples (HG02486, HG02572, HG02622, HG02886, HG03516, HG03540, HG03579) and FASTA sequences from selected low-complexity regions

(n = 27) are available at Zenodo (https://doi.org/10.5281/zenodo.7392259) or at the IGSR FTP site (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/publications/202212_Porubsky _GenomeResearch). All custom scripts are available in the Supplemental Code and at Zenodo (https://doi.org/10.5281/zenodo.7392259).

Human Pangenome Reference Consortium (HPRC)

Haley J. Abel,[12] Lucinda L. Antonacci-Fulton,[13] Mobin Asri,[14] Gunjan Baid,[15] Carl A. Baker,[16] Anastasiya Belyaeva,[15] Konstantinos Billis,[17] Guillaume Bourque,[18,19,20] Silvia Buonaiuto,[21] Andrew Carroll,[15] Mark J.P. Chaisson,[22] Pi-Chuan Chang,[15] Xian H. Chang,[14] Haoyu Cheng,[23,24] Justin Chu,[23] Sarah Cody,[13] Vincenza Colonna,[21,25] Daniel E. Cook,[15] Robert M. Cook-Deegan,[26] Omar E. Cornejo,[27] Mark Diekhans,[14] Daniel Doerr,[28,29] Peter Ebert,[28,29,30] Jana Ebler,[28,29] Evan E. Eichler,[16,31] Jordan M. Eizenga,[14] Susan Fairley,[17] Olivier Fedrigo,[32] Adam L. Felsenfeld,[33] Xiaowen Feng,[23,24] Christian Fischer,[25] Paul Flicek,[17] Giulio Formenti,[32] Adam Frankish,[17] Robert S. Fulton,[13,34] Yan Gao,[35] Shilpa Garg,[36] Erik Garrison,[25] Nanibaa' A. Garrison,[37,38,39] Carlos Garcia Giron,[17] Richard E. Green,[40,41] Cristian Groza,[42] Andrea Guarracino,[25,43] Leanne Haggerty,[17] Ira M. Hall,[44,45] William T. Harvey,[16] Marina Haukness,[14] David Haussler,[14,31] Simon Heumos,[46,47] Glenn Hickey,[14] Kendra Hoekzema,[16] Thibaut Hourlier,[17] Kerstin Howe,[48] Miten Jain,[49] Erich D. Jarvis,[31,32,50] Hanlee P. Ji,[51] Eimear E. Kenny,[52] Barbara A. Koenig,[53] Alexey Kolesnikov,[15] Jan O. Korbel,[17,54] Jennifer Kordosky,[16] Sergey Koren,[55] HoJoon Lee,[51] Alexandra P. Lewis,[16] Heng Li,[23,24] Wen-Wei Liao,[44,45,56] Shuangjia Lu,[44] Tsung-Yu Lu,[22] Julian K. Lucas,[14] Hugo Magalhães,[28,29] Santiago Marco-Sola,[57,58] Pierre Marijon,[28,29] Charles Markello,[14] Tobias Marschall,[28,29] Fergal J. Martin,[17] Ann McCartney,[55] Jennifer McDaniel,[59] Karen H. Miga,[14] Matthew W. Mitchell,[60] Jean Monlong,[14] Jacquelyn Mountcastle,[32] Katherine M. Munson,[16] Moses Njagi Mwaniki,[61] Maria Nattestad,[15] Adam M. Novak,[14] Sergey Nurk,[55] Hugh E. Olsen,[14] Nathan D. Olson,[59] Benedict Paten,[14] Trevor Pesout,[14] Adam M. Phillippy,[55] Alice B. Popejoy,[62] David Porubsky,[16] Pjotr Prins,[25] Daniela Puiu,[63] Mikko Rautiainen,[55] Allison A. Regier,[13] Arang Rhie,[55] Samuel Sacco,[64] Ashley D. Sanders,[65] Valerie A. Schneider,[66]

[12]Division of Oncology, Department of Internal Medicine, Washington University School of Medicine, St. Louis, MO 63110, USA
[13]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO 63108, USA
[14]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA 95064, USA
[15]Google LLC, Mountain View, CA 94043, USA
[16]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA
[17]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[18]Department of Human Genetics, McGill University, Montreal, Québec H3A 0C7, Canada
[19]Canadian Center for Computational Genomics, McGill University, Montreal, Québec H3A 0G1, Canada
[20]Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto 606-8501, Japan
[21]Institute of Genetics and Biophysics, National Research Council, Naples 80111, Italy
[22]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA
[23]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[24]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02215, USA
[25]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN 38163, USA
[26]Arizona State University, Barrett and O'Connor Washington Center, Washington, DC 20006, USA
[27]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA 95064, USA
[28]Institute for Medical Biometry and Bioinformatics, Medical Faculty, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany
[29]Center for Digital Medicine, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany
[30]Core Unit Bioinformatics, Medical Faculty, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany
[31]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
[32]Vertebrate Genome Laboratory, The Rockefeller University, New York, NY 10065, USA
[33]National Institutes of Health (NIH)–National Human Genome Research Institute, Bethesda, MD 20892, USA
[34]Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA
[35]Center for Computational and Genomic Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
[36]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Copenhagen DK-2200, Denmark
[37]Institute for Society and Genetics, College of Letters and Science, University of California, Los Angeles, Los Angeles, CA 90095, USA
[38]Institute for Precision Health, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA
[39]Division of General Internal Medicine and Health Services Research, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA
[40]Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA 95064, USA
[41]Dovetail Genomics, Scotts Valley, CA 95066, USA
[42]Quantitative Life Sciences, McGill University, Montreal, Québec H3A 0C7, Canada
[43]Genomics Research Centre, Human Technopole, Milan 20157, Italy
[44]Department of Genetics, Yale University School of Medicine, New Haven, CT 06510, USA
[45]Center for Genomic Health, Yale University School of Medicine, New Haven, CT 06510, USA
[46]Quantitative Biology Center (QBiC), University of Tübingen, 72076 Tübingen, Germany
[47]Biomedical Data Science, Department of Computer Science, University of Tübingen, 72076 Tübingen, Germany
[48]Tree of Life, Wellcome Sanger Institute, Hinxton, Cambridge CB10 1SA, UK
[49]Northeastern University, Boston, MA 02115, USA
[50]Laboratory of Neurogenetics of Language, The Rockefeller University, New York, NY 10065, USA
[51]Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA
[52]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[53]Program in Bioethics and Institute for Human Genetics, University of California, San Francisco, San Francisco, CA 94143, USA
[54]European Molecular Biology Laboratory, Genome Biology Unit, 69117 Heidelberg, Germany
[55]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA
[56]Division of Biology and Biomedical Sciences, Washington University School of Medicine, St. Louis, MO 63110, USA
[57]Computer Sciences Department, Barcelona Supercomputing Center, 08034 Barcelona, Spain
[58]Departament d'Arquitectura de Computadors i Sistemes Operatius, Universitat Autònoma de Barcelona, 08193 Barcelona, Spain
[59]Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20877, USA
[60]Coriell Institute for Medical Research, Camden, NJ 08103, USA
[61]Department of Computer Science, University of Pisa, Pisa 56127, Italy
[62]Department of Public Health Sciences, University of California, Davis, Davis, CA 95616, USA
[63]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA
[64]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA 95064, USA

Baergen I. Schultz,[33] Kishwar Shafin,[15] Jonas A. Sibbesen,[67] Jouni Sirén,[14] Michael W. Smith,[33] Heidi J. Sofia,[33] Ahmad N. Abou Tayoun,[68,69] Françoise Thibaud-Nissen,[66] Chad Tomlinson,[13] Francesca Floriana Tricomi,[17] Flavia Villani,[25] Mitchell R. Vollger,[16,70] Justin Wagner,[59] Brian Walenz,[55] Ting Wang,[13,34] Jonathan M.D. Wood,[48] Aleksey V. Zimin,[63,71] and Justin M. Zook[59]

Competing interest statement

E.E.E. is a scientific advisory board (SAB) member of Variant Bio. The following authors have previously disclosed a patent application (no. EP19169090) relevant to Strand-seq: A.D.S., J.O.K., T.M., and D.P. The other authors declare no competing interests.

Acknowledgments

We thank Tonia Brown for assistance in editing this manuscript. T.M. and P.E. acknowledge the support and computational infrastructure provided by the Centre for Information and Media Technology at Heinrich Heine University Düsseldorf. This work was supported, in part, by grants from the National Institutes of Health (NIH; grants 5R01HG002385 and 5U01HG010971 to E.E.E. and 1U01HG010973 to E.E.E. and T.M.) and by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 956229 to T.M. E.E.E. is an investigator of Howard Hughes Medical Institute. This article is subject to HHMI's Open Access to Publications policy. HHMI laboratory heads have previously granted a nonexclusive CC BY 4.0 license to the public and a sublicensable license to HHMI in their research articles. Pursuant to those licenses, the author-accepted manuscript of this article can be made freely available under a CC BY 4.0 license immediately upon publication.

Author contributions: Conceptualization and design were by D.P., M.R.V., and E.E.E. Assembly gap, inversion resolution, and structurally complex region analysis was by D.P. Contig end enrichment analysis was by M.R.V. Production of the HGSVC assemblies was by P.E. and T.M. Production of the HPRC assemblies was by the HPRC group. Strand-seq data generation was by P.H., A.D.S., C.S., and J.O.K. Bioinformatics support was by W.T.H. and A.N.R. Organization of the tables and Supplemental Materials was by D.P. Display items were by D.P. and M.R.V. Resources were by HPRC, G.H., B.P., and E.E.E. Manuscript writing was by D.P., M.R.V., and E.E.E., with input from all authors.

References

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526:** 68–74. doi:10.1038/nature15393

Altemose N, Logsdon GA, Bzikadze AV, Sidhwani P, Langley SA, Caldas GV, Hoyt SJ, Uralsky L, Ryabov FD, Shew CJ, et al. 2022. Complete genomic and epigenetic maps of human centromeres. *Science* **376:** eabl4178. doi:10.1126/science.abl4178

Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. 2022. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185:** 3426–3440.e19. doi:10.1016/j.cell.2022.08.004

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10:** 1784. doi:10.1038/s41467-018-08148-z

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18:** 170–175. doi:10.1038/s41592-020-01056-5

Cheng H, Jarvis ED, Fedrigo O, Koepfli K-P, Urban L, Gemmell NJ, Li H. 2022. Haplotype-resolved assembly of diploid genomes without parental data. *Nat Biotechnol* **40:** 1332–1335. doi:10.1038/s41587-022-01261-x

Chin C-S, Khalak A. 2019. Human genome assembly in 100 minutes. bioRxiv doi:10.1101/705616

Coe BP, Witherspoon K, Rosenfeld JA, van Bon BWM, Vulto-van Silfhout AT, Bosco P, Friend KL, Baker C, Buono S, Vissers LELM, et al. 2014. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet* **46:** 1063–1071. doi:10.1038/ng.3092

Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, et al. 2011. A copy number variation morbidity map of developmental delay. *Nat Genet* **43:** 838–846. doi:10.1038/ng.909

Delcher AL, Phillippy A, Carlton J, Salzberg SL. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* **30:** 2478–2483. doi:10.1093/nar/30.11.2478

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, et al. 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372:** eabf7117. doi:10.1126/science.abf7117

Ebler J, Ebert P, Clarke WE, Rausch T, Audano PA, Houwaart T, Mao Y, Korbel JO, Eichler EE, Zody MC, et al. 2022. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat Genet* **54:** 518–525. doi:10.1038/s41588-022-01043-w

Falconer E, Hills M, Naumann U, Poon SSS, Chavez EA, Sanders AD, Zhao Y, Hirst M, Lansdorp PM. 2012. DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat Methods* **9:** 1107–1112. doi:10.1038/nmeth.2206

Garg S, Fungtammasan A, Carroll A, Chou M, Schmitt A, Zhou X, Mac S, Peluso P, Hatas E, Ghurye J, et al. 2021. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat Biotechnol* **39:** 309–312. doi:10.1038/s41587-020-0711-0

Garrison E, Guarracino A, Heumos S, Villani F, Bao Z, Tattini L, Hagmann J, Vorbrugg S, Marco-Sola S, Kubica C, et al. 2023. Building pangenome graphs. bioRxiv doi:10.1101/2023.04.05.535718

Guarracino A, Buonaiuto S, de Lima LG, Potapova T, Rhie A, Koren S, Rubinstein B, Fischer C, Human Pangenome Reference Consortium, Gerton JL, et al. 2023. Recombination between heterologous human acrocentric chromosomes. *Nature* **617:** 335–343. doi:10.1038/s41586-023-05976-y

Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, Sahinalp SC. 2010. mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat Methods* **7:** 576–577. doi:10.1038/nmeth0810-576

Jarvis ED, Formenti G, Rhie A, Guarracino A, Yang C, Wood J, Tracey A, Thibaud-Nissen F, Vollger MR, Porubsky D, et al. 2022. Semi-automated assembly of high-quality diploid human reference genomes. *Nature* **611:** 519–531. doi:10.1038/s41586-022-05325-5

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37:** 540–546. doi:10.1038/s41587-019-0072-8

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34:** 3094–3100. doi:10.1093/bioinformatics/bty191

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26:** 589–595. doi:10.1093/bioinformatics/btp698

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079. doi:10.1093/bioinformatics/btp352

Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. 2023. A draft human pangenome reference. *Nature* **617:** 312–324. doi:10.1038/s41586-023-05896-x

[65]Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, 10115 Berlin, Germany
[66]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
[67]Center for Health Data Science, University of Copenhagen, 2200 Copenhagen, Denmark
[68]Al Jalila Genomics Center of Excellence, Al Jalila Children's Specialty Hospital, Dubai, UAE
[69]Center for Genomic Discovery, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai, UAE
[70]Division of Medical Genetics, University of Washington School of Medicine, Seattle, WA 98195, USA
[71]Center for Computational Biology, Johns Hopkins University, Baltimore, MD 21218, USA

Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovykh MA, Koren S, Nurk S, Mercuri L, Dishuck PC, Rhie A, et al. 2021. The structure, function and evolution of a complete human chromosome 8. *Nature* **593:** 101–107. doi:10.1038/s41586-021-03420-7

Lu H, Giordano F, Ning Z. 2016. Oxford Nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinformatics* **14:** 265–279. doi:10.1016/j.gpb.2016.05.004

Noyes MD, Harvey WT, Porubsky D, Sulovari A, Li R, Rose NR, Audano PA, Munson KM, Lewis AP, Hoekzema K, et al. 2022. Familial long-read sequencing increases yield of *de novo* mutations. *Am J Hum Genet* **109:** 631–646. doi:10.1016/j.ajhg.2022.02.014

Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S. 2020. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res* **30:** 1291–1305. doi:10.1101/gr.263566.120

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science* **376:** 44–53. doi:10.1126/science.abj6987

Patterson M, Marschall T, Pisanti N, van Iersel L, Stougie L, Klau GW, Schönhuth A. 2015. WhatsHap: weighted haplotype assembly for future-generation sequencing reads. *J Comput Biol* **22:** 498–509. doi:10.1089/cmb.2014.0157

Pendleton AL, Shen F, Taravella AM, Emery S, Veeramah KR, Boyko AR, Kidd JM. 2018. Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication. *BMC Biol* **16:** 64. doi:10.1186/s12915-018-0535-2

Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, et al. 2018. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* **36:** 983–987. doi:10.1038/nbt.4235

Porubský D, Sanders AD, van Wietmarschen N, Falconer E, Hills M, Spierings DCJ, Bevova MR, Guryev V, Lansdorp PM. 2016. Direct chromosome-length haplotyping by single-cell sequencing. *Genome Res* **26:** 1565–1574. doi:10.1101/gr.209841.116

Porubsky D, Ebert P, Audano PA, Vollger MR, Harvey WT, Marijon P, Ebler J, Munson KM, Sorensen M, Sulovari A, et al. 2021. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat Biotechnol* **39:** 302–308. doi:10.1038/s41587-020-0719-5

Porubsky D, Höps W, Ashraf H, Hsieh P, Rodriguez-Martin B, Yilmaz F, Ebler J, Hallast P, Maria Maggiolini FA, Harvey WT, et al. 2022. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell* **185:** 1986–2005.e26. doi:10.1016/j.cell.2022.04.017

Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, Eichler EE, Phillippy AM, Koren S. 2023. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol* doi:10.1038/s41587-023-01662-6

R Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Sanders AD, Falconer E, Hills M, Spierings DCJ, Lansdorp PM. 2017. Single-cell template strand sequencing by strand-seq enables the characterization of individual homologs. *Nat Protoc* **12:** 1151–1176. doi:10.1038/nprot.2017.029

Sanders AD, Meiers S, Ghareghani M, Porubsky D, Jeong H, van Vliet MACC, Rausch T, Richter-Pechańska P, Kunz JB, Jenni S, et al. 2020. Single-cell analysis of structural variations and complex rearrangements with tri-channel processing. *Nat Biotechnol* **38:** 343–354. doi:10.1038/s41587-019-0366-x

Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, et al. 2020. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol* **38:** 1044–1053. doi:10.1038/s41587-020-0503-6

Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, Hurst JA, Stewart H, Price SM, Blair E, Hennekam RC, et al. 2006. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat Genet* **38:** 1038–1042. doi:10.1038/ng1862

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, 1000 Genomes Project, et al. 2010. Diversity of human copy number variation and multicopy genes. *Science* **330:** 641–646. doi:10.1126/science.1197005

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH-Y, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526:** 75–81. doi:10.1038/nature15394

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. 2015. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31:** 2032–2034. doi:10.1093/bioinformatics/btv098

Vollger MR, Logsdon GA, Audano PA, Sulovari A, Porubsky D, Peluso P, Wenger AM, Concepcion GT, Kronenberg ZN, Munson KM, et al. 2019. Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *Ann Hum Genet* **84:** 125–140. doi:10.1111/ahg.12364

Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, Diekhans M, Sulovari A, Munson KM, Lewis AP, et al. 2022. Segmental duplications and their variation in a complete human genome. *Science* **376:** eabj6965. doi:10.1126/science.abj6965

Vollger MR, Dishuck PC, Harvey WT, DeWitt WS, Guitart X, Goldberg ME, Rozanski AN, Lucas J, Asri M, Abel HJ, et al. 2023. Increased mutation and gene conversion within human segmental duplications. *Nature* **617:** 325–334. doi: 10.1038/s41586-023-05895-y

Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, Phillippy AM, Popejoy AB, Asri M, Carson C, Chaisson MJP, et al. 2022. The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604:** 437–446. doi:10.1038/s41586-022-04601-8

Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37:** 1155–1162. doi:10.1038/s41587-019-0217-9

Received September 19, 2022; accepted in revised form December 7, 2022.