Science
AAAS

Supplementary Materials for

The complete sequence of a human genome

Sergey Nurk *et al*.

Corresponding authors: Evan E. Eichler, eee@gs.washington.edu; Karen H. Miga, khmiga@ucsc.edu; Adam M. Phillippy, adam.phillippy@nih.gov

Science **376**, 44 (2022)
DOI: 10.1126/science.abj6987

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

Materials and Methods
Table of Contents

Molecular Methods

Chromosome spreads and Fluorescent In-Situ Hybridization (FISH)

For the preparation of chromosome spreads, cells were blocked in mitosis by the addition of Karyomax colcemid solution (0.1 μg/ml, Life Technologies) for 6-7h and collected by trypsinization. Collected cells were incubated in hypotonic 0.4% KCl solution for 12 min and pre-fixed by addition of Methanol:Acetic acid (3:1) fixative solution (1% total volume). Pre-fixed cells were collected by centrifugation and then fixed in Methanol:Acetic acid (3:1). Spreads were dropped on a glass slide and incubated at 65°C overnight. Before hybridization, slides were treated with 0.1mg/ml RNAse A (Qiagen) in 2xSSC for 45 minutes at 37°C and dehydrated in a 70%, 80%, and 100% ethanol series for 2 minutes each. Slides were denatured in 70% formamide/2X SSC solution pre-heated to 72°C for 1.5 min. Denaturation was stopped by immersing slides in 70%, 80%, and 100% ethanol series chilled to -20°C. Labeled DNA probes were denatured separately in a hybridization buffer by heating to 80°C for 10 minutes before applying to denatured slides. A fluorescently labeled probe for human rDNA (BAC clone RP11-450E20) was obtained from Empire Genomics (Williamsville, NY). Acrocentric chromosome paints were obtained from Oxford Gene Technology (Cambridge, UK). WaluSat DNA probe was PCR-amplified from CHM13 genomic DNA, labeled with biotin-dUTP by nick-translation reaction, and detected with fluorescently conjugated streptavidin post hybridization. Specimens were hybridized to the probe under a glass coverslip or HybriSlip hybridization cover (GRACE Biolabs) sealed with the rubber cement or Cytobond (SciGene) in a humidified chamber at C37°C for 48-72hours. After hybridization, slides were washed in 50% formamide/2X SSC 3 times for 5 minutes per wash at 45°C, then in 1x SSC solution at 45°C for 5 minutes twice and at room temperature once. Slides were mounted in Vectashield containing DAPI (Vector Laboratories).

Estimating rDNA copy number from FISH images

Wide-field images of chromosomal spreads were acquired on Nikon ECLIPSE Ti2 microscope equipped with Photometrics Prime 95B sCMOS camera and a 100x Plan Apo Lambda 1.45 NA objective. Acrocentric chromosome identities were assigned based on chromosome paints and morphological features. Individual rDNA loci were segmented via a selected threshold, and lines (at least 12 pixels wide) were created across each segmented rDNA locus to generate fluorescent intensity profiles. The averages of three ending values on both ends of each intensity profile were used to subtract the local background. The sum of background-subtracted intensity profile values represented the integrated intensity of each locus. The sum of all integrated intensities of all rDNA loci represented the total amount of rDNA per cell. The fraction of this total fluorescent intensity was calculated for each rDNA locus. The total rDNA copy number was estimated from Illumina sequencing data to be 405 copies per diploid genome (Fig. S9). This number was multiplied by two as mitotic cells have a doubled genome (810 copies). The fraction of the total rDNA fluorescence intensity was used as a fraction of the total rDNA copy number to determine the number of copies on each identifiable chromosome and divided by two to get the final copy number per sister chromatid (Fig. S1). As CHM13 is

homozygous, the rDNA copy numbers are similar on both haplotypes. Chromosomes 14 and 15 are exceptions, with the two haplotypes having a large difference in copy number. The shorter Chromosome 15 haplotype was confirmed by a single spanning ONT UL read (Fig. S2).

Ancestry analysis

We used RFMix v2.03-r0 (https://github.com/slowkoni/rfmix) to infer the local ancestry of the CHM13 genome (*62*). As a set of reference samples for ancestry, we used the 2,504 individuals sequenced in the Phase 3 release of the 1000 Genomes Project (*24*). Phased SNP genotypes for this reference set were generated by the 1000 Genomes Consortium through alignment and calling variants against the GRCh38 reference (*63*), and included only biallelic SNVs (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20181203_biallelic_SNV/). We obtained a genetic map for GRCh38, originally generated on build 35 of the human reference genome by the HapMap Consortium (*64*), from Beagle (http://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/).

CHM13 variants were called on GRCh38 with dipcall (https://github.com/lh3/dipcall) (*65*), through direct alignment of the CHM13 v1.0 assembly to the GRCh38 reference (*25*). For variable sites in the 1000 Genomes dataset that were not called as variants in CHM13, we assumed that CHM13 carried the reference allele. All CHM13 genotypes were represented as phased homozygous diploid genotypes (either reference/reference or alt/alt), which resulted in two sets of identical inferred local ancestry segments for each haplotype. We ran RFMix with default parameters (conditional random field spacing = 50 SNPs; random forest window size = 5 SNPs), grouping the 1000 Genomes reference panel into superpopulations (African, Ad Mixed American, East Asian, European, Southeast Asian). Since males are haploid for variants on the X chromosome, we included only female individuals in the reference panel for local ancestry inference for the X chromosome. Regions that overlapped with centromeres or were marked as inaccessible for the 1000 Genomes Project (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/working/20160622_genome_mask_GRCh38/) were excluded from results (UCSC Genome Browser; https://genome.ucsc.edu/cgi-bin/hgTables). The non-pseudoautosomal regions (PARs) of the X chromosome were inaccessible for this analysis due to the absence of variant calls outside the PARs. For identification of Neanderthal-introgressed haplotypes in CHM13, we used IBDmix (https://github.com/PrincetonUniversity/IBDmix), which detects introgressed archaic sequences in a set of individuals by identifying haplotypes that are identical-by-descent to a reference archaic sample (*66*). Of the three available Neanderthal individuals with high-coverage sequences, Vindija Neanderthal was chosen as the reference archaic individual due to its inferred closest relation to the Neanderthal population that admixed with modern humans (*67*). Local ancestry analysis indicated that the majority of the CHM13 genome is composed of European ancestries; we used European samples from the 1000 Genomes Project as the background set of modern individuals in addition to CHM13. These samples were sequenced to 30x coverage by

the New York Genome Center, and variants were called on the GRCh38 reference (*68*). We required Neanderthal-introgressed haplotypes identified in CHM13 to have a LOD score of at least 4 and a length of at least 50 kb. We removed introgressed segments using bedtools within regions deemed inaccessible by the 1000 Genomes Project.

Whole genome assembly

Limitations of the current approach

Our assembly string graph construction and tangle resolution procedures were specifically targeted to CHM13, which is derived from a hydatidiform mole and is largely homozygous (*18*). Thus, some of our design and heuristic decisions (e.g. arbitrary choice of haplotypes, bubble collapse) may not be applicable to assembling diploid genomes. Manual resolution would also be complicated by more heterozygous genomes due to shorter, more ambiguous nodes resulting from long homozygous regions with no single path traversing all nodes. Lastly, the basic string graph approach as described in Myers 2005 (*26*) has some fundamental deficiencies, arising from the initial removal of contained reads, making it a poor model for assembling datasets with varying read lengths, especially if the underlying genome is diploid/polyploid. See Fig. S11 for an example. These issues were largely avoided in this study due to the relatively uniform size of the available HiFi reads and the homozygous nature of CHM13.

HiFi graph construction

Read processing

PacBio's HiFi sequencing offers a compromise of length with a median accuracy over 99.9% (*15*, *69*). HiFi 20 kbp libraries providing 32.4X average coverage of CHM13 (Table S1) were used to construct a string graph. The graph construction started by applying slightly reshuffled modules of the HiCanu assembler to find and analyze overlaps between HiFi reads. For more details on individual steps see Nurk et al. (*16*). First, the reads were homopolymer-compressed (*70*). Then all >99% identity overlaps (dovetail and contained) longer than 500 bp were identified. Each read was subjected to a conservative correction based on the overlap pile-ups (see "Overlap Error Adjustment" section in (*16*)). Unlike in HiCanu, here we introduced the corrections to the homopolymer-compressed read sequences rather than only adjusting identity scores. We then recomputed overlaps (>500 bp and >99% identity) on the corrected reads and the overlap alignment identity was adjusted by masking differences in microsatellite array regions to account for previously observed increases in errors in these regions in HiFi reads (*16*). The need to apply and recompute overlaps stems from a limitation of HiCanu but is also common in other overlapping tools which limit the number of overlaps between a pair of reads. Currently, HiCanu stores at most two overlaps for every pair of reads (matching and inverted orientation) with higher identity overlaps preferred. In rare cases this can lead to the true overlap being dropped in favor of a (usually) shorter but higher identity overlap. Re-computing overlaps on corrected reads partially mitigated this issue. Additionally, HiCanu's overlap search settings were adjusted (OvlMerThreshold=20000) to correctly identify overlaps within large repetitive arrays (e.g. centromere, rDNA). All overlaps with adjusted identity below 100% were then

discarded. Finally, the BOGART module of (Hi)Canu was used to identify and discard reads with structural errors based on the analysis of the overlap pile-ups (*16*, *71*).

Initial construction

The resulting homopolymer-compressed reads and overlaps were then provided to a modified version of miniasm (*27*) to construct a Myers string graph (with a 1 kbp overlap threshold).

The modification fixed a problem in the pseudocode of the transitive overlap reduction algorithm in Myers 2005 (*26*), which has made its way into the miniasm implementation. The pseudocode was missing a final verification that the identified overlap is actually redundant with respect to the other two overlaps based on their sizes. Without this check, an overlap could be deemed transitive even if it implied an incompatible arrangement of the reads.

Notably, miniasm's GFA format includes information about the read composition of the individual graph nodes corresponding to the non-branching paths (unitigs) in Myers' string graph. Using this information, each node in the graph had a coverage estimate assigned to it to facilitate downstream analysis. Previous methods to assign coverage to string graphs (e.g. the A-stat (*4*)) are not applicable to shorter unitigs consisting of just a few reads. Thus, we implemented a different approach. Each read is assigned a coverage value by considering the read's overlaps and taking a minimum depth across all positions. Each graph node is assigned the median value of all the reads comprising it.

All further string graph processing within miniasm (e.g. graph simplification, tip removal, bubble removal and 'relatively weak' link pruning) was disabled.

Iterative graph simplification

We used custom graph simplification procedures to replace miniasm's graph simplifications. This was implemented as a set of modular tools on top of public GFA libraries (*72–74*) available at https://github.com/snurk/sg_sandbox.

Graph processing steps:

- *Tip removal.* Dead-end nodes shorter than a specified threshold (set to 25 kbp excluding overlaps to an adjacent node) were removed. Initial rounds had a limit on the number of reads forming a tip that is iteratively increased. This minimizes the amount of correct sequence clipped at both the telomeres and "coverage gap" regions.

- *Superbubble collapsing.* Thresholds were set at 2 kbp on both the longest path and the maximal length difference between alternative paths. Each superbubble was replaced with the path having the maximum minimal overlap between its nodes. Note that the aggressive 2 kbp length difference threshold is a legacy of earlier stages of pipeline development. However, it did not affect the representation of the repetitive regions.

Firstly, the limitations of the procedure prevent it from simplifying tangles of high complexity (e.g. most HSats). Secondly, early-processing-stage subgraphs were spot-checked around the nodes that were used multiple times within the traversals in an attempt to verify perfect repeat representation. A notable exception was the HSat3 region on Chromosome 9 where variation between genomic repeat copies was lost. This region spans a recent multi-megabase duplication, leading to chains of distinctive bubbles, representing variation between the two copies. During reconstruction of this region an alternate version of the HiFi string graph was built with the difference threshold set to 5 bp (see "Resolution of Chr6 and Chr9" section).

- *Low coverage node removal.* Removal of short nodes with assigned coverage value below an iteratively increased threshold (2-3-4-5X).

- *'Weak' edge pruning.* Edges corresponding to overlaps smaller than an iteratively increasing (2-4-6-8 kbp) threshold were removed. Before each threshold increase, other procedures (e.g. superbubble collapsing, tip removal etc.) were launched. Notably a handpicked absolute threshold was preferred over the miniasm strategy of pruning the overlaps based on their ratio against the size of the longest overlap for the same (side of the) node (read).

- *'Simple' bulge removal.* Removal of short nodes that have an alternative path of similar length (subject to additional conditions). This step is complementary to superbubble removal in tangled regions, where no clear superbubble subgraph can be identified.

- *Removal of some unusable edges.* An edge was considered unusable if it could not be a part of a traversal which uses all of the 'unique' and reliable (classified based on length and coverage) nodes (of a particular graph region) exactly once. At present, we only covered several cases forming distinct commonly occurring subgraphs, leaving the development of a more general procedure for future work.

The graph is compacted after every procedure by merging the non-branching paths in the graph into longer nodes. The progressive increase of the overlap threshold allows the avoidance of higher levels of repeat entanglement potentially arising from using a small overlap threshold, while leading to better preservation of continuity compared to using a single larger threshold. This strategy is similar in spirit to the progressive increase of the k-mer size, commonly used by de Bruijn graph based assemblers (*75*).

Based on subsequent chromosome reconstruction attempts, the resulting graph was manually adjusted in a few regions in order to restore some edges. In one case, Canu chose to store an overlap inconsistent with the graph, and this overlap was manually replaced with an alternate overlap (Canu can store only one overlap per read pair orientation). With the exception of the fragmentation caused by absent HiFi read coverage, and bubbles representing larger

heterozygous / polymorphic indels, the resulting assembly graph had relatively few artifacts (e.g. redundant or non-genomic nodes, or missing edges). These artifacts reflected minor issues in our processing strategies and parameters as well as inherent deficiencies of string graphs built from reads of varying lengths (see ”Limitations of current approach”).

Final graph overview

To facilitate manual inspection of the graph, all graph nodes were aligned to GRCh38 (*1*) using mashmap (*76*) and assigned tentative chromosome labels (these alignments were not used to inform chromosome reconstruction). This showed that graph regions corresponding to different chromosomes are almost never connected by graph edges. Moreover, the connected components have a mostly linear structure, suggesting that the genome has few perfect repeats longer than 8 kbp (edges shorter than this are pruned during “Weak Edge Pruning”, above) on different chromosomes or at distant positions of the same chromosome. Edges connecting nodes of the acrocentric chromosomes were an exception, potentially pointing to recent recombination events across distal rDNA junction regions. Another exception is in the HSat3 region on Chromosome 9, which spans a recent multi-Mbp tandem duplication, consistent with the 9qh+ karyotype of CHM13 (*77*)(Fig. S3), resulting in edges between the two copies.

Manual graph resolution and gap filling

HiFi-only resolution

Reconstruction of some regions required the use of ONT UL reads (see”ONT-based tangle resolution”). In many cases, genomic paths through tangles could be unambiguously identified by analysis of the graph structure when considering estimated node multiplicities (based on coverage). This included cases where there was only a single traversal visiting every node (Fig. S4A) or the coverage information indicated the appropriate traversal (Fig. S4B). While handling bubbles representing heterozygous / polymorphic indels (see Fig. S4C,D for the example), we generally preferred the longer alternative (we could pick a slightly shorter path if it removed fewer nodes from the graph or if its coverage was at least two times higher than the alternate). Somewhat surprisingly, this approach allowed us to obtain draft reconstructions for tangles consisting of dozens of nodes (Fig. S4C,D) without (or with very limited) usage of the ONT UL. We first considered resolutions of the most complicated tangles (e.g. those representing HSAT regions on Chromosomes 4 and 16) as low confidence, but subsequent ONT and HiFi based validation has mostly supported our initial resolutions.

ONT-based tangle resolution

When graph information alone was insufficient, alignments of homopolymer-compressed Oxford Nanopore ultra-long (ONT UL) reads were used (*14*, *21*) to inform the reconstruction. Here, all available homopolymer-compressed ONT UL reads (longer 100 kbp) were aligned to the complete assembly graph using GraphAligner v1.0.12 (*28*). In some cases, these alignments were sufficient to provide reliable evidence for a particular path through the tangle. Unfortunately, GraphAligner’s paths could not be used to reliably disambiguate traversal of more complex tangles. Moreover, due to GraphAligner limitations some of the highly tangled

regions did not produce any alignments. When multiple alignment paths of similar overall score are present in the graph, GraphAligner has limited guarantees of picking the optimal one. In these cases, a more "brute-force" strategy was implemented that considered all candidate paths from the particular region of the graph. Here, ONT UL reads were aligned to every corresponding sequence using Winnomap v2.0 (*78*, *79*) and identity scores and break-points compared to identify reads exhibiting substantial differences in alignment quality between the candidate reconstructions. Alignment reports for these "informative" reads were then inspected to select the best-supported candidate path (Fig. S5).

Resolution of Chr6 and Chr9

Unfortunately, several regions in the graph could not be resolved using the methods outlined in the previous sections. For instance, the core of the rDNA arrays remained unresolved. Additionally, the Chromosome 6 centromere and Chromosome 9 HSat3 array could not be easily resolved using the HiFi string graph and semi-manual analysis of ONT UL alignments. During Chromosome 6 centromere reconstruction, information about the higher-order repeat organization obtained by CentroFlye (*80*) from ONT UL data was used to inform the choice of graph traversal. For the reconstruction of the Chromosome 9 HSat3 array (involving a large-scale recent duplication) MBG (*33*) was used to build a sparse de Bruijn graph from homopolymer-compressed HiFi reads. This graph was resolved by aligning the HiFi and ONT UL reads, identifying unique nodes based on their length and estimated coverage, and connecting them based on the read alignments. An alternate version of the HiFi string graph was built with the superbubble collapsing threshold set to 5 bp (see 'Initial Construction') to preserve minor differences between the duplicated sequences. This graph was then searched for a traversal corresponding to the draft sequence using GraphAligner.

Consensus and subsequent analysis

Consensus

After identifying the genomic graph traversals associated with a particular chromosome, their sequences were generated following the layout/consensus paradigm.

Throughout the analyses in previous sections, sequences were associated with graph nodes and paths (particularly for ONT UL read alignments). Prior to this point, they were obtained as simple concatenations of the homopolymer-compressed substrings of individual reads (joined according to the overlaps sizes). For the generation of the final consensus sequence, the graph traversal was translated into a backbone read layout—an orientation-aware concatenation of the chains of reads comprising each of the nodes along a path. Then, using procedures from the HiCanu assembler, the contained reads were incorporated into the layout, requiring 100% adjusted overlap identity, with the layout of the raw (uncompressed and uncorrected) read sequences being inferred as in (*16*). The consensus sequence was then computed via HiCanu's POA-graph of the resulting layout (*16*, *71*). Upon validation of intermediate results using TandemTools (*37*), the logic and the parameters were adjusted to produce accurate layouts and consensus, even in highly repetitive genomic regions (in particular

centromeres). A limitation of the Canu codebase, allowing a read to be used only once per layout, forced special handling of the nodes used more than once in the same traversal. The default strategy was to randomly distribute the reads comprising such nodes between their occurrences. In cases where there were not enough (non-contained) reads for this strategy to produce a valid layout, the traversal was split at the duplicated node, consensus sequences produced for each piece, and finally merged.

GA-gaps were filled as in the 'Gap Filling' section. The per-base consensus accuracy (QV) was estimated to be Q70.2 by Merqury (*81*) based on the analysis of 21-mer spectra from the combination of HiFi and PCR-free Illumina reads (after filtering low copy-count 21-mers, for details see (*34*)).

Gap filling

Several chromosomes were represented by multiple disconnected components in the HiFi assembly graph, primarily due to the previously described (*16*) deficiency in HiFi sequencing, leading to depleted (or absent) HiFi coverage in (GA)-rich microsatellite regions. To mitigate the problem, ONT UL read alignments (aligned with winnowmap v1.1 (*78*, *79*)) were used to identify the graph unitigs flanking each coverage gap. After consensus sequences were generated for all recovered graph traversals (see above), gap 'patching' was performed using the previously published v0.7 assembly (*14*) as the source of the missing sequence. While gaps could have been filled *de novo* using spanning ONT UL reads, patching with v0.7 assembly sequence (based on the same ONT UL data) had the benefit of being extensively polished by other sequencing technologies. Patch coordinates were identified by aligning consensus sequences to the v0.7 assembly using minimap2 (*82*) with the command `minimap2 -H -x asm20`. Alignments were subject to strict criteria on size (> 4 kbp), identity (>98%) and number of allowed unaligned flanking bases (200bp). All 25 gaps across the genome were successfully patched, with a maximal patch size of 6.5 kbp.

Assembly of the rDNA arrays

The five rDNA arrays on the acrocentric chromosomes 13, 14, 15, 21 and 22 are highly repetitive regions consisting of long tandem repeats with a repeat motif size of approximately 45 kbp, which further consists of other tandem repeat units. It was observed that most rDNA arrays consisted of a repeating major morph with a few distinct repeat units close to the PJ and DJ regions. Heterozygous structural variants were observed near the distal boundaries of the rDNA arrays on chromosomes 14, 15, and 21, including a megabase-scale heterozygous deletion within the Chromosome 15 array (Figs. S1 and S2). In all cases we chose to include the variant that was most similar to the canonical rDNA unit and took care not to mix reads from different haplotypes. Given their multi-megabase size, additional heterozygous copy number variation within the interior of the arrays could not be ruled out. However, based on the assembly graphs, we can exclude the presence of additional, non-rDNA, sequences within the arrays. Based on this observation, model array sequences were created by filling in the gaps between confidently assembled boundary repeat units of the arrays with copies of the major morph. The number of

copies was estimated using both molecular and bioinformatic techniques, which resulted in consistent estimates (Figs. S1, S9, and S10).

A custom pipeline was used to analyze the rDNA arrays and recover consensus sequences of distinct repeat 'morphs'. HiFi reads were first recruited based on a previous rDNA consensus sequence (*47*). The reads were then separated by chromosome via building a sparse de Bruijn graph constructed from recruited, homopolymer-compressed reads with MBG (*33*) (k=3501), in which five rDNA arrays corresponded to separate connected components (Fig. S12A). Array-specific sparse DBGs with k=201 were then built from the uncompressed HiFi reads. The resulting chromosome-specific graphs revealed a prominent central loop structure representing the 45 kbp rDNA unit surrounded by unique distal and proximal junctions (Fig. S12B). ONT reads were assigned to a particular array and corrected by aligning them to the chromosome-specific graphs with GraphAligner (e.g. each aligned read sequence was replaced by the alignment path in the graph). All fragments representing array-specific rDNA repeats were extracted from the corrected ONT reads, and grouped based on the pairwise edit distances through single-linkage clustering with a distance threshold of 200 (corresponding to a 99.5% identity). A consensus sequence for individual clusters, which was considered to represent distinct repeat morphs, was built with SPOA (*83*). Each morph represents a set of one or more nearly identical rDNA units, and most morphs are distinguishable by variable-length repeats within the intergenic spacer (Fig. S35). Note, this approach was not sensitive enough to discern small SNP-level variations below the clustering threshold. For each array, the major morphs were identified, i.e. the morph corresponding to the highest number of ONT read fragments, and polished with Racon (*83*) using array-specific HiFi reads. Morphs included in the v1.1 assembly were further polished via manual inspection of the alignments of mapped HiFi and ONT reads. Information of ONT reads spanning between different morphs was used to build a graph of morphs to represent the structure of the rDNA arrays (Fig. S12C) showing the simple structure of chromosomes 14 and 22 and the complex mosaic structure of chromosomes 13, 15 and 21.

To accurately reconstruct the boundary repeat units of the rDNA arrays, ONT reads were extracted that anchored to the unique sequence in the PJ/DJ regions surrounding the rDNA arrays and separated by chromosome, PJ/DJ arm, and haplotype (in cases of known heterozygosity). Racon was used to build a consensus for each group of ONT reads. To further correct the ONT-based consensus sequences using HiFi reads they were aligned to an updated HiFi string graph and substituted by the alignment path sequence. This approach allowed us to accurately recover several copies of the rDNA repeat unit at the ends of every array.

The rDNA array on chromosome 15 had a more complex structure (Fig. S12C). First, its length differed between the two haplotypes, with the shorter array comprising only 7 repeat units spanned by a single 450 kbp ONT read (Fig. S2). Second, its longer haplotype appeared to consist of multiple copies of three distinct morphs interleaved with each other. Thus, we included the three major morphs as three distinct blocks each containing the estimated number of copies.

However, this does not accurately reflect the true interleaved ordering of the morphs and should be considered only a model until longer sequencing reads enable a more accurate reconstruction of this array in the future.

Polishing

Given the high quality of this draft, a conservative polishing approach was taken. Vertebrate telomeric sequences (TTAGGG) were identified as in the Vertebrate Genomes Project (*84*), identifying one missing telomere on the p-arm of Chromosome 18. Telomeric reads were manually identified and added, increasing the assembly by 4,862 bp (*34*). Subsequently, short-reads were aligned to the assembly using bwa-mem v0.7.15 (*85*) and long-reads using Winnowmap2 v1.11 (*78*, *79*). Single-nucleotide variants (SNVs) were identified using the "hybrid" model in DeepVariant v1.0 (*86*), combining HiFi and short-read data. PEPPER-DeepVariant v0.3 (*87*) was used to identify SNVs from ONT UL data. Variants were filtered to eliminate poorly supported variants and this list of SNVs was further filtered using Merfin (*88*) which ensured the SNV change does not introduce *k*-mers not present in the Illumina or HiFi data. Larger structural variants (SVs) were identified using Parliament2 (*89*) for short-reads and using Sniffles (*90*) v1.0.12 from long reads (CLR, HiFi, ONT). SVs with low support (<30% of reads supporting ALT allele) were removed (*34*). The three technology-specific call sets were merged using Jasmine v1.0.2 (*91*) and all variants >=30bp supported by at least two technologies were manually inspected using IGV (*92*). In total, 993 SNV and 3 SV corrections were made (*34*). The Merqury-estimated QV increased from Q70.2 to Q72.6. The initial polishing filtered SNPs near chromosome ends due to low coverage and ONT sequencing strand bias, leading to lower quality calls in telomeric sequence. Therefore, we ran a second round of polishing after re-training the existing model of PEPPER to allow variants to be called with support from only a single strand (*34*). A custom script was used to filter the telomeric variants and select those which decreased the Levenshtein distance to the canonical telomere k-mer. Using this method we made 454 telomere edits in the v1.0 assembly (*34*). The Merqury-estimated QV increased further to Q73.9 (67.86 from PCR-Free Illumina-only; 69.80 from HiFi-only), exceeding the original Q40 definition of "finished" sequence (*93*) and meeting the Q60 standard of the VGP (*84*). More details and specific parameter settings are described in (*34*).

All changes made from v0.9 to v1.0 and v1.0 to v1.1 are available as VCF files linked from the CHM13 project GitHub (https://github.com/marbl/CHM13).

HG002 ChrX Assembly and Validation

HG002 Whole genome assembly

The same assembly pipeline was used (see "Whole genome assembly" section) to reconstruct Chromosome X from the GIAB (*94*) HG002 cell line (*15*). The HPRC (https://humanpangenome.org/hg002/) HiFi libraries (Table S1) were used to build a string graph and ONT UL data (*95*) to manually resolve tangles. A GA-rich region in the p-arm pseudoautosomal region (PAR) was patched with a Flye (*96*) trio-binned (*97*) and medaka polished assembly. The complete Chromosome X is available at NCBI (Table S1). The assembly

used for analysis in this and companion papers was missing approximately 2 Mbp short of the p-arm telomeric sequence as the GA gaps were not yet patched. This assembly is available from the CHM13 github site at: https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/HG002/assemblies/HG002.chrX_v0.7.fasta.gz.

HG002 Molecular Methods

HG002 Pulsed field gel electrophoresis

Alpha satellite array sizes were estimated by PFGE and Southern blotting using established methods (*14*, *98*). High molecular weight DNA from 10^7-10^8 was embedded in 1% low melting point agarose plugs and digested with restriction enzymes that cut infrequently within alpha satellite DNA, releasing the DXZ1 array as one or a few large fragments. HMW DNA in one-half of an agarose plug was digested overnight with 20U of enzyme and run on a 1% agarose gel. Saccharomyces cerevisiae and Hansenula wingei chromosomes embedded in agarose were used as size standards (Bio-Rad CHEF DNA Size Markers). Gels were run at 3V/cm for 48-51 hours at 14°C in 1X TAE buffer, using switch times of 250 seconds (initial) – 900 seconds (final). CHM13, containing a previously sized DXZ1 array, was used as a control (*14*) (Fig. S6).

HG002 Southern blotting

After electrophoresis, gels were stained with ethidium bromide and imaged using a UV light source. Gels were depurinated with 0.25 M HCl for 12 minutes at room temperature, then washed twice for 15 minutes each in denaturing buffer (1.5 M NaCl, 0.5 M NaOH). DNA was transferred to HyBond-N+ membrane (GE Healthcare/Amersham) for 48 hours in fresh denaturing buffer. DNA was UV-crosslinked to membranes using the auto-crosslink setting on a Stratagene Stratalinker prior to hybridization.

A 500bp fragment spanning monomers 9-12 of DXZ1 was generated by PCR (*99*) and labeled overnight at 37°C with digoxygenin-11-dUTP using DIG High Prime (Sigma-Aldrich). Labeling reactions were purified using the High Pure PCR purification kit (Roche).

Membranes were pre-hybridized for 45-60 minutes in glass hybridization bottles containing 20mL ExpressHyb buffer (Clontech) at 63°C. Pre-hybridization buffer was replaced with 20mL of fresh ExpressHyb containing 300-350ng of labeled probe that has been denatured at 95°C for 10 minutes. The probe was allowed to hybridize to the membrane at 63°C overnight in a hybridization oven. Membranes were washed at 68°C twice for 20 minutes in 2X SSC/0.1% sodium dodecyl sulfate (SDS), followed by a single high stringency wash in 0.2X SSC/0.1% SDS for 15 minutes at 68°C. Membranes were blocked in 1X Western blocking reagent (Roche) in maleic acid buffer (0.1 M maleic acid, 0.15 M NaCl, pH 7.5) for 60-90 minutes at room temperature, then incubated for 60 minutes in blocking buffer with anti-digoxygenin-alkaline phosphatase (Roche, 1:2000). Chemiluminescent detection was performed using 5mL of CDP-Star ready-to-use reagent (Tropix). Membranes were imaged on a G:Box using GeneSys

software (Syngene) for direct image analysis. Images were adjusted (leveled to curves) and labeled in Adobe Photoshop.

HG002 ddPCR

Genomic DNA was isolated using the DNeasy Blood & Tissue Kit (Qiagen). DNA was quantified using Qubit Fluorometer with Qubit dsDNA HS Assay (Invitrogen). Primers, gDNA concentrations, and restriction enzymes are in Tables S2 and S3. EvaGreen ddPCR reactions were performed using the manufacturer's protocol (Bio-Rad). Mastermixes were simultaneously prepared for HPRT1 and the gene of interest which were then incubated for 15 minutes to allow for enzymatic restriction digestion. Statistics were performed using the confidence interval calculated by the Quantasoft software and applying it to Taylor's expansion (Fig. S7).

HG002 Alignment-based validation

TandemTools (*37*) was run to validate the HG002 centromeric satellite assembly using ONT data. The coverage plot (Fig. S8A) showed no suspicious coverage dips or spikes, except for two peaks at 2.8 and 3.1 Mbp corresponding to LINE elements. The plot (Fig. S8B) showed high base-calling quality across the whole centromere (usually represented by k-mers forming no clumps) and suggested an absence of collapsed repeats (usually represented by k-mers forming multiple clumps). Neither this analysis nor comparison to CHM13 revealed any errors or high frequency heterozygous sites in the centromeric satellite assembly.

Assembly validation

Our assembly strategy was iteratively refined by validation of intermediate results, most notably by TandemTools (*37*). We identified deficiencies in the pipeline (e.g. missing high-coverage overlaps and low consensus quality, 'Read Processing' and 'Consensus' sections) which were corrected before building the final chromosome reconstructions. We validated the final assembly using a mixture of techniques and sequencing data, as detailed below. Remaining known issues have been catalogued and are available at https://github.com/marbl/CHM13-issues.

ddPCR copy number validation

ddPCR reactions were performed using unique primers designed for amplicons in targeted regions (Tables S2 and S3). Each reaction consists of 10 uL 2x ddPCR QX200 Evagreen Supermix, 0.2 uL of restriction enzyme for fragmentation, 1 uL 10 uM primer mix, 1 uL of 0.1-1 ng CHM13 DNA template and 7.8 uL with nuclease free water. Mastermixes were then emulsified with Evagreen droplet generator oil (Bio-Rad) using a QX200 droplet generator according to the manufacturer's instructions. After droplet generation, thermocycling was performed with the following parameters: 10 min at 95°C, 40 cycles consisting of a 30-s denaturation at 94°C and a 60-s extension at 59°C, followed by 10 min at 98°C and a hold at 4 °C. Control reactions without the DNA were performed to rule out non-specific amplification.

Following PCR amplification, the 96-well plate was transferred to a QX200 droplet reader (Bio-Rad). Positive droplets were automatically determined by the QuantaSoft software. Concentrations reported where copies/µL of the final ddPCR reaction were adjusted according to

the TBP1 single copy gene present at Chromosome 6. The normalized copy number values for 28S repeated array were calculated as follows: [(28S target copies/µL)/(TBP1 copies/µL)] x 10 (Fig. S10A). The normalized copies for other arrays were calculated as for 28S, using the single-copy gene listed in Table S3 (Fig. S10B,C).

Alignment-based validation

Agreement of the assembly with the data used to generate it is a commonly used validation method (*100*). We mapped all available data (HiFi, ONT) to the assembly using Winnowmap v2.01 (*78*, *79*) and PCR-free Illumina data using bwa v0.7.15 (*85*). PCR duplicate-like redundancies were removed using biobambam2 `bamsormadup` (v2.0.87) (*101*) with default parameters. For long-read analysis, non-primary alignments were filtered using `samtools view -F 256` (*102*). The resulting read coverage was evenly distributed across all technologies, with the notable exception of some human satellite classes (see 'Technology-specific sequencing biases' for details). We used these alignments to generate summary statistics (min, max, mean, median) for 10 kbp non-overlapping windows of the assembly for each of: alignment length, read length, identity, and MAPQ.

Strand-seq validation

We validated the assembly using Strand-seq (*103*, *104*) data from CHM13 (*20*, *21*) and compared it to GRCh38. We first aligned paired-end Strand-seq data to both CHM13 and GRCh38 assemblies using BWA-MEM (version 0.7.15). Next we used breakpointR (*105*) to detect regions with recurrent changes in strand directionality across multiple Strand-seq libraries. Regions with the majority of reads mapped in minus orientation are suggestive of unresolved inversion or contig misorientation, while regions where we see a mixture of plus and minus reads are suggestive of possibly collapsed or low mappability (*106*). Overall we observe no large inversion or misorientation (Fig. S13A) as well as no chromosome mis-assignment (Fig. S13B) events in CHM13 assembly.

The Strand-seq alignments allowed us to verify chromosome assignments even in repetitive regions, such as the acrocentric chromosomes. Fig. S14 shows the mapping of Strand-seq reads across Chromosome 15. In contrast to GRCh38, which has no reads mapped to the short arm, CHM13 shows a consistent read orientation across the entire chromosome, confirming the coloring in Fig. S13B and the correctness of the acrocentric reconstructions.

Hi-C validation

We used Hi-C data to validate the long-range structure of the assembly. Hi-C reads were aligned to the CHM13v1.0 assembly using the VGP Hi-C mapping and visualization pipeline (*84*). In brief, both ends of a read pair were mapped independently using BWA-MEM (*85*) with the parameter `-B8`, and filtered when mapping quality was <10. Chimeric reads were trimmed from the restriction site onward, leaving only the 5′ end. The filtered single-read alignments were then rejoined as paired read alignments. Alignments were converted with PretextMap (https://github.com/wtsi-hpag/PretextMap) and visualized with PretextView

(https://github.com/wtsi-hpag/PretextView) and HiGlass (*107*). The resulting Hi-C interaction matrix showed a clean assembly, with strong interactions within chromosomes and few strong interactions between chromosomes (Fig. S15), supporting the overall structure of the assembly.

BAC-based validation

We validated the assembly using previously sequenced and newly released BACs for CHM13 (*20*, *21*), including corrections made by HiCanu (*16*, *95*). 644 out of 647 BAC sequences had a continuous mapping to the assembly at Q41.94 (averaged across all aligned bases). To confirm the correctness of the assembly in the regions corresponding to the three remaining BACs, we selected all ONT UL reads with primary alignments to Chromosome 17 and Chromosome X of the CHM13v1.0 assembly at the locations of the partial BAC alignments. We also included a "resolved" BAC (AC279712.1) from Chromosome X as a control. Selected ONT reads were then aligned to the BACs in question with Winnowmap v2.0.1 (the command `winnowmap -z150 -t 16 -ax map-ont -H -k 19`, retaining only primary alignments (`samtools view -F 256`). Alignments were inspected in IGV. The three "unresolved" BACs (AC279506.1, AC279581.1, AC279712.1) show uneven coverage and an increase in variant positions, indicating mismatches versus the ONT data (Fig. S16-18). In contrast, the control BAC exhibits even coverage and no variant sites (Fig. S19). As the assembly did not use ONT data for consensus, this provides independent confirmation of its structure. It is likely that the relatively low QV (compared to the Merqury estimated values) is due to limitations of BAC quality as opposed to the assembly.

K-mer based validation

Merfin (*88*) analysis confirmed that assembly k-mer multiplicity is consistent with both the HiFi and PCR-free sequencing data (Fig. S21), with a notable improvement from the v1.0 to the v1.1 assembly due to the filled in rDNA arrays.

Assembly completeness

Assembly completeness was evaluated using both *k*-mer statistics and mapping-based statistics. Using PCR-free Illumina and HiFi data, Merqury (*81*) completeness for the assembly was 99.1% and 99.8%, respectively. Based on primary HiFi alignments, 97.5% of the reads (97.5% of bases) have an alignment with a MAPQ=60, >= 99% identity, and covering >= 90% of the read length. In contrast, for CHM13 v0.7 (*14*), only 92.7% of reads (92.6% of bases) meet this criteria. Some reads were expected not to map full length or at high identity due to heterozygous variants present in the CHM13 cell line but not in the assembly. Specific regions such as simple sequence repeats in HiFi data also show an elevated error rate (*16*). Thus, the criteria was relaxed to include reads with an alignment with MAPQ >= 20, > 90% identity, and covering > 50% of the read length. Additionally, reads assigned to the rDNA arrays, corresponding to the 5 known gaps in CHM13v1.0, were ignored. Only 29,320 reads (524 Mbp) remained unaligned to the assembly. Assuming 32.4X coverage, this is only 15 Mbp of potentially missing bases (0.5% of the genome) and is likely an overestimate due to chimeric, erroneous, or unassigned (to the rDNA) reads. Comparatively, the v0.7 assembly has 154,951

reads (2,823 Mbp), or 87 Mbp of missing bases (2.85% of the genome). Both alignment-based and k-mer based analyses support that CHMv1.0 is highly complete.

Technology-specific sequencing biases

Despite overall uniform coverage by HiFi read mappings (with the known exception of GA-rich sequence), we observed that some multi-megabase long regions within the CHM13v1.0 assembly showed consistent excess or depletion of coverage. These corresponded to human satellite (HSat) regions (Fig. S29). While some small errors likely remain in the complex HSat arrays, the observed coverage discrepancies are very unlikely to originate from potential assembly errors, and instead point to sequencing bias. First, in all cases, coverage excess and depletion events were consistent with the satellite's class (increased coverage corresponded to HSat2/3, and depleted coverage to HSat1A). Second, analysis of HiFi read alignments in the regions of increased coverage revealed relatively few sites with a high-frequency of second-most common bases (Fig. S22), and also did not reveal extensive coverage spikes (beyond the uniform increase in the entire region). These observations indicated the absence of large-scale repeat collapses (*14*, *38*). Third, coverage anomalies were inconsistent between HiFi and ONT sequencing libraries with ONT read alignments showing no increase of coverage for HSat2/3 arrays, and an even larger depletion (approximately 50%) of HSat1A arrays coverage (Fig. S22). To further investigate this phenomenon we evaluated various statistics (identity, length by strand, etc) for ONT reads mapped across instances of HSat1A, HSat2 and HSat3 satellite arrays and visualized them with IGV (*92*) (Figs. S23 to 25).

To summarize the properties of PacBio and ONT reads from different types of satellite regions, a mapping-independent strategy was implemented. In addition to the three HSat classes, alpha satellite (AlphaSat) was added as an example satellite with no HiFi coverage abnormalities. For each satellite class of interest, namely HSat1A, HSat2, HSat3, and AlphaSat relevant ONT reads (both Bonito and Guppy base-called), HiFi reads, and raw Pacbio subreads from which HiFi sequences were generated were recruited. Raw subreads were considered to assess the performance of the SMRT sequencing step separately from the subsequent subread consensus step (generating HiFi reads). Sequence assignment was performed using pre-identified sets of class and strand-specific marker *k*-mers (*k*=21). Namely, a 1 kbp long sequence was assigned to a particular satellite class if it contained 5 or more class-specific *k*-mers. A read/subread was assigned according to the majority rule across the assignments of its non-overlapping 1 kbp windows, with ties broken randomly. Relevant summary statistics, involving assigned reads and their class-assigned 1 kbp windows are presented in Table S4.

The enrichment of HSat2 in HiFi reads (>53%) is more pronounced than in the raw PacBio subreads (<45%, Table S4), a finding that is consistent with HSat2 assigned polymerase reads having the highest HiFi conversion rates across considered satellite classes (57.37% vs 44.08%, 50.52% and 52.59% for HSat1A, HSat3, and AlphaSat, respectively). Polymerase reads containing a subread assigned to a particular class counted as successfully converted if it resulted in a HiFi read assigned the same class. However, consistent values across HiFi and raw PacBio

subreads in Table S4 show that the computational consensus step is unlikely to be the major factor of observed HSat enrichment/depletion in the PacBio HiFi dataset.

In ONT data, HSat1A and HSat2 were associated with shorter read lengths than AlphaSat and surrounding regions (Table S4, Figs. S23 and S24). We hypothesize that the AT-rich HSat1A may be unstable during library preparation and shears preferentially, with shorter molecules leading to lower coverage by both ONT data (due to shorter read lengths) and HiFi data (due to fewer molecules passing strict size selection). ONT signal processing methods have known biases in AT-rich repeats (*108*), which could lead to fewer reads successfully passing quality filters and even incorrect detection of molecule boundaries. ONT data associated with HSat2 exhibits a significant strand bias (Table S4 and Fig. S24), with a high depletion only in sequencing reads coming from the 'reverse' strand ("GAATG"-unit) vs the 'forward' strand of the repeat array ("CATTC"-unit). Note that the entire HSat2 array on Chr1 was inverted with respect to the canonical unit orientation. However, the higher accuracy of ONT Guppy reads and regular coverage of the forward strand masks these effects in 'unstranded' mapping-based coverage plots (Fig. S22). ONT reads were also depleted in AlphaSat while HiFi is consistent with expectation.

The biases exhibited by ONT read sets differed across Guppy and Bonito base-callers. While Bonito base-calling positively affected (reduced) AlphaSat and HSat2 depletion and reduced the strand bias, there was a negative impact on the identity of the reads from HSat1A.

Finally, CLR kinetic information was inspected for PacBio polymerase reads originating from different types of satellite regions (Fig. S26). Kinetic information was collected for all ZMWs with at least 3 passes (but may not have produced a HiFi read) and had at least one subread assigned to the particular satellite class based on *k*-mer analysis. Unfortunately, this analysis did not lead to conclusive insights into the origin of HSat enrichment phenomena.

HSat2,3 arrays were associated with slightly longer sequencing times and polymerase lengths in PacBio data, compared to AlphaSat and whole-genome stats (Fig. S26). We hypothesize that these regions might be 'easier' for the PacBio polymerase to sequence, which would lead to more reads surviving the initial pre-extension step, and in turn contribute to the observed enrichment. HSat1A arrays are depleted in the longest (>250 kbp) polymerase reads, and are associated with slightly shorter sequencing time (vs AlphaSat and whole-genome).

Here, many complex human satellites have been assembled for the first time (*30*). The potential causes leading to the observed coverage biases were briefly explored, but more analysis is needed in future. It is unsurprising that machine-learning methods relying on learning from known sequence contexts, e.g. ONT base-calling and HiFi subread consensus, performed poorly in these regions. We hope that the availability of a finished human reference along with our observations will lead to improvements in the quality of both ONT and PacBio reads originating from challenging satellite genomic regions in the near future.

Alpha-satellite and human satellite 1,2,3 validation

TandemTools (*37*) provided orthogonal validation for the assembly of alpha and human satellite arrays, excluding rDNA (Table S5 and Fig. S27). Quality assessment with both accurate PacBio HiFi and error-prone ONT reads revealed positions in the assembly with consistent discrepancies between mapped reads and the assembly in distances between consecutive rare *k*-mers (default $k = 19$), a *k*-mer was considered rare if its frequency in the assembly did not exceed 10 (*37*). A list of regions where the discrepancy was supported by at least 30% of covering reads and may indicate heterozygous sites (*34*) was compiled. All putative events within the centromeric satellite DNA were further manually inspected in IGV for validity and catalogued at https://github.com/marbl/CHM13-issues. Two sites with 100% deviated reads (on Chromosomes 18 and 19) corresponded to a low coverage region (Chr18) and a potential 2.4 kbp insertion in the assembly (Chr19). The insertion in Chromosome 19 was not corrected as the alternate sequence had ambiguous mapping support. Besides that possible error, arrays of centromeric satellite DNA do not contain large structural mis-assemblies (longer than 100 bp).

The proportion of monomers of each live higher-order repeat (HOR) in sequencing reads versus the same proportion calculated from the HOR annotation (*30*) in our v1.0 assembly were also compared. The expected length of AS monomers was ≈ 170 bp. The number of the live HORs was 19, including S2C13/21 and S2C14/22 each located on 2 chromosomes and S1C1/5/19 located on 3 chromosomes. In these cases, the HOR counts were divided by 2 for doubles and by 3 for the triple. The sister HORs of the live HORs were not included (e.g. S3C17H1-B and C). Only the near full-length monomers (length >= 160 bp) were counted to reduce the risk of erroneous monomer classification. Monomer classification/identification was performed by HumAS-HMMER-HOR (https://github.com/enigene/HumAS-HMMER) described in (*109*) (for SF1 HORs only) and in (*30*) (all known HORs covered). Here, a HMMER platform (*110*) was used to identify (as the best match in a set of standards) each monomer of about 80 known HORs and about 30 classes of monomers in AS monomeric layers. Note that for monomers of homogeneous HORs only one representative was selected so the best match was equivalent to the best alignment. For divergent HORs, the standard was formed using a multi-alignment of representative monomers and therefore constitutes a true HMM. Conversion of the HMMER output file into a bed file was performed using the `hmmertblout2bed` script (https://github.com/enigene/hmmertblout2bed) (*109*).

Monomers were counted in both the PCR-free Illumina 250 bp data and the HiFi 20 kbp libraries. In the short read sample, a total of 8,314,086 HOR monomers were identified and 2,045,562 HOR monomers remained post length filtering. In the long read sample a total of 3,792,752 HOR monomers were identified and 3,720,342 HOR monomers remained post length filtering. In the assembly, a total of 426,773 HOR monomers were identified and 425,732 HOR monomers remained post length filtering (Fig. S28). There was a good agreement between the proportion of monomers in the sequencing data and the assembly for both data types.

HiFi data was aligned and the frequency of secondary alleles plotted as in (*14*). Reads were aligned with the following command: `pbmm2 align --log-level DEBUG --preset SUBREAD --min-length 5000` and filtered with `samtools -F 2308`. Plots were generated using NucFreq's `NucPlot.py` script (https://github.com/mrvollger/NucFreq)(*38*) for selected regions (Fig. S29). Since CHM13 is homozygous (with few exceptions), assembly loci with high second most frequent allele counts are a signal of possible misassembly (*38*). As expected, the assembly had a low level of secondary alleles across all satellite arrays, with only a handful of potential variants identified, and had even coverage, with the exception of sequencing artifacts in HSat regions (see 'Technology-specific biases' section). The rDNA arrays had depleted coverage and spikes at the borders due to being model sequences. Note that simpler rDNA arrays (see 'Assembly of the rDNA arrays' section) on chromosomes 14, 22 showed even coverage through the array while model-based resolutions (chromosomes 13, 15, 21) showed a decrease in coverage and an increase in secondary allele frequency.

Genome analysis

Assembly tracks

UCSC Genome Browser (*54*) assembly hubs have been released for both the v1.0 (http://genome.ucsc.edu/cgi-bin/hgTracks?genome=t2t-chm13-v1.0&hubUrl=http://t2t.gi.ucsc.edu/chm13/hub/hub.txt) and v1.1 (http://genome.ucsc.edu/cgi-bin/hgTracks?genome=t2t-chm13-v1.1&hubUrl=http://t2t.gi.ucsc.edu/chm13/hub/hub.txt) assemblies.

These tracks were used to generate Figures 1, 3, 4, and 5 of the main text. Specifically, segmental duplication and censat annotations were computed as in (*30*, *42*) and plotted as non-overlapping regions. SD density in 10 kbp windows was computed using the kpPlotDensity (*111*) function and colored by log of the window density. Sequence in CHM13 but not in GRCh38 was identified based on whole-genome LastZ (*112*) alignment of GRCh38 and CHM13, filtering for any regions without 1 Mbp of synteny (*42*). A more conservative set of regions not covered by GRCh38 was computed using Winnowmap v2.0.1 to align GRCh38 (excluding ChrY, patches, and alts) to CHM13v1.1 with the command `winnowmap -ax asm20 -H --MD chm13.v1.1.fasta GCA_000001405.28_GRCh38.p13_genomic_noalt_noY.fasta > out.sam`. Primary alignments were converted to paf and filtered for primary only via `k8 paftools.js sam2paf -p out.sam > out.paf` and converted to unmapped regions by inverting the list of mappings w/MAPQ>0: `cat out.paf |awk '{if ($12 > 0) print $6"\t"$8"\t"$9}'|bedtools sort -i - |bedtools merge -i - |bedtools complement -i - -g chm13.sizes`. GRCh38 (*1*) issue XML files were downloaded from the GRC (ftp://ftp.ncbi.nlm.nih.gov/pub/grc/human/GRC/Issue_Mapping/) and filtered to exclude issues resolved in GRCh38 or a patch or marked as 'Variant' or 'GRCHousekeeping', resulting in 191 issues. Regions were padded by 1 Mbp, lifted over with

the command `liftOver -bedPlus=3 -minMatch=0.5 hg38_issues.bed hg38.t2t-chm13-v1.0.all.chain.gz chm13_lifted_issues.bed unmapped.bed`, and trimmed back by 1 Mbp. A total of 43 issues could not be lifted over. Genes were identified from the combined CAT and LiftOff annotations (see 'Annotation' below).

The computed ancestry regions (see "Ancestry Analysis"), in GRCh38 coordinates, were lifted over from GRCh38 to CHM13 with the command liftOver -bedPlus=3 -minMatch=0.75 ancestry.bed hg38.t2t-chm13-v1.0.all.chain.gz chm13_v1.0_lifted_ancestry.bed unmapped.bed.CHM13. CHM13 v1.0 bed files were lifted over to v1.1 using the command `liftOver <input v1.0 bed> v1_to_v1.1.chain <output v1.1.bed> unmapped.bed`

Annotation

The Cactus (*113*) alignment between the v1.0 assembly and the primary contigs of GRCh38 (*1*), with chimp as an outgroup, was created with the following command:

```
cactus aws:us-west-2:t2t-jobstore-chm13 cactus-config-
chm13-t2t.draft_v1.v2.txt t2tChm13.draft_v1.v2.hal --
maxCores 80 --binariesMode local
```

Using the following config file for cactus:

```
(Chimp:0.00655,(GRCh38:0.0005,CHM13:0.0005));
Chimp   GCF_002880755.1_Clint_PTRv2_fixed.fa
GRCh38  GRCh38.primary.fa
CHM13   t2t-chm13-v1.0.fa
```

Iso-Seq reads were aligned using minimap2 (*82*) using the following command:

```
minimap2 -ax splice -f 1000 --sam-hit-only --secondary=no -
-eqx -K 100M -t 4 --cap-sw-mem=3g mmdb/0.mmi
iso_fastas/0_0.fasta
```

Stringtie2 (*114*) assembled the transcriptome using available Iso-Seq reads:

```
stringtie -p 8 chm13_1.t2t.sorted.bam.filtered.bam -L >
chm13_1.t2t.TM.stringtie.gtf
```

CAT (*39*) v2.2.1 (commit `96e7550f22387a669f0b98dfc0c94be825192e24`) was run in TransMap (*115*) mode using the following command:

```
luigi --module cat RunCat --hal=t2tChm13.draft_v1.v2.hal --
target-genomes='("CHM13",)' --ref-genome=GRCh38 --
workers=10 --config=cat.t2t.draft_v1.full.isoseq.config --
```

```
work-dir work-chm13-t2t --out-dir out-chm13-t2t --local-
scheduler --assembly-hub --maxCores 5 --binary-mode local
```

Using GENCODEv35 (*41*) and following config file for CAT:

```
[ANNOTATION]
GRCh38 = gencode.v35.annotation.gff3.noPAR
CHM13 = CHM13.TM.stringtie.merged.gff3
[ISO_SEQ_BAM]
CHM13 =
data/chm13_1.t2t.sorted.bam,data/chm13_2.t2t.sorted.bam
```

We removed genes from the CAT annotation that had overlapping annotations from multiple genes in the same family, leaving the gene that was correct based on synteny.

The Liftoff (*40*) annotation was created with the following command using version 1.6.0:

```
liftoff chm13.draft_v1.0.fasta GRCh38.fa -sc 0.95 -copies -
g gencode.v35.annotation.gff3 -polish -chroms chroms.txt
```

To create the final annotation, we complemented the CAT result with missed GENCODE genes and putative additional paralogs (with minimum sequence identity of 95%) from the Liftoff annotation. Only predictions that did not overlap any CAT annotations were added.

The annotation set on the v1.0 assembly was lifted over to the v1.1 assembly using liftover with the command

```
liftOver -gff CHM13.combined.v4.gff3 v1_to_v1.0423.chain
CHM13.combined.v4.liftover.v1.1.gff3 unmapped.txt
```

The rDNAs were annotated by mapping an assembly of an rDNA unit isolated from chromosome 21 (*47*) onto v1.1 with Liftoff. Using GenBank entry KY962518.1, the rDNA sequence was obtained and a gff3 file created with the coordinates of the 45S, 18S, 5.8S, and 28S subunits. Liftoff was then run with the following command to annotate all rDNAs within the assembly

```
liftoff chm13.draft_v1.1.fasta KY962518.1.fasta -g
KY962518.1.gff3 -copies -sc 0.95 -mm2_options="-N 300"
```

All annotations that have been lifted over and that overlapped the added rDNA regions were removed. The rDNA annotations (876 genes) were added to create a final annotation set.

In the final annotation set, 63,494 genes were annotated (19,969 protein-coding, 27,026 non-coding RNA, 15,799 pseudogenes, 652 Immunoglobulin/T-cell receptor segments, and 48 genes without a Biotype from StringTie2). In terms of transcripts, 233,615 transcripts were

annotated (86,245 protein-coding, 57,196 non-coding RNA, 15,997 pseudogenes, 674 Immunoglobulin/ T-cell receptor segments, 56 from StringTie2, and 73,447 other types). Further breakdown by biotype is provided by Tables S6 and S7.

Through this process 469 frameshifts in 387 genes were identified in CHM13 versus GENCODE v35. IDs of the frameshifted genes and transcripts can be found in Table S8. An example is shown in Fig. S30.

From the GENCODE v35 annotation, 263 genes were missing from the v1.1 assembly gene set (63 protein-coding, 67 non-coding RNA, 94 pseudogenes, and 39 Immunoglobulin/ T-cell receptor gene segments). In terms of transcripts, 1,708 transcripts from GENCODE v35 annotation were missing from the v1.1 assembly annotation (829 protein-coding, 256 non-coding RNA, 109 pseudogenes, 39 Immunoglobulin/ T-cell receptor gene segments, and 475 other). IDs of the missing genes and transcripts can be found in Tables S9 and S10, respectively. There are a number of reasons why genes may not have been annotated by either the CAT or Liftoff pipelines. First, the gene may truly be absent in the v1.1 assembly, and, most likely, the CHM13 genome. Fig. S31 gives an example of *ORM1* gene, falling within a 6,762 bp deletion on Chr9 in the v1.1 assembly relative to GRCh38

This 'true' gene loss is especially likely in case of genes associated with segmental duplications within copy number variable regions, such as paralogs from CT45/47 gene clusters on Chromosome X which have previously been validated as correct in the CHM13 assembly (*14*). Second, GENCODE genes within the repetitive genomic regions could be missed due to alignment failure. With CAT, this can happen when the genomic sequence of CHM13 did not align (or aligned poorly) to GRCh38. Alignment of the GENCODE transcripts to the assembly (done by TransMap in CAT, and Minimap2 in Liftoff) could also fail. Many such genes and transcripts might be successfully annotated in future. Lastly, GRCh38 may have assembly errors that create false duplications and 'missing' copies. For example, the region on Chromosome 21 shown in Fig. S32, has been flagged by prior studies (*7*, *116*) and is not consistent with copy-number variation seen in the human population (*25*). Overall, at least 23 of those genes correspond to presumed errors in the GRCh38.

Genes and transcripts exclusive to CHM13 are in Tables S11 and S12. The identities of the exclusive protein-coding genes against their closest match in GENCODE are in Table S13. Out of 3,604 genes 'exclusive' to CHM13 annotation, 2,680 genes (140 protein-coding), accounting for 3,258 transcripts, were 'putative paralogs' reported by Liftoff (*40*) and most of the remaining ones (876 out of 924, accounting for 876 'exclusive' transcripts) encode rRNA subunits. While some of the putative new genes reported here may also be present in GRCh38, only 1,251 extra putative paralog genes (16 protein-coding) with 1,350 transcripts were identified in GRCh38 by Liftoff using the same parameters as for CHM13v1.1 analysis. Moreover, 1,956 and (99 protein coding) of the genes exclusive to our CHM13 annotation fall within the regions with no GRCh38 primary alignments.

Mappability

In an effort to compare the mappability of this assembly to the GRCh38 reference both a variety of sequencing technologies were compared (differing by read lengths and accuracy) and an aligner-independent strategy was implemented. First, alignments of different CHM13 sequencing read sets to v1.0 assembly (filtered by MAPQ >=60) were used to estimate the average length of perfectly matching segments for each sequencing technology. These characteristic perfect runs were 129 bp, 500 bp, and 33 bp for Illumina, HiFi, and ONT (Guppy v3.6.0 base-called), respectively. To assess the mappability of a genome by a particular read type we used GenMap (*117*) to find positions of unique k-mers in the genome, with k set to the technology's characteristic perfect run size. A sliding window equal to the read size was used to determine if a genomic region was "mappable". Any window with at least 1 kbp, or 33% of the read length for reads < 1 kbp, was considered "mappable". The resulting mappability statistics across different technologies and read lengths is summarized in Table S14.

To find positions of multi-copy *k*-mers in the CHM13v1.0 and GRCh38 for a wide range of *k* (powers of 2+1 from 33 to 131,073 bp) GenMap was used. At *k*=131,073 bp, all CHM13v1.0 *k*-mers were unique and at *k*=65,537 the only repetitive *k*-mers were located within the Chromosome 9 HSat3 array draft reconstruction (Fig. S29). Contrastingly, GRCh38 had megabases of non-unique regions spread across all chromosomes at both *k*=65,537 bp and *k*=131,037 bp, even when gap bases were not considered. Fig. S38 presents several *k*-mer based repeat statistics across different genomes and values of *k*. Note that rDNA repeat units were excluded from *k*-mer analysis since these were absent in CHM13v1.0 and are model-based in the current assembly (see "Assembly of the rDNA arrays").

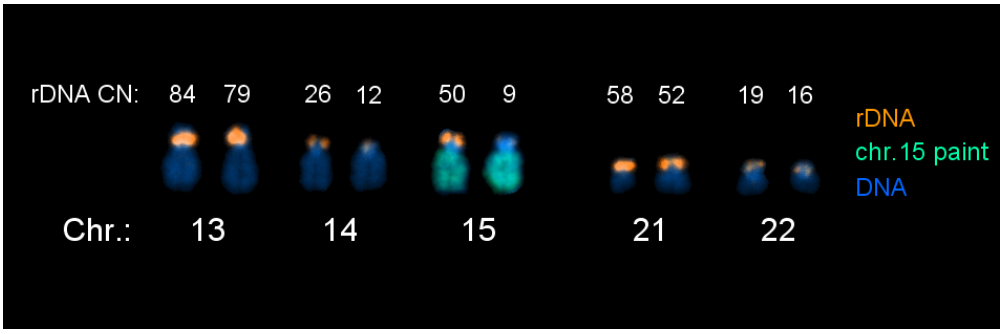
FRG1 Analysis

All *FRG1* paralogs identified in T2T-CHM13 and GRCh38 are listed in Table S15.

For CHM13 paralogs, sequences were aligned using MAFFT v7 (*118*) and analyzed using MEGA X (*119*). A phylogenetic tree was inferred using the Neighbor-Joining method (*120*) with 500 bootstrap replicates. The evolutionary distances were computed using the Kimura 2-parameter method (*121*). This analysis involved 27 nucleotide sequences. All ambiguous positions were removed for each sequence pair (pairwise deletion option). There were a total of 128,204 positions in the final dataset. Additionally, we applied a maximum likelihood method using RAxML (*122*) and the general time reversible+G model (*123*) as matched by jModelTest2. RAxML was run on the PTHREADS version and rapid bootstrap approximation to generate 100 bootstrap replicates. Consensus trees were then generated in Geneious3 and visualized in FigTree4 (Fig. S39). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Despite some of the paralog branches differing between approaches, both trees strongly supported that the *FRG1BP4~10* paralogs shared the same ancestral origin with *FRG1DP* and share identity >96.5% (Table S16).

Transcript abundance was estimated in transcripts per million (TPM) with Salmon v1.3.0 (*124*), using the CHM13 transcriptome and the v1.0 assembly as a "decoy" sequence to account for reads mapping to unannotated sequences. TPM values were aggregated to the gene level using the R package tximport (*125*). In addition, TPM values per gene were calculated from alignments of RNA-seq, PRO-seq, and Iso-seq data and filtered on singly unique markers in the CHM13 assembly. Coordinates were lifted over to the v1.1 assembly. Full details are in (*49*). The *FRG1* paralogs show varying levels of expression (Table S17).

A



B

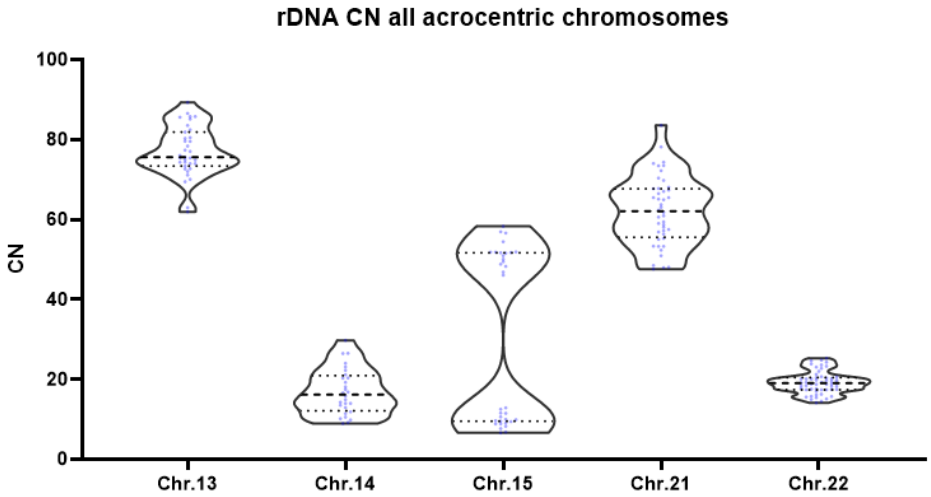


Fig. S1: Estimation of CHM13 rDNA copy numbers. (A) Karyogram of acrocentric chromosomes from a CHM13 chromosome spread labeled by Fluorescent In-Situ Hybridization (FISH) with rDNA probe (orange) and Chromosome 15 paint (green). Chromosomes were counter-stained by DAPI. Chromosome 15 was identified by paint, other chromosomes were identified based on morphology. Estimated rDNA copy numbers per chromatids (rDNA CN) are shown on top of the acrocentric chromosomes with rDNA labeled in orange. Chromosomes 14 and 15 have larger differences in copy numbers, with the rest being within 5 copies. (B) Quantification of the rDNA copy number based on FISH. Chromosomal spreads from CHM13 cells were labeled with rDNA probe and acrocentric chromosome-specific paints. The rDNA copy numbers per specific chromatids were calculated from the fraction of the total fluorescent intensity of the rDNA signal on all chromosomes in a given spread and the Illumina sequencing estimate of the total copy number of rDNA repeats in the CHM13 genome. In addition to rDNA loci on painted chromosomes, all rDNA loci on other identifiable chromosomes were included.

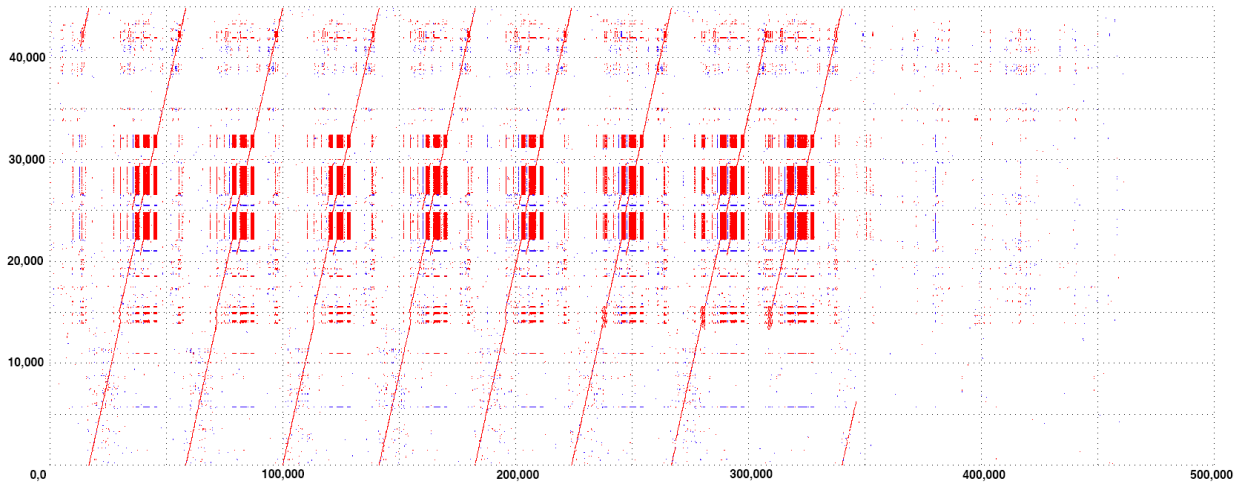


Fig. S2: MUMmer (*126*) alignment of a canonical rDNA copy to a single ONT read. The y-axis corresponds to the canonical rDNA unit (*47*) and the x-axis to a single CHM13 ONT read. Forward matches are in red and reverse complement matches in blue. The read has seven copies of the 45S gene cluster and shows a consistent structure within its repeat units. It is anchored in a non-rDNA sequence at the start (x-axis 0-10kb) and the end (x-axis 350 kbp-450 kbp). The contained rDNA units match assembled morphs from chromosome 15.

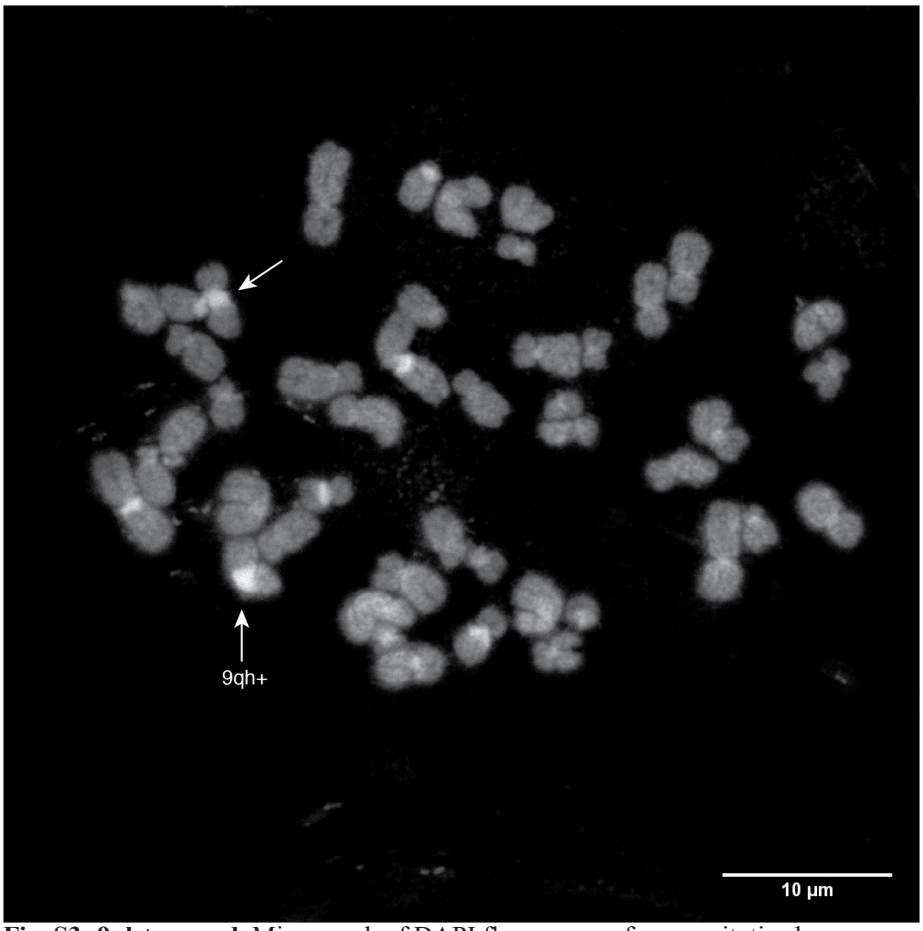


Fig. S3: 9qh+ spread. Micrograph of DAPI fluorescence from a mitotic chromosome spread from the CHM13 cell line at passage 10 (*14*). Arrows denote the large pericentric Human Satellite 3 array on Chromosome 9. Bar, 10 µm.

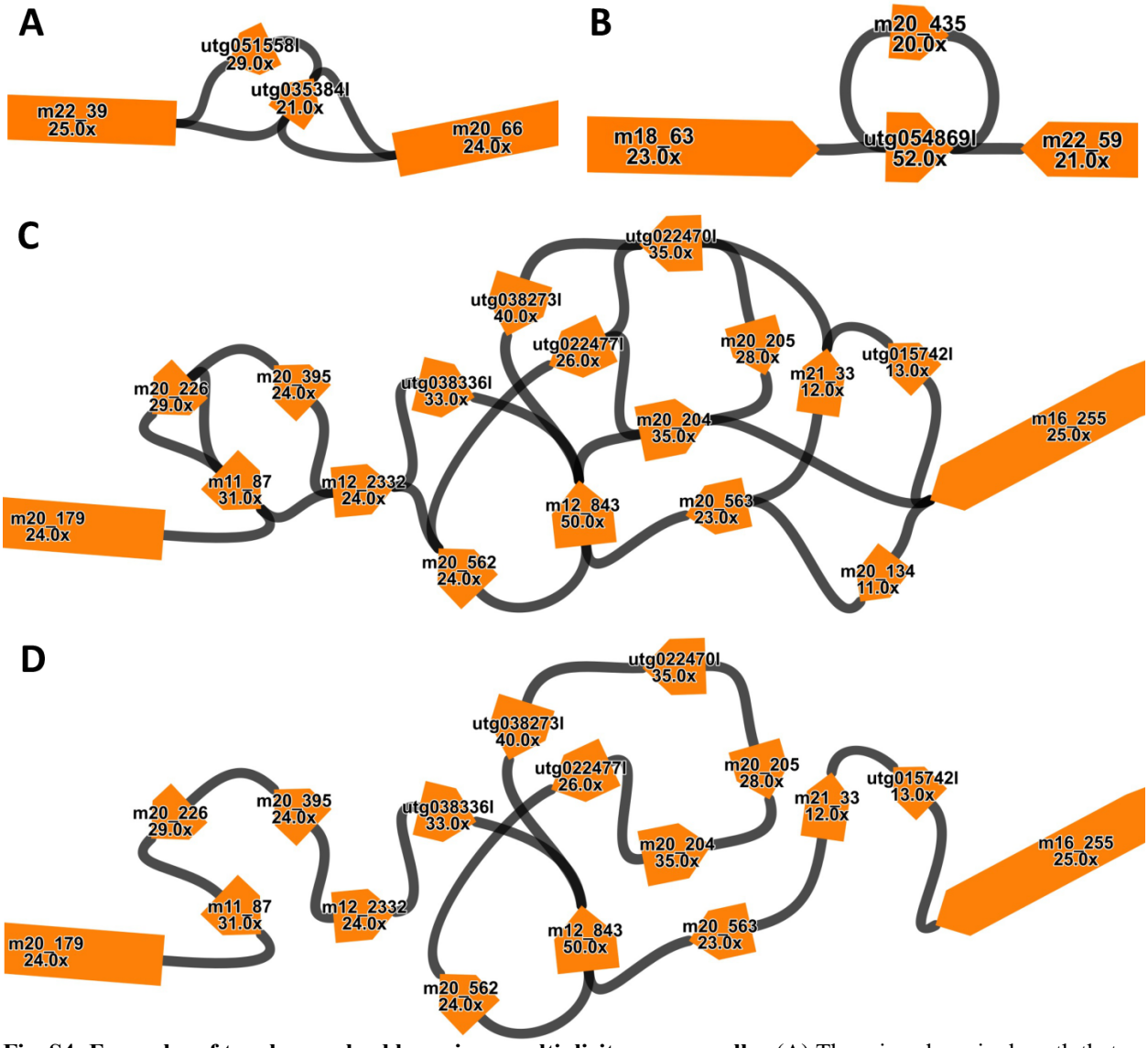


Fig. S4: Examples of tangles resolved by unique multiplicity-aware walks. (A) There is only a single path that goes through all the nodes exactly once, as would be expected given their coverage relative to the large unique nodes. The path is m22_39-,utg051558l-,utg035384l-,m20_66+. (B) Given the two-fold increase in the coverage of a single node, the genomic reconstruction should traverse the loop once. (C-D) Subgraph corresponding to the AMY region on chr1. (C) The original graph is shown without simplification from the automated graph construction pipeline. (D) We start by removing the heterozygous (4 bp insertion) variant represented by m20_134. We proceed with the iterative removal of edges that can not be a part of a genomic traversal assuming that it exists in the graph and visits all nodes with coverage < 40X only once. For example, note that the node utg022477ll has a single outgoing edge (leading to the node m20_562) which we have to use for utg022477l to be a part of the traversal. Since m20_562 is expected to have multiplicity one, we can remove the other edge outgoing from it (leading to m12_2332). Repeated application of this strategy brings us to the final simplified subgraph, which has a single unambiguous traversal using node m12_843 (cov 50X) twice.

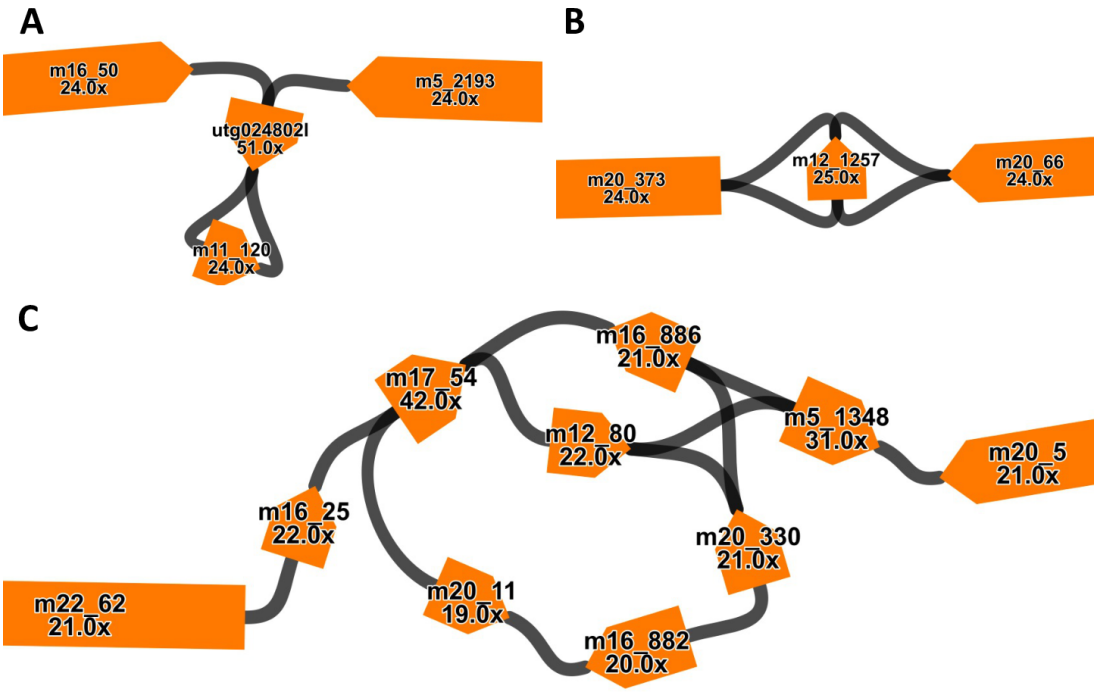


Fig. S5: Examples of tangles resolved using ONT alignments. (A) Example of a tangle that cannot be resolved based exclusively on the HiFi reads. The node m11_120 can be incorporated into traversal in either the forward or the reverse direction. While such "hairpins" can potentially represent a perfect (reverse-complement) palindrome, in all the situations that we studied this was not the case. The correct orientation of the "hairpin" node was identified based on the sequence-to-graph alignments of ONT UL reads by GraphAligner (*28*). (B) Another example with an ambiguous orientation for the center node (m12_1257). (C) An example subgraph where GraphAligner alignment paths were insufficient to unambiguously identify the genomic traversal. The tangle was resolved by generating two candidate traversals (namely m22_62-, m16_25+, m17_54+, m16_886-, m20_330-, m16_882+, m20_11-, m17_54+, m12_80+, m5_1348+, m20_5- and m22_62-, m16_25+, m17_54+, m12_80+, m20_330-, m16_882+, m20_11-, m17_54+, m16_886-, m5_1348+, m20_5-) and analyzing Winnowmap (*78*, *79*) alignments of the ONT UL reads against each of the corresponding sequences. Nodes and edges not part of the candidate traversals were excluded for clarity.

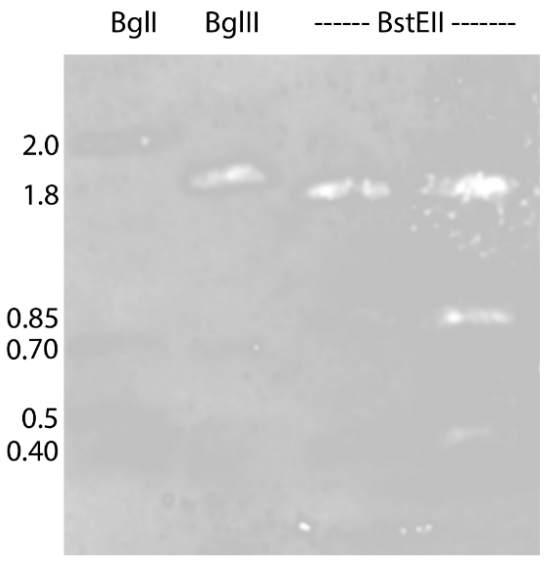


Fig. S6: PFGE digestion of DXZ1 confirms assembly size in HG002. Average estimates were 3.2 Mbp for BgII (2.0 + 0.7 + 0.5, n=3); 3.05 Mbp for BgIII (1.9 + 0.7 + 0.45, n=1); and 3.05 Mbp for BstEll (1.8 + 0.85 + 0.40, n=4). The digest sizes were consistent with an in-silico digest of the assembly.

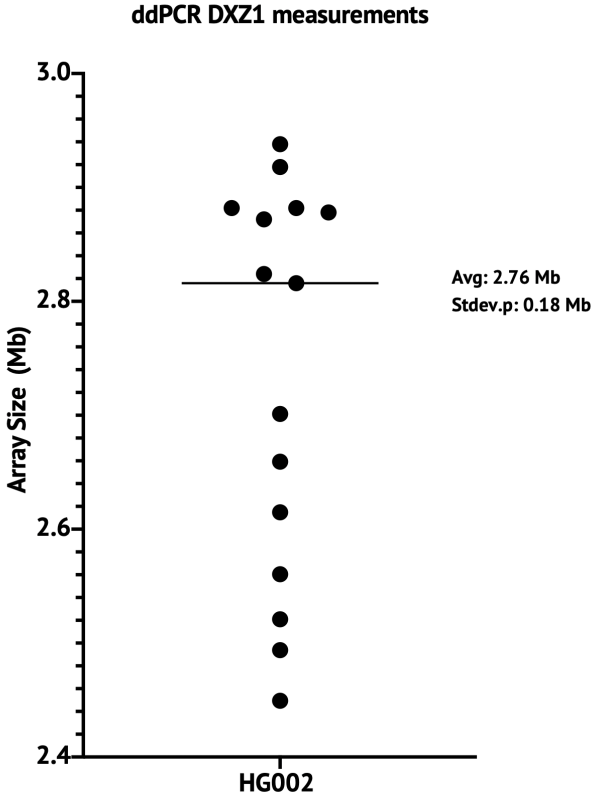


Fig. S7: ddPCR estimate of 1N DXZ1 copy number in HG002. Mean ± s.d. is shown. Droplet digital PCR rDNA copy number estimates for HG002 cell line calculated by DXZ1 measurements normalized to the single copy gene HPRT1 is consistent with the assembly size of 3.1 Mbp.

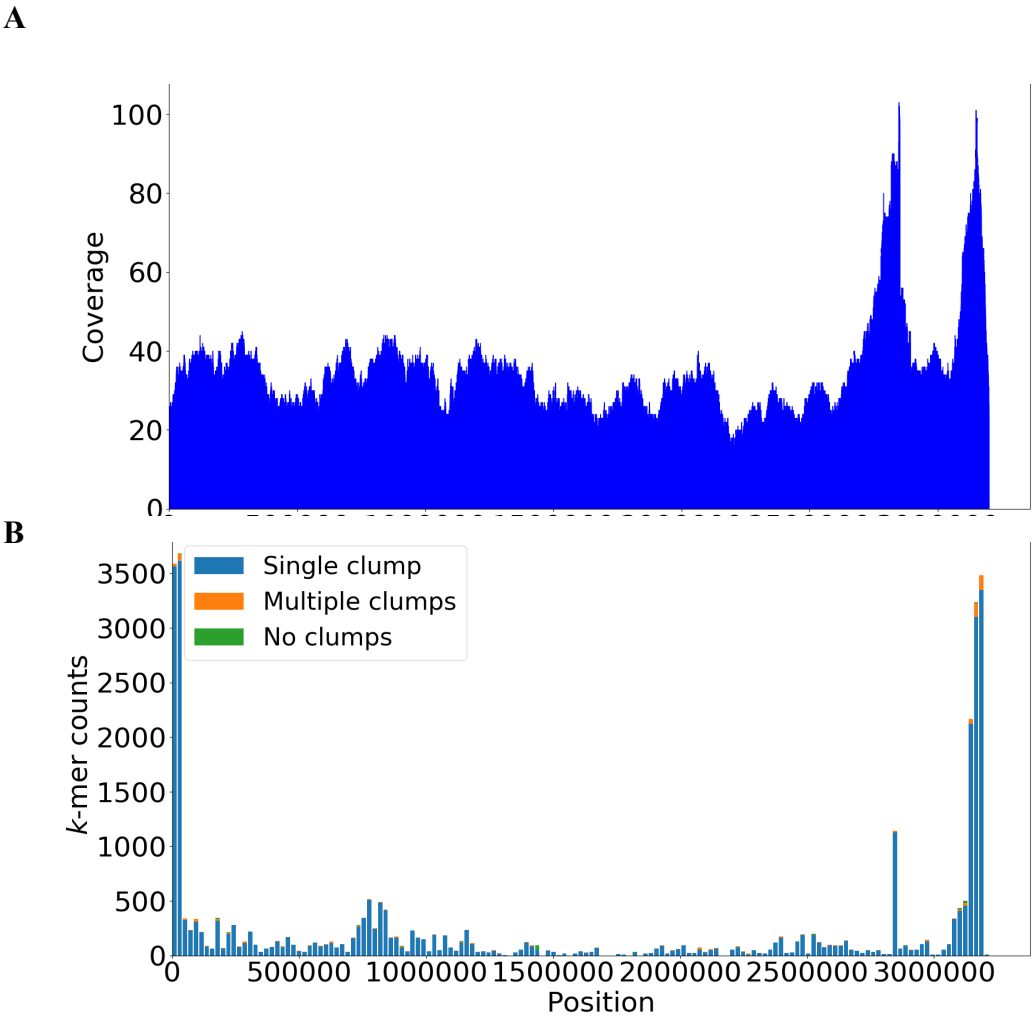


Fig. S8: TandemTools validation of CENX in HG002. (A) Coverage by ONT mapped reads to the assembly, with the position set to 0 at the start of the centromeric satellite. There are two coverage peaks at 2.8 and 3.1 Mbp corresponding to LINE elements but the rest of the array shows even coverage. (B) k-mer validation of the centromeric array. Pileups of multiple clumps, as defined by TandemTools (*37*). which would be indicative of mis-assemblies, are absent from the figure.

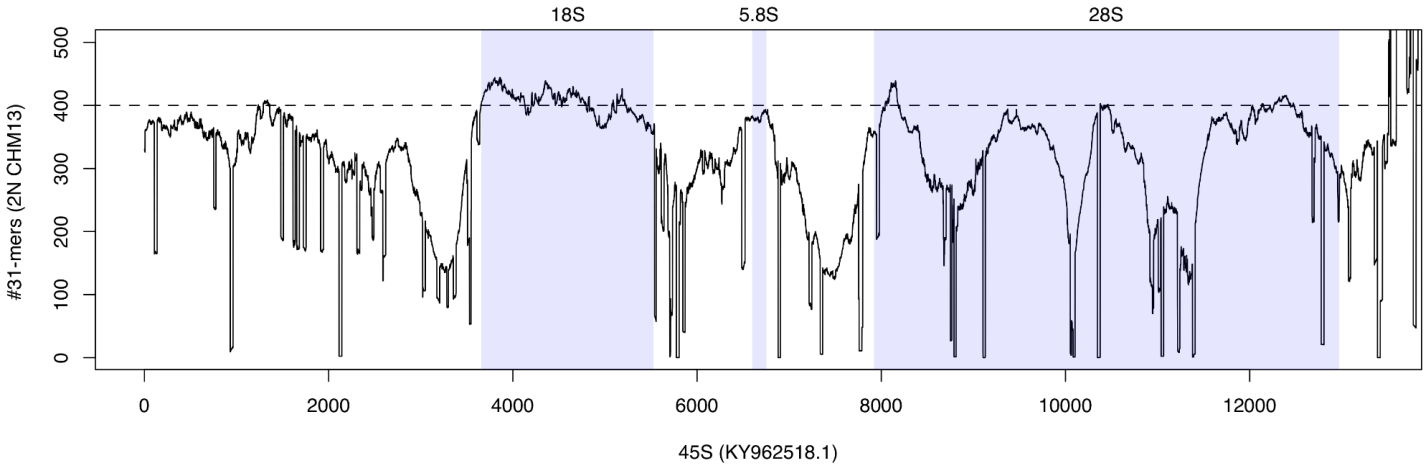


Fig. S9: 45S *k*-mer count distribution in CHM13. Diploid copy number was estimated from ILMN *k*-mers. The x-axis shows the canonical rDNA unit, and the y-axis the estimated (diploid) count of those *k*-mers in the CHM13 genome. Total ILMN sequencing depth was estimated to be 98X based on the modal *k*-mer copy number (Fig. S20), or 49X per haplotype, so the diploid copy number was estimated as the raw *k*-mer count divided by 49. Sharp drops in copy number correspond to small variants, while the smoother dips are due to larger indels and %GC bias in the ILMN data. The 18S gene shows relatively uniform coverage due to its high degree of conservation in the genome and more typical %GC (56%). This plot confirms a diploid copy number of ~400 rDNA units, or ~200 per haplotype, in the CHM13 genome, consistent with the FISH and ddPCR estimates.

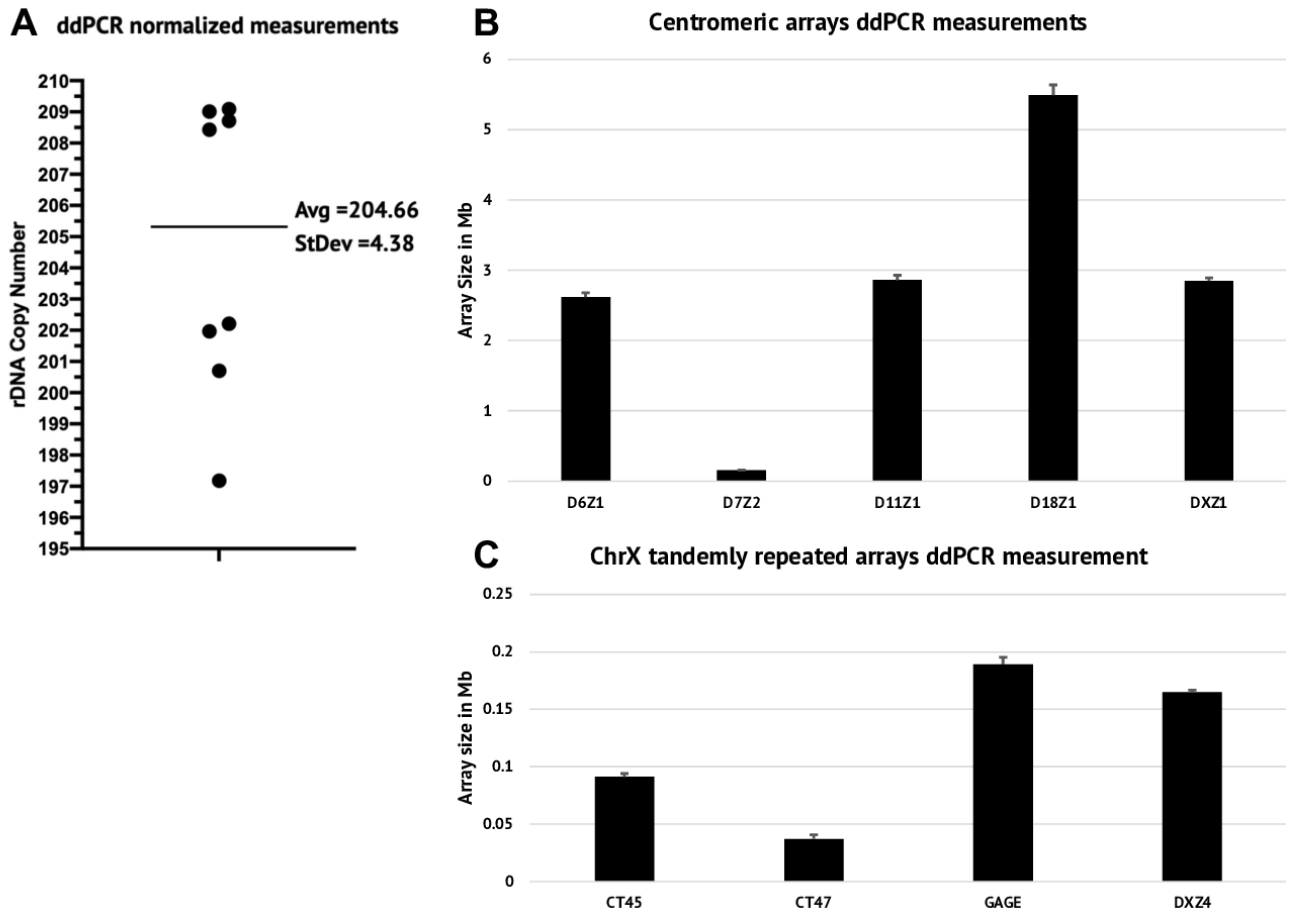


Fig. S10: ddPCR estimate of 1N copy number in CHM13. (A) Droplet digital PCR rDNA copy number estimates for CHM13 cell line calculated by 28S measurements normalized to the single copy gene TBP1. Mean ± s.d. is shown. The estimate is consistent with the FISH estimates in Fig. S1. (B) Higher-order repeat (HOR) sizing from ddPCR. All estimates are consistent with annotated (*30*) sizes in the assembly (D6Z1 = S1C6H1L = 2.77 Mbp; D7Z2 = S5C7H2 = 0.20 Mbp; D11Z1 = S3C11H1L = 3.38 Mbp; D18Z1 = S2C18H1L = 4.97 Mbp; DXZ1 = S3CXH1L = 3.11 Mbp). (C) Tandem repeat sizes on Chromosome X from ddPCR. All estimates are consistent with the previously published (*14*) and current reconstruction (CT45 = 0.15 Mbp; CT47 = 0.04 Mbp; GAGE = 0.19 Mbp; DXZ4 = 0.17 Mbp).

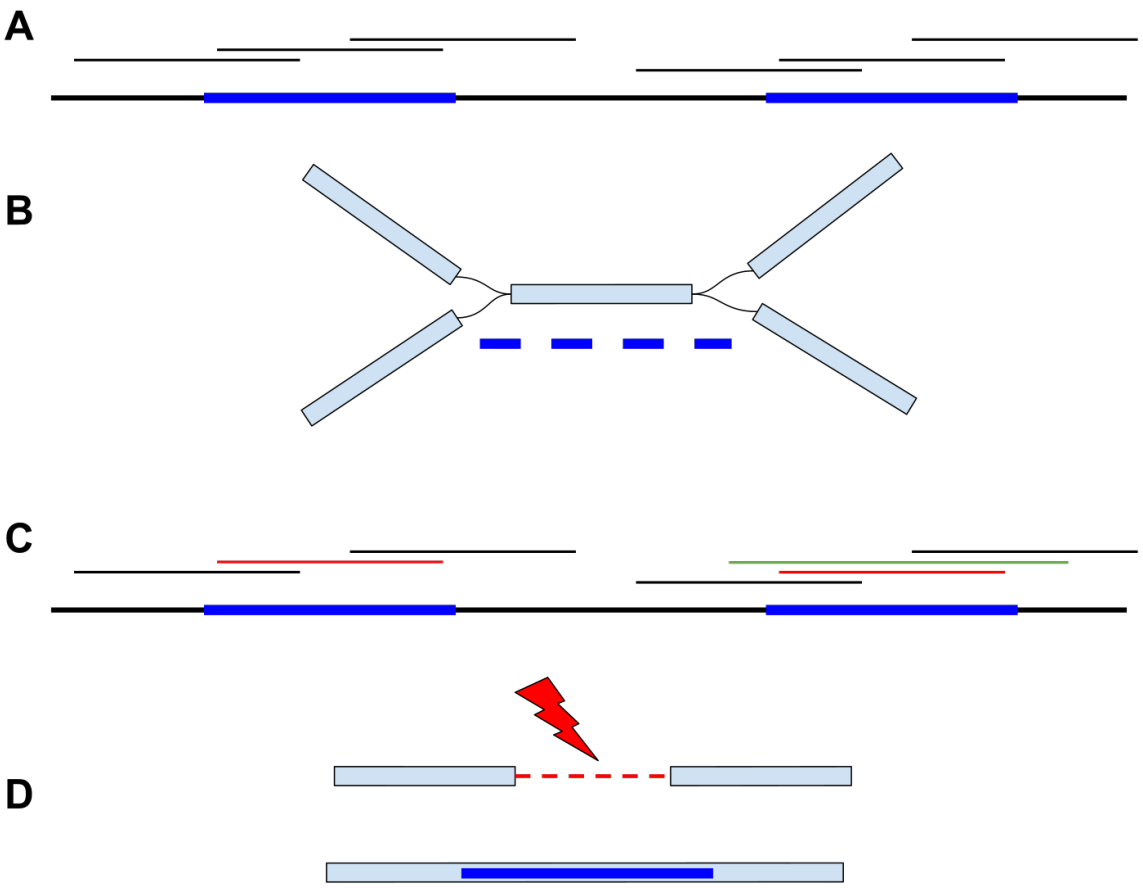


Fig. S11: Fragmentation of a string graph constructed from reads of varying lengths. (A) A genome with a two-copy repeat (shown in blue) sequenced by equal-length reads denoted by thin black lines. (B) Schematic bidirectional condensed string graph constructed from the input reads showing an unspanned repeat with branching. (C) Same genome sequenced by reads of varying lengths, the longer green read spanning one of the blue repeat copies contains reads originating from both repeat instances (red), leading to their exclusion from string graph construction. (D) The string graph constructed from this read set has a unitig spanning one of the repeat copies, while the second repeat copy is not represented by any continuous path. Note that more serious assembly issues might follow if the arising "dead-end" unitigs at the repeat boundary are further discarded by graph simplification procedures.

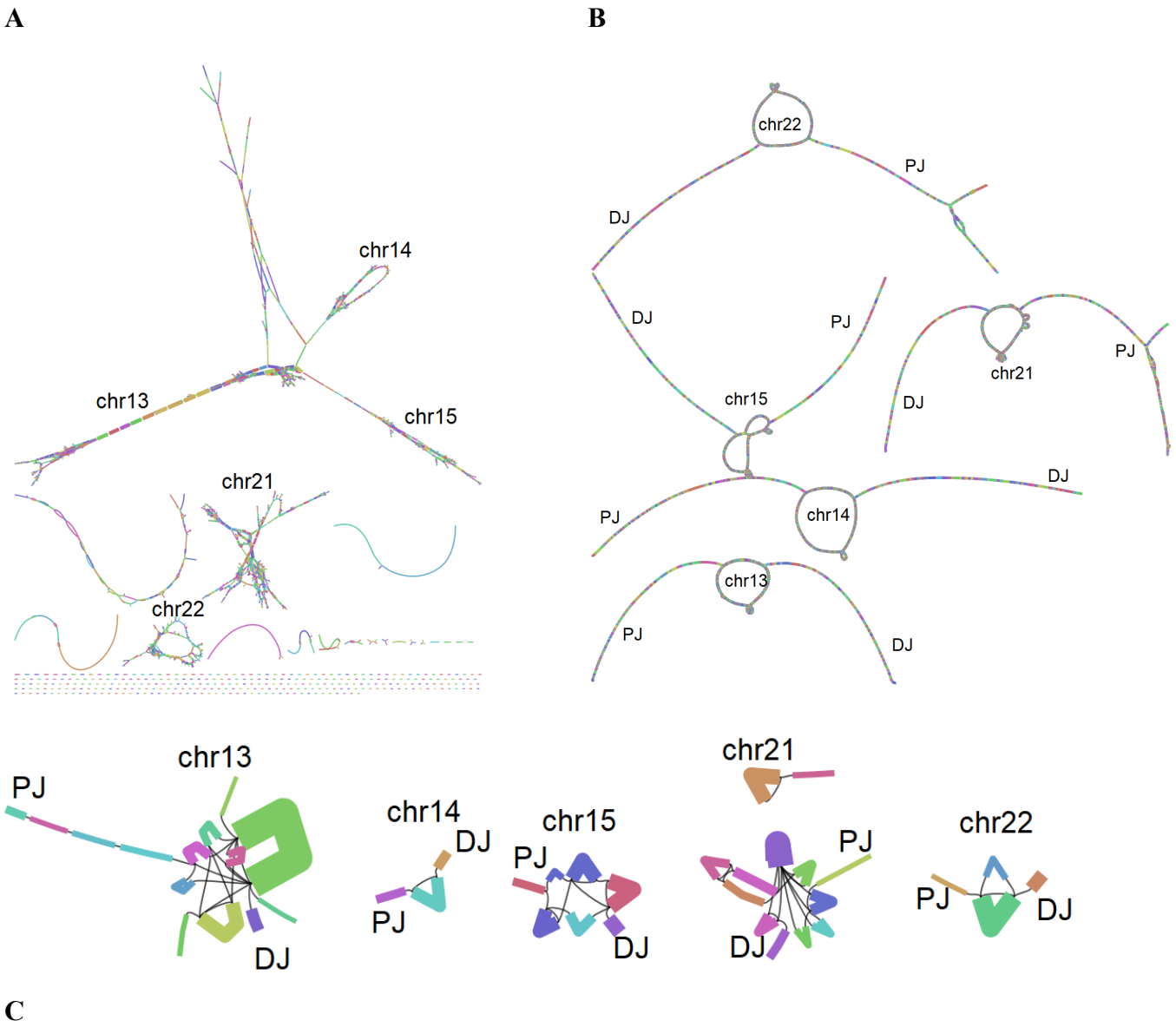


Fig. S12: Overview of rDNA assembly process. (A) A minimizer graph of rDNA HiFi reads with k=3501. Due to chromosome specific variation, the graph has five large clusters, one per chromosome. The chromosome specific variation enables clustering the HiFi reads by chromosome based on shared k-mers between the reads and the sequences of the nodes in each cluster. The graph is fragmented due to systematic coverage bias and sequencing errors in the HiFi reads, especially within the GA-rich regions of the intergenic spacer. (B) Minimizer graphs built from chromosome specific HiFi reads with k=201. The loop structure represents the rDNA array, including both morph-specific variation and sequencing errors. (C) A graph built from clustered ONT reads. Each node represents one cluster, and the thickness of the nodes corresponds to ONT read coverage. Edges are added whenever at least two ONT reads support a connection between the clusters. The clusters recover structurally unique rDNA morphs, although morphs with only minor SNP variation may be collapsed to the same cluster. The graph shows that chromosomes 14 and 22 consist of one major morph repeating multiple times while chromosomes 13, 15, and 21 have a more complex mosaic structure.

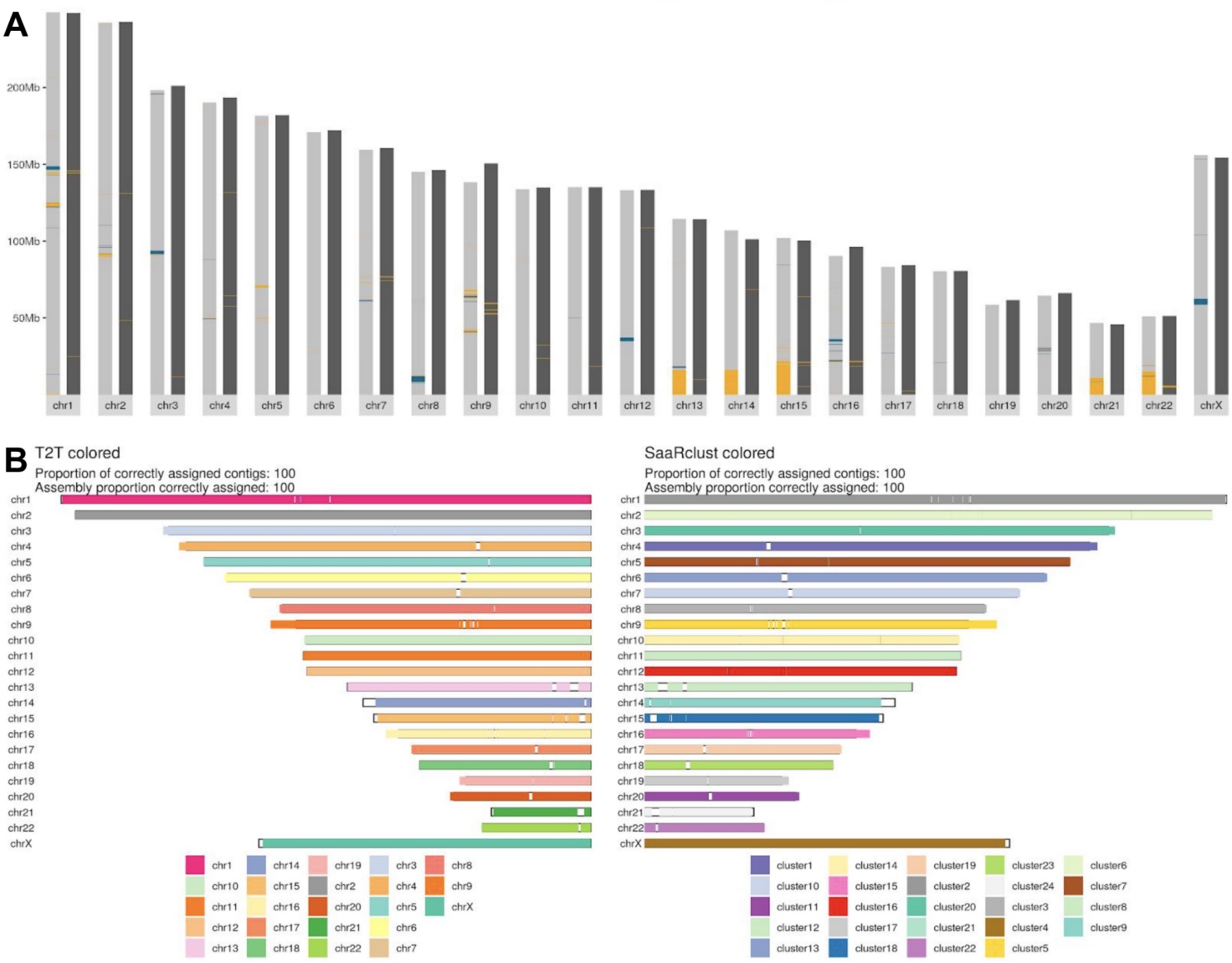


Fig. S13: Strand-seq supports correct orientation and chromosome assignments. (A) Comparison of GRCh38 and a preliminary version of the CHM13 assembly. GRCh38 is plotted as light gray bars and CHM13 as dark gray bars. Changes in assembly directionality (shown in blue) might point to an unresolved inversion or misoriented contigs. Other highlighted assembly features are low mappability (or possibly collapsed) regions (shown in yellow). There are few inverted regions in CHM13 in contrast with GRCh38 (0.0002 Mbp vs 28.41 Mbp). GRCh38 has a high count of low mappability regions, including the short arms of all acrocentric chromosomes which correspond to large gaps. In contrast, in CHM13, the low mappability regions are reduced to 8.99 Mbp from 101.10 Mbp in GRCh38. (B) Ideogram shows assembled contigs aligned to CHM13 colored based on cluster identity determined by SaaRclust (*106*, *127*). In an ideal scenario there is a single color per chromosome. The left ideogram is colored by contig names in the assembly. As expected, there is a single color per chromosome. Right ideogram is colored based on the cluster ID assigned to each 200 kbp window in a contig. In case of large scale chromosomal mis-joins, we expect to see two or more colors assigned to a contig. None of the CHM13 contigs show this, all being consistent with a single chromosome bin.

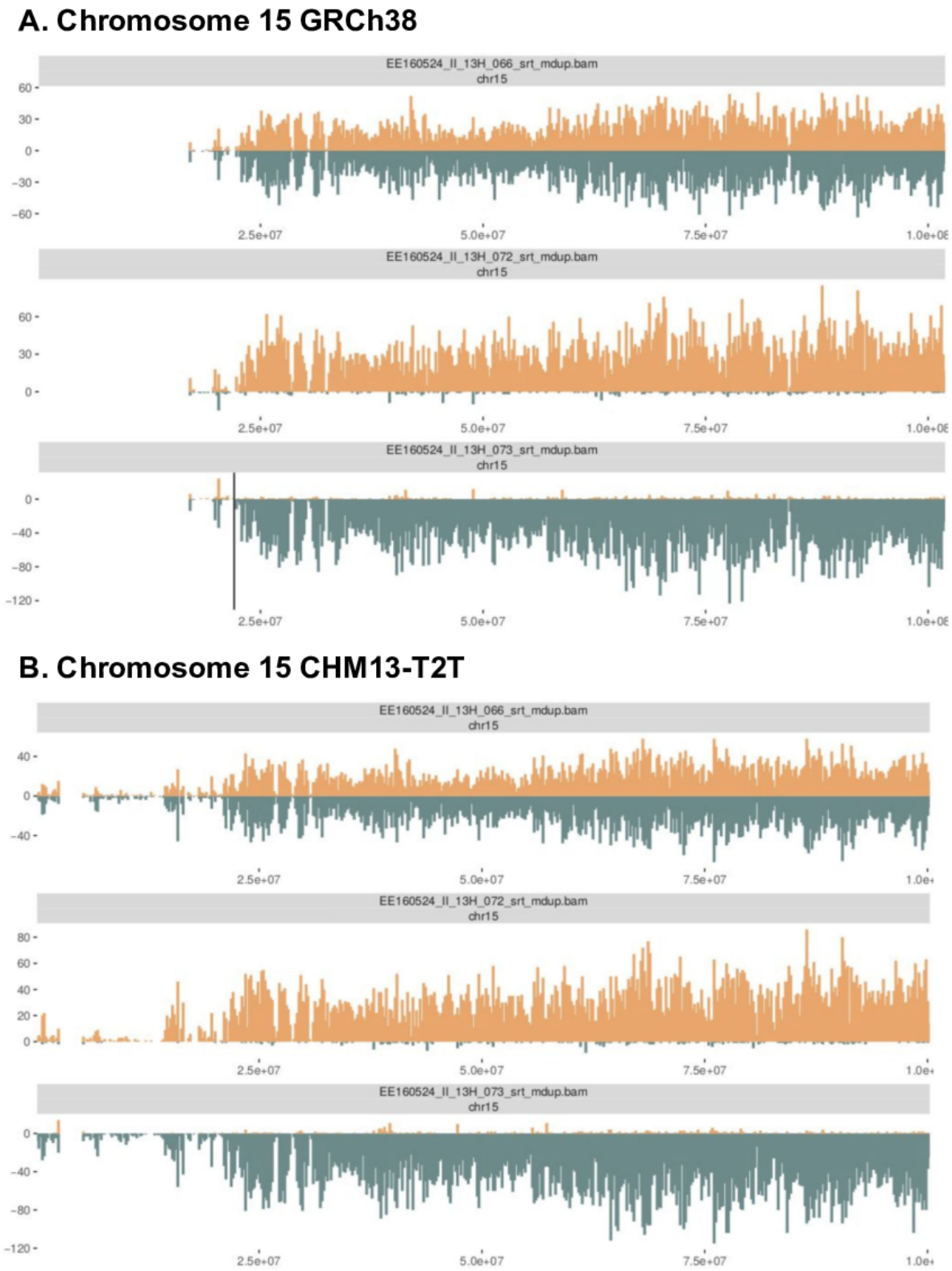


Fig. S14: Mapping of short Strand-seq reads in rDNA loci of Chromosome 15. We visualize the mapping of Strand-seq reads to Chromosome 15 of GRCh38 (A) and CHM13 (B). For the same three Strand-seq libraries we count Strand-seq reads mapping to either plus strand (teal) or minus strand (orange) of the reference genome (GRCh38 or CHM13) in 200 kbp bins reported as vertical bars (plus - below zero; minus - above zero). Only reads with MAPQ >10 are plotted. As expected, there are no reads mapping to GRCh38 from 0-15 Mbp. In contrast, there are enough reads mapping to CHM13 to confirm the orientation and assignment of the short arm of this chromosome.

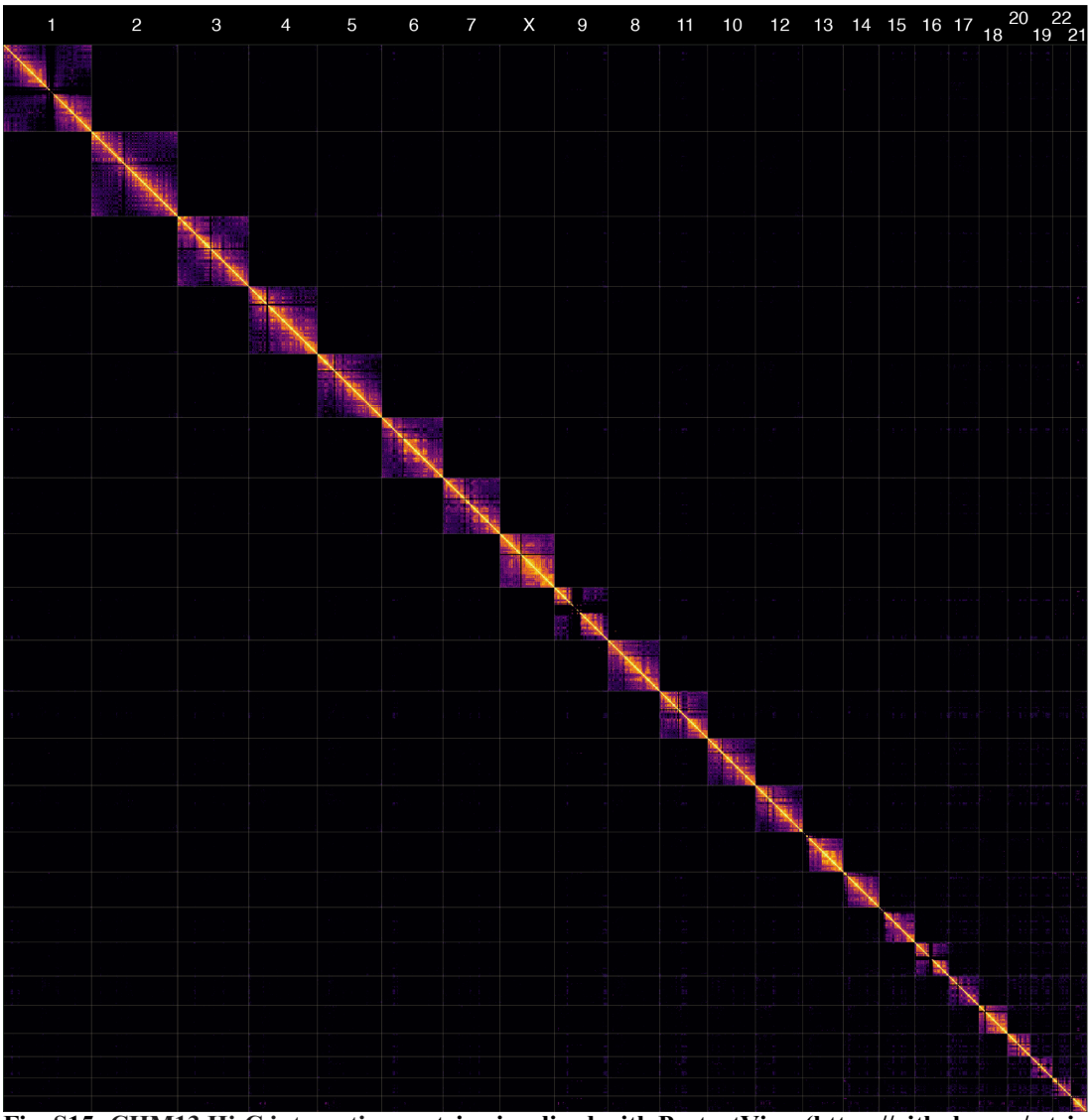


Fig. S15: CHM13 Hi-C interaction matrix visualized with PretextView (https://github.com/wtsi-hpag/PretextView). There are strong signals within chromosomes. There is no off-diagonal or no chross-chromosomal signal indicating the chromosomes are complete and not chimeric. Black regions within chromosomes (e.g. Chromosomes 1 or 9) indicate repetitive regions where no short-reads could be confidently anchored. The short arms of the acrocentric chromosomes (13, 14, 15, 21, 22) show sufficient signal to confirm the correct assignment of each short arm to the chromosome.

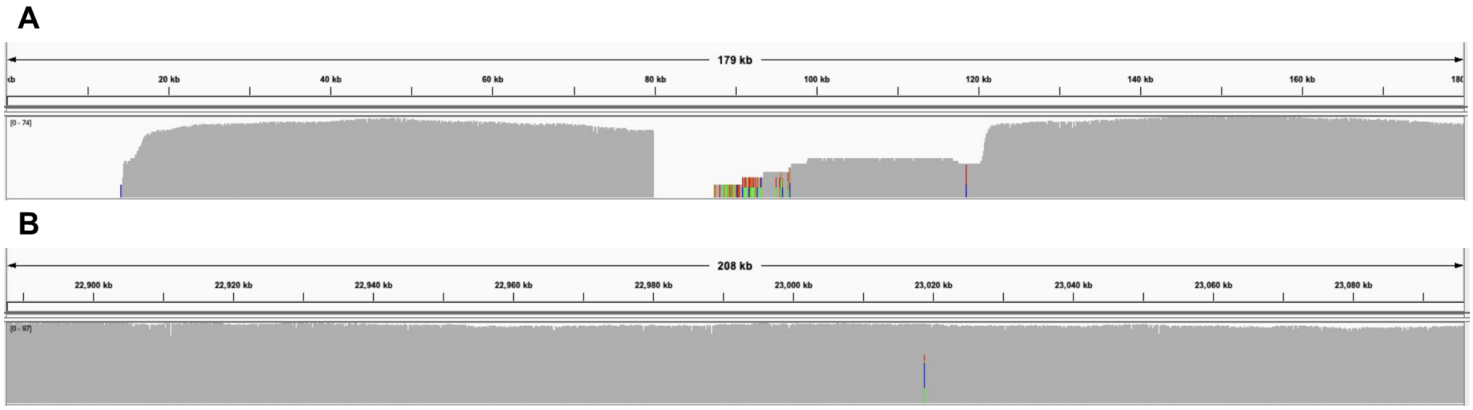


Fig. S16: Concordance of ONT reads with BAC AC279506.1 and corresponding assembly region. This 180,858 bp long BAC aligned incompletely, breaking at 20 kbp (BAC:19,686-180,858 aligned to chr17:22,937,576-23,096,790 in rev strand). (A) Alignments for the first 100 kbp of the BAC show low coverage at the start (before 20 kbp) with an increase in variant positions. There is also low coverage in the BAC at 80 kbp. (B) In contrast, the corresponding position in the assembly (chr17:22,900,000-23,080,000) shows no such coverage dropouts and no increase in variant calls.

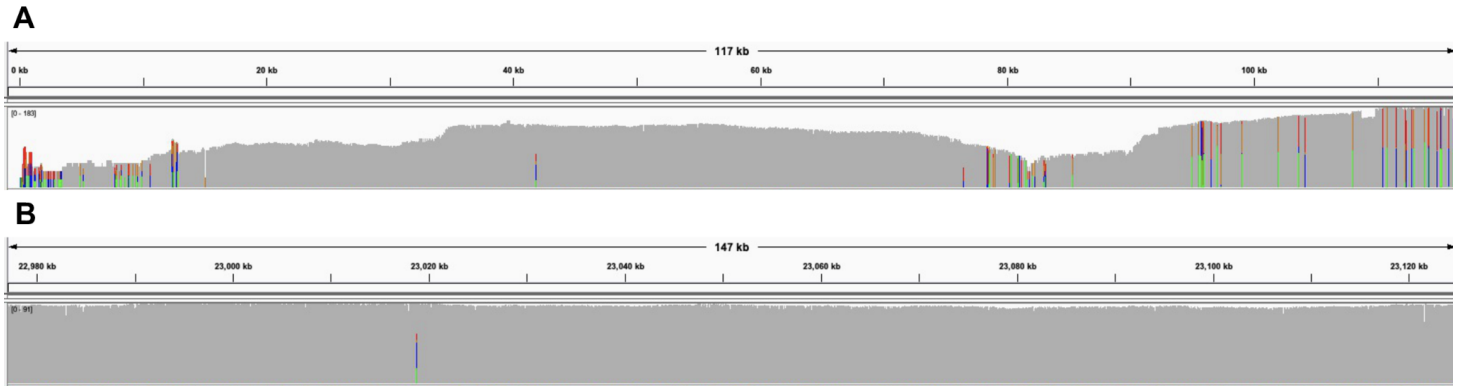


Fig. S17: Concordance of ONT reads with BAC AC279581 and corresponding assembly region. This 117,950 bp long BAC aligned incompletely, breaking at 94 kbp (BAC:25-93,747 to chr17:22,9770,000-23,070,901 in fwd strand). (A) Alignments for the BAC show low coverage at the start with another drop in at 80-90 kbp and an increase in variant positions after 90 kbp. (B) In contrast, the corresponding position in the assembly (chr17:22,980,000-23,120,000) shows no such coverage dropouts and no increase in variant calls.

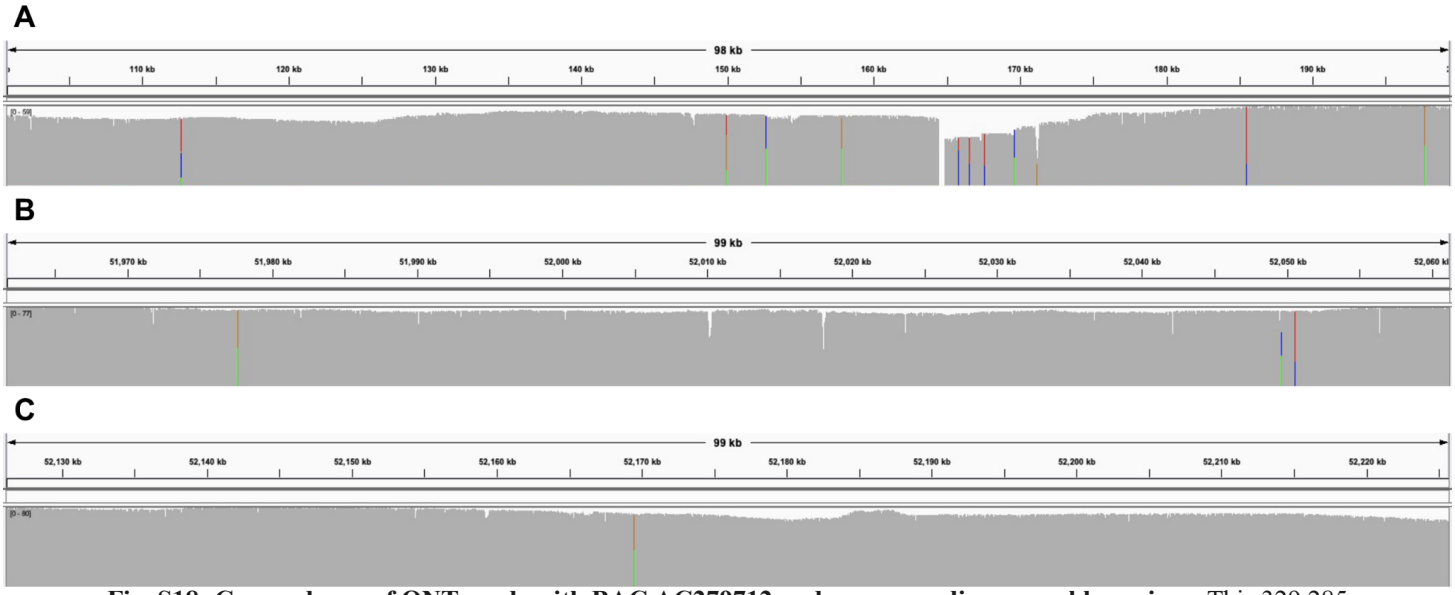


Fig. S18: Concordance of ONT reads with BAC AC279712 and corresponding assembly regions. This 329,285 bp long BAC aligned in two pieces to the same location in the assembly (BAC:4-164,475 and BAC:164,810-329,291 aligned to chrX:52,011,667-52,176,139 in fwd strand). (A) Alignments for the 110-190 kbp region of the BAC show a dropout in coverage at 165 kbp, corresponding to the break between the two alignments. (B) The start position of the alignment in the assembly (chrX:51,970,000-52,060,000) shows no such coverage dropouts and no increase in variant calls. (C) The end position of the alignment in the assembly (chrX:52,130,000-52,220,000) shows even coverage and no variant artifacts to indicate a repeat collapse.

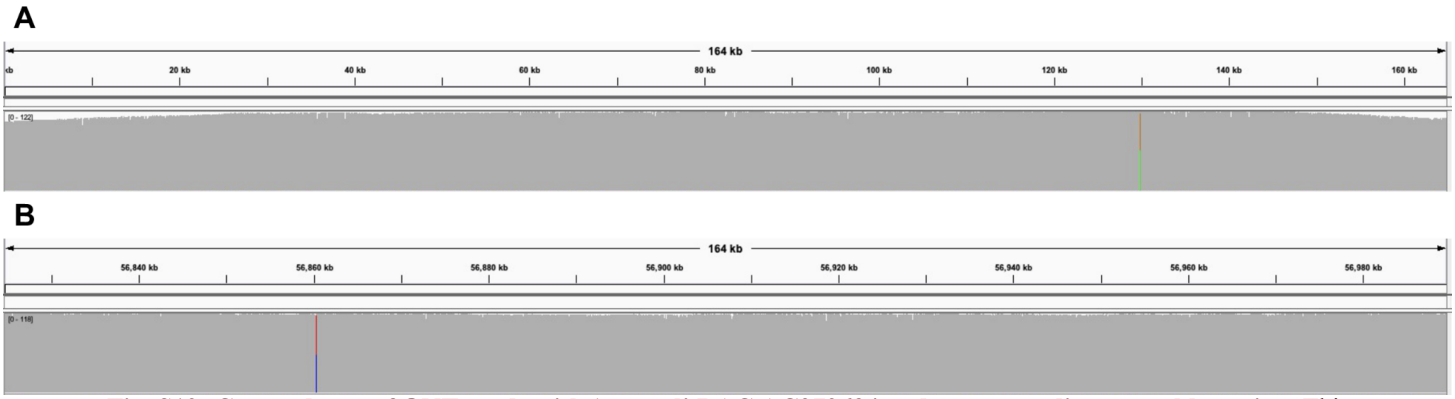


Fig. S19: Concordance of ONT reads with 'control' BAC AC279634 and corresponding assembly region. This 165,682 bp long BAC aligned in one piece (BAC:0-165,682 to chrX:56,824,611-56,990,290 in rev strand). (A) Alignments for the BAC show even coverage and no variant positions. (B) The corresponding position in the assembly (chrX:56,830,000-56,980,000) shows even coverage and no increase in variant calls.

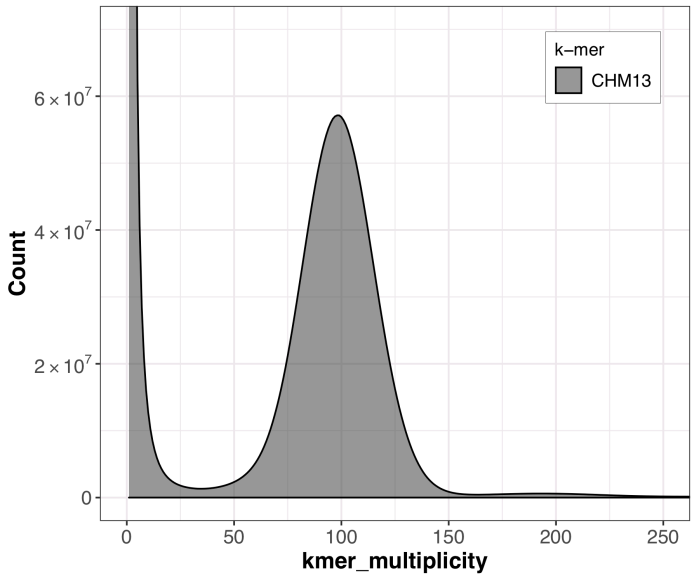


Fig. S20: Distribution of k-mer counts from PCR-free Illumina data. *k*-mer multiplicities were computed with meryl (*81*) and plotted with merqury (*81*).

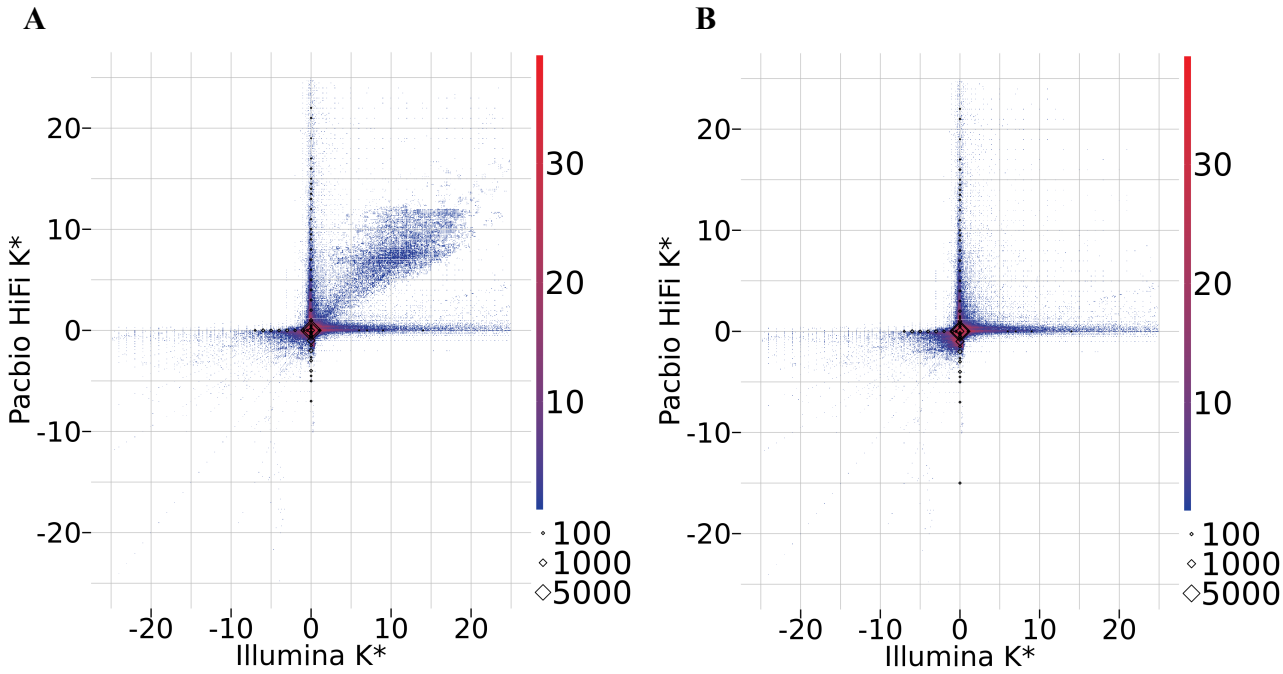


Fig. S21: Merfin k* statistics showing assembly concordance with HiFi and PCR-free sequencing data. Positive values indicate the k-mers have a higher copy in the sequencing data than the assembly (collapse in the assembly) while negative values indicate a lower copy in the sequencing data (expansion in the assembly). (A) *k*-mer multiplicities within v1.0 are mostly consistent with HiFi and PCR-free data (k* close to 0), with a cloud of *k*-mers collapsed in the assembly. (B) corresponding plot for the v1.1 assembly indicates that the collapsed rDNA regions in the assembly have been correctly resolved.

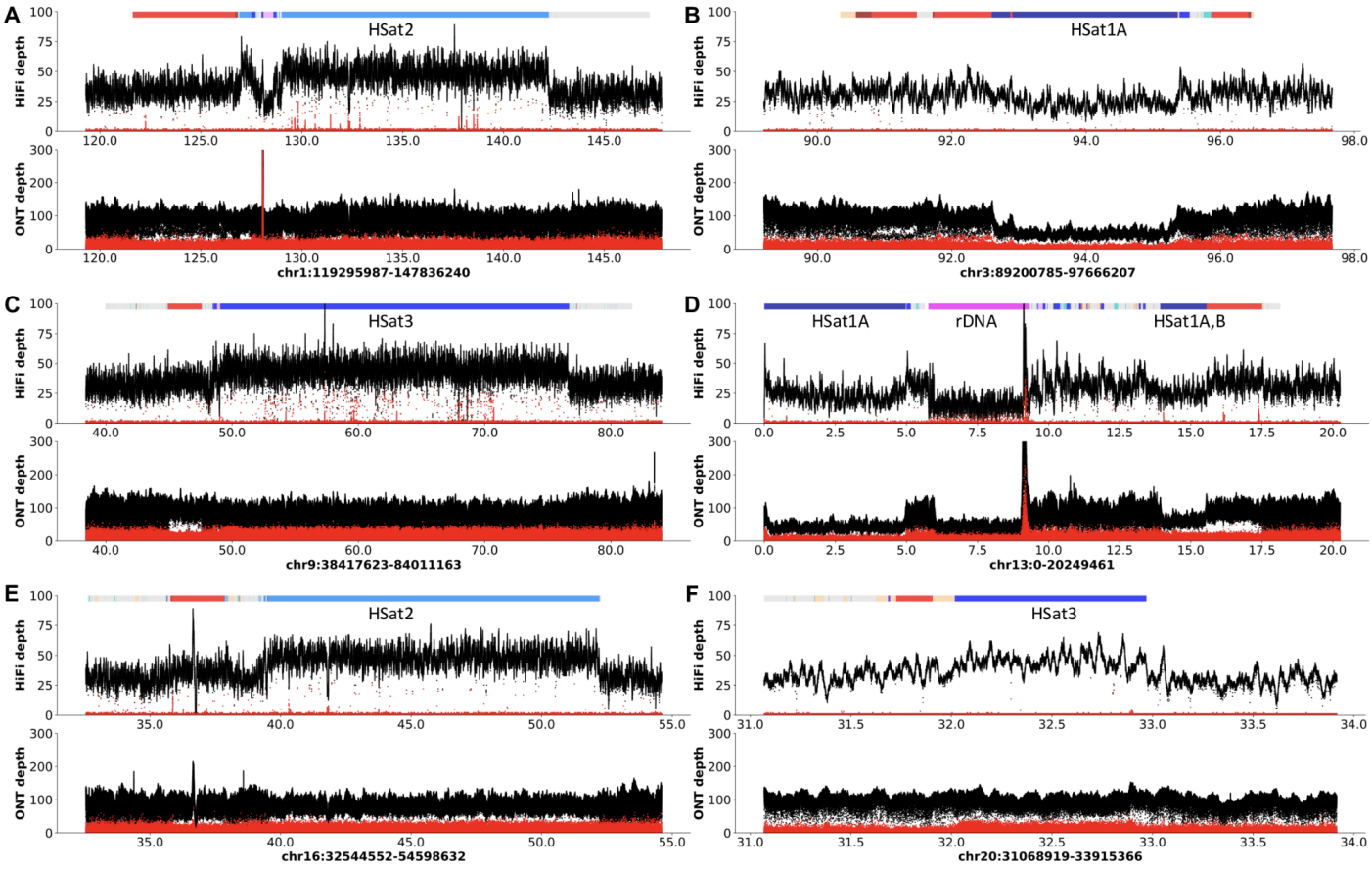
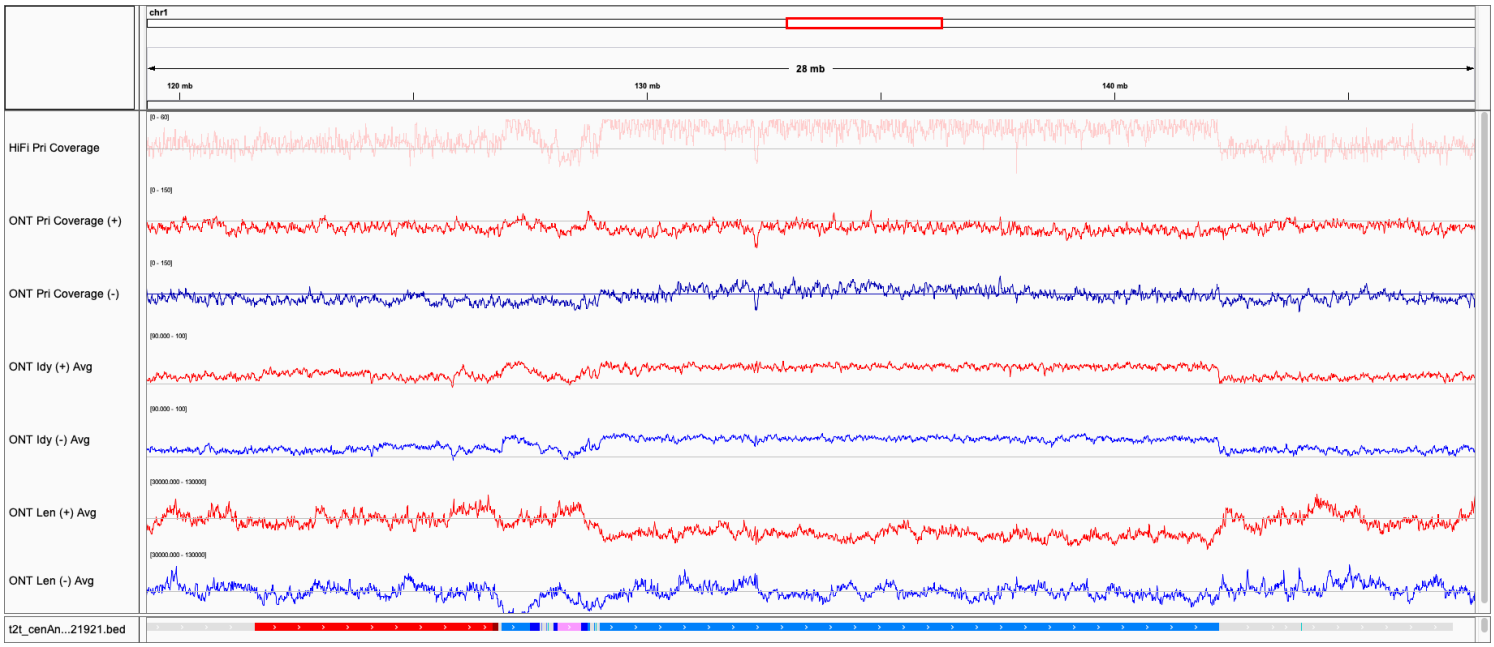


Fig. S22: HiFi alignments for genomic regions containing HSat arrays (with repeat annotation track showing HSat1 and HSat3 in dark blue, HSat2 in light blue, and AlphaSat in red). Read coverage for HiFi and ONT mapped data is shown on the Y axis in black and second most frequent allele (not in the assembly) in red. Assembly position is shown on the X axis in Mbp. (A) Chr1:119295987-147836241 with the HSat2 region in light blue (B) Chr3:89,200,785-97,666,208 with HSat1A in dark blue (C) Chr9:38,417,623-84,011,164 with HSat3 in dark blue (D) Chr13:0-20,249,462 with HSat1A and HSat1A,B regions in dark blue and rDNA in pink (E) Chr16:32,544,552-54,598,633 with HSat2 region in light blue (F) Chr20:31068919-33915366 with the HSat3 region in dark blue. Repeat collapse events are commonly manifested by spikes of alternative base frequency, arising when reads originating in non-identical genomic repeat copies get mapped to the same assembly region (*38*). The low number of alternative-base frequency spikes across most HSat2/3 arrays (with the only exception of the HSat3 array on chr9 and the model-based rDNA on chr13), comparable with their density in the surrounding regions, indicates that the elevated coverage is unlikely to arise from the collapse of genomic repeat copies within the assembly. In contrast with HiFi, ONT read alignments do not display elevated coverage across the HSat2/3 arrays on Chr1, Chr9, Chr16, and Chr20 and reveal even more depletion in HSat1A array coverage on Chr3 and Chr13. The inconsistency in coverage between sequencing technologies is indicative of technology biases. While HSat3 on Chr9 shows an elevated rate of alternative-base frequency spikes we note that assembly of this region is difficult to evaluate as it is one of the most repetitive regions in the genome (see 'Mappability' section) and many artifacts may stem from mapping ambiguities rather than assembly issues.

Fig. S23: HiFi and ONT read statistics within HSat1 regions on Chr3 and Chr13. (A) read statistics for the HSat1A region on chr3 (same region as on Fig. S22B, satellite array marked in dark blue in the annotation track). Top to bottom: coverage depth by 'primary' HiFi alignments, coverage depth by 'primary' ONT alignments by strand, identity of ONT read alignments by strand, and ONT read length by strand. Statistics were averaged in non-overlapping 10 kbp windows (see 'Alignment-based validation' section). The gray lines correspond to genome-wide means for each track. HiFi identity and read lengths are excluded because of their near-uniform identity and the tight library selection prior to sequencing. (B) Same statistics within HSat1A and HSat1A,B regions on chr13 (same as in Fig. S22D, satellite array marked in dark blue in the annotation track). Coverage depletion can be seen across both ONT and HiFi reads for all shown satellite arrays. Note the decreased read length of the ONT reads mapping to the satellite arrays.

A



B

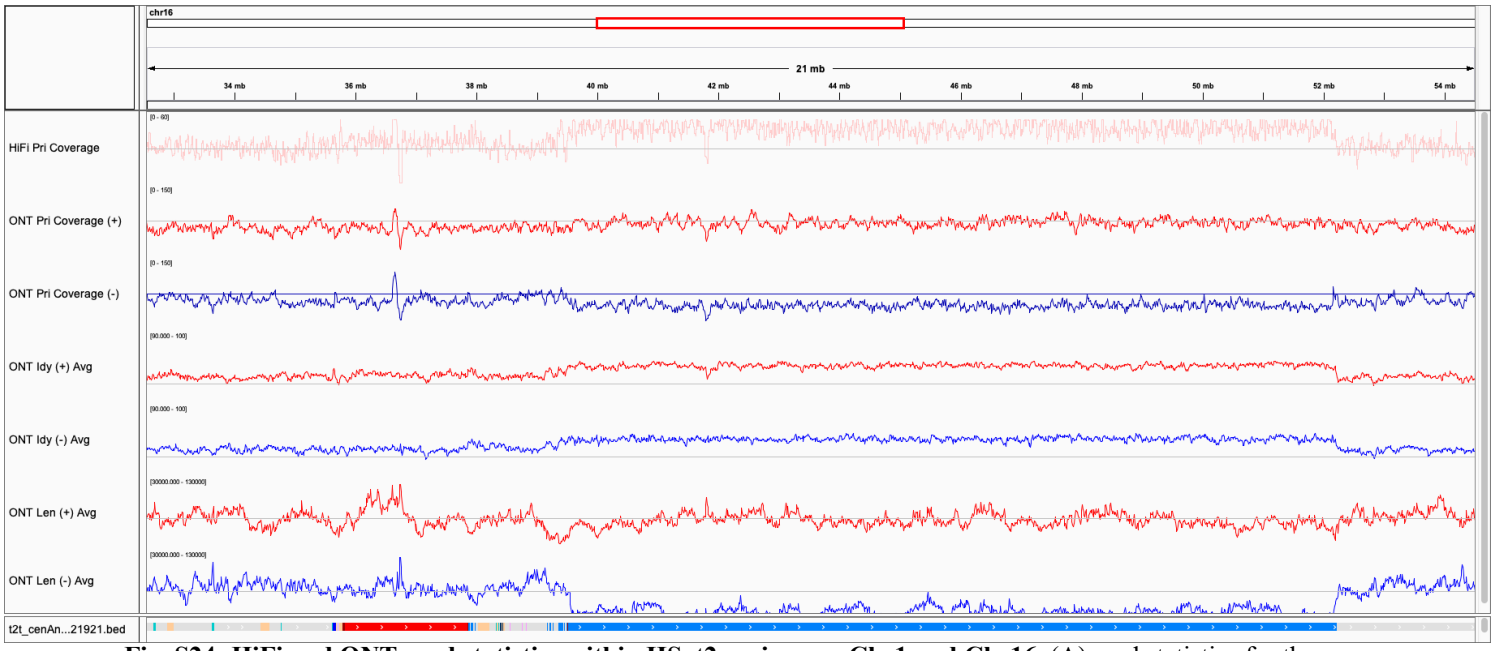
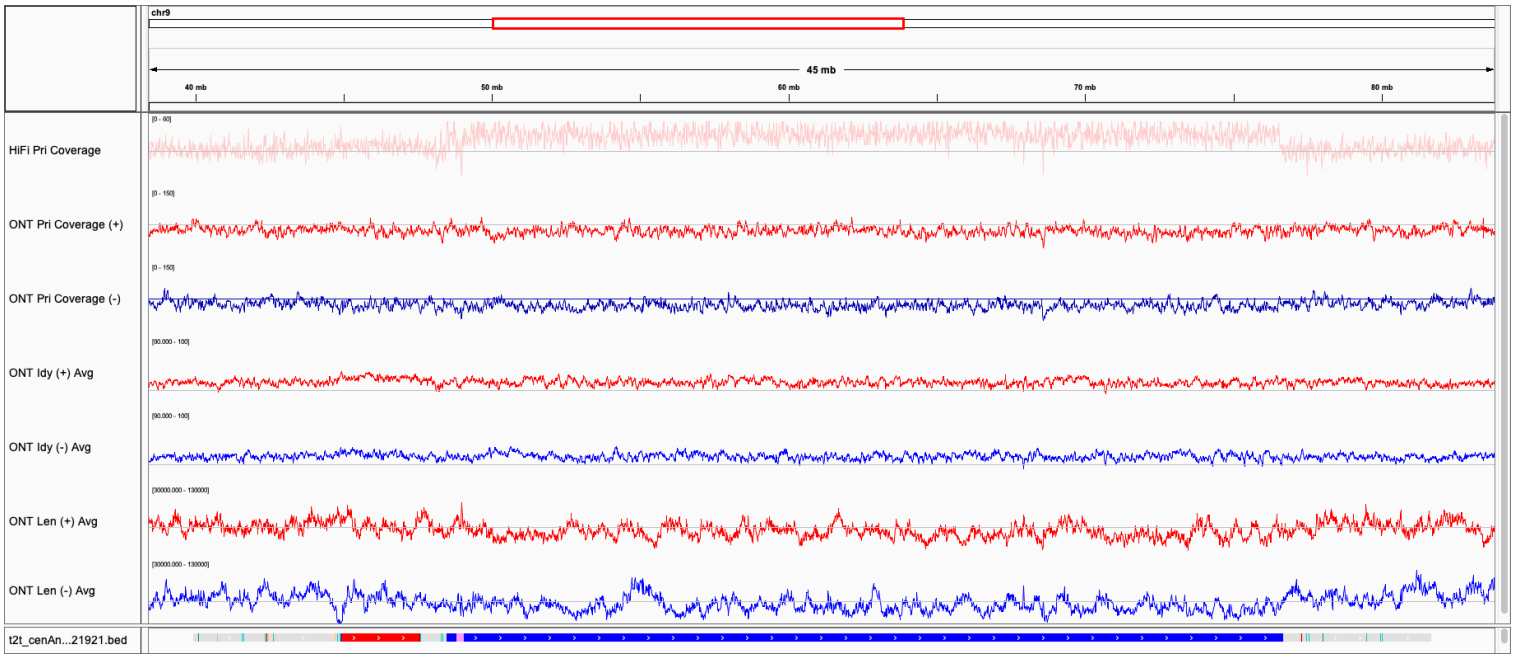


Fig. S24: HiFi and ONT read statistics within HSat2 regions on Chr1 and Chr16. (A) read statistics for the HSat2 region on Chr1 (same region as on Fig. S22A, satellite array marked in light blue on the annotation track). Note that the HSat2 array on Chr1 is in reverse orientation, with respect to the canonical unit, so its 'forward' strand corresponds to the 'reverse' strand of the array on Chr16. (B) Same statistics within the HSat2 region on Chr16 (same region as on Fig. S22E, satellite array marked in light blue on the annotation track). Note the drop of ONT read length on one strand (the negative strand of the repeat array) and an increase in identity on both strands.

A



B

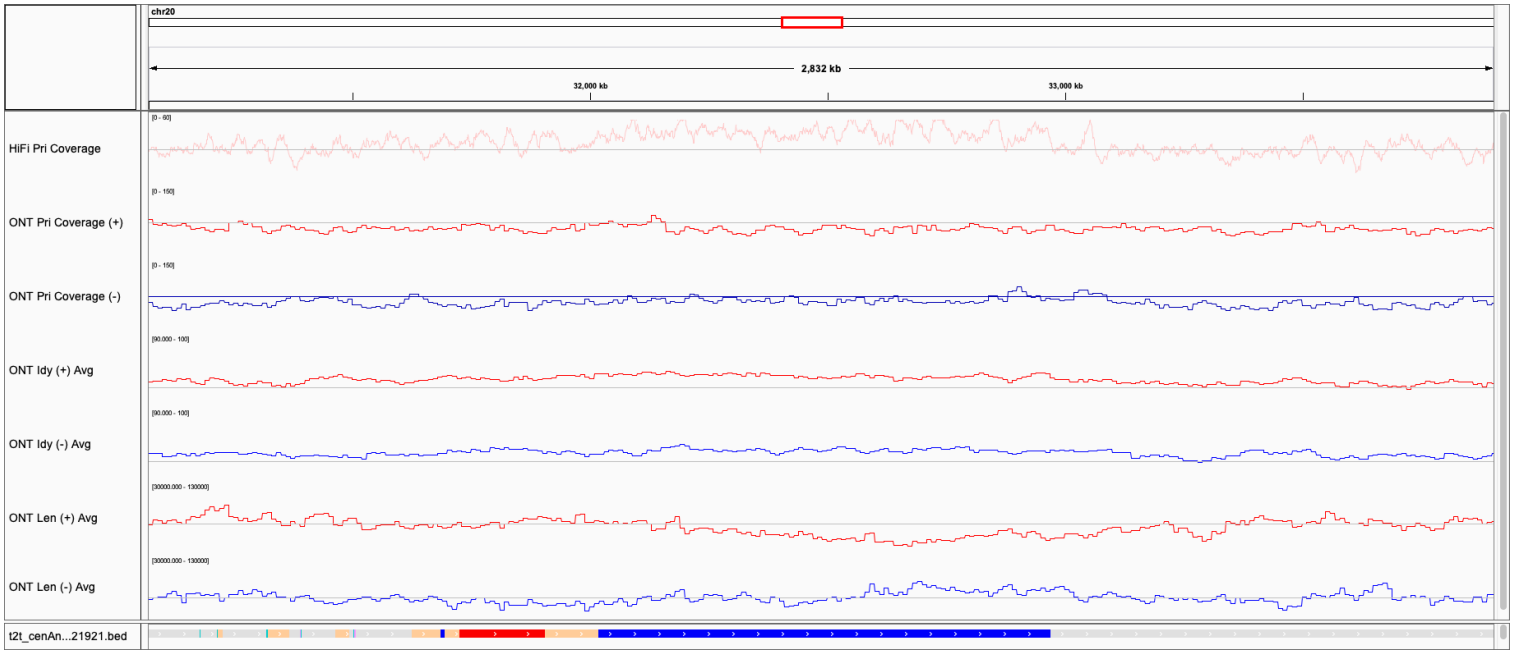


Fig. S25: HiFi and ONT read statistics within an HSat3 region on Chr9 and Chr20. (A) read statistics for the HSat3 region on Chr9 (same as in Fig. S22C, satellite array marked in dark blue in the annotation track). Overall, HSat3 tracks exhibit similar behavior to HSat2 in HiFi data but are closer to even coverage in ONT data and have less strand bias than HSat2. (B) Same statistics within the HSat3 region on Chr20 (same as in Fig. S22F, satellite array marked in dark blue in the annotation track).

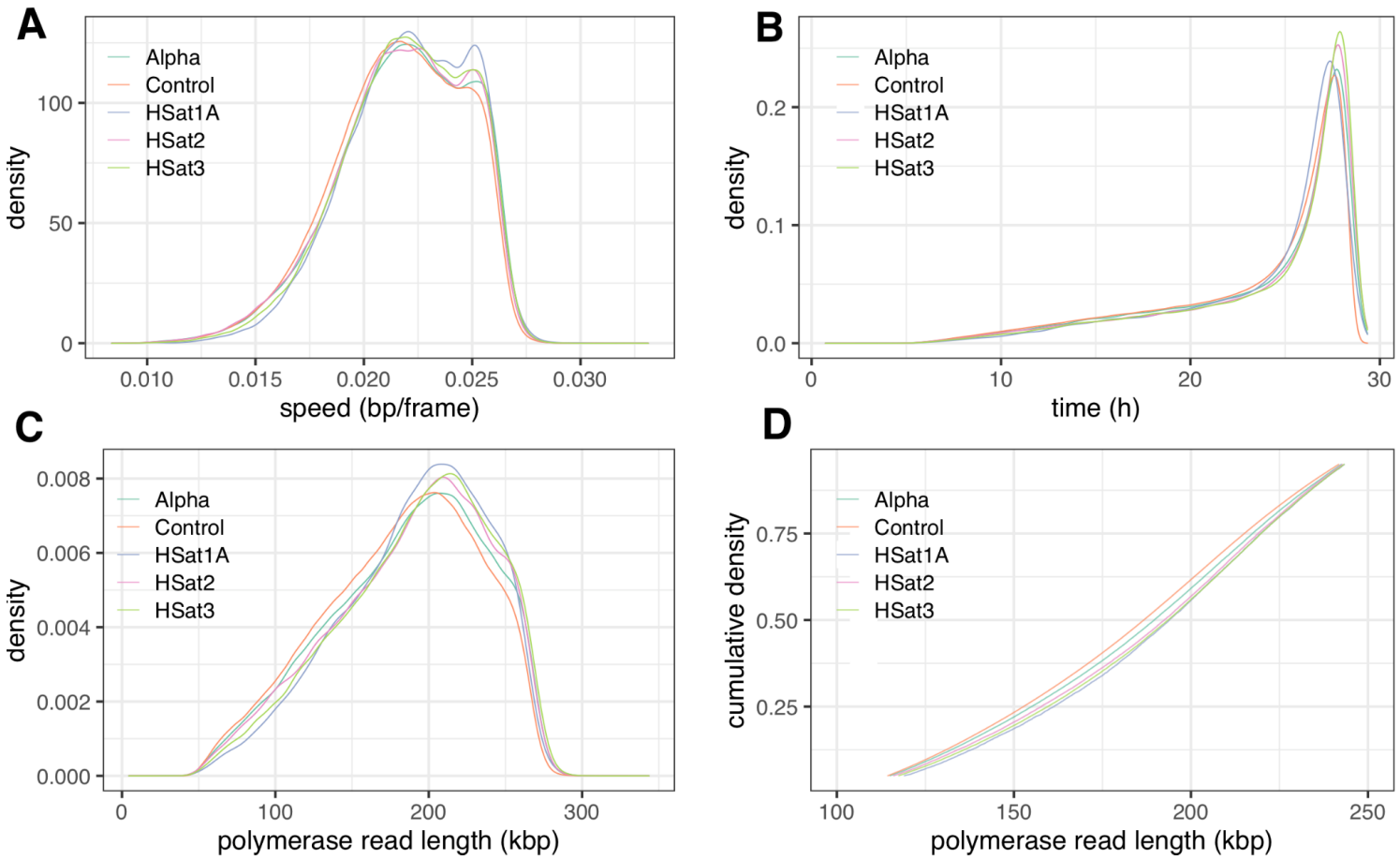


Fig. S26: Kinetic information for reads from HSat and AlphaSat regions as well as all reads (Control). The plots from top to bottom show (A) speed of sequencing, (B) total time sequencing a molecule, and polymerase read length ((C) regular and (D) cumulative). In all plots AlphaSat behaves similarly to the 'Control'. (A) No drastic difference in sequencing speeds was observed between the considered classes (AlphaSat in teal, HSat1A in blue, HSat2 in pink, and HSat3 in green), with all showing a bi-modal distribution. However, satellite reads are elevated in the second (faster) peak relative to control, with HSat1A having the largest increase. (B) The reading time for a polymerase read is longer for HSat2,3 and shorter for HSat1A, consistent with observations about polymerase read lengths. (C/D) The polymerase reads are of similar length, with a slight shift to longer reads in HSat2,3. HSat1A reads have higher density of reads in the 150-250 kbp range but then drop off and have lower fraction of polymerase reads longer than 250 kbp than AlphaSat or HSat2,3. Satellite reads are shifted to longer lengths when compared to the control.

Fig. S27: TandemTools validation. The x-axis is chromosome position (from Table S5) and y-axis is percent deviated reads (0-100%). Only 7 sites have >50% deviated reads.

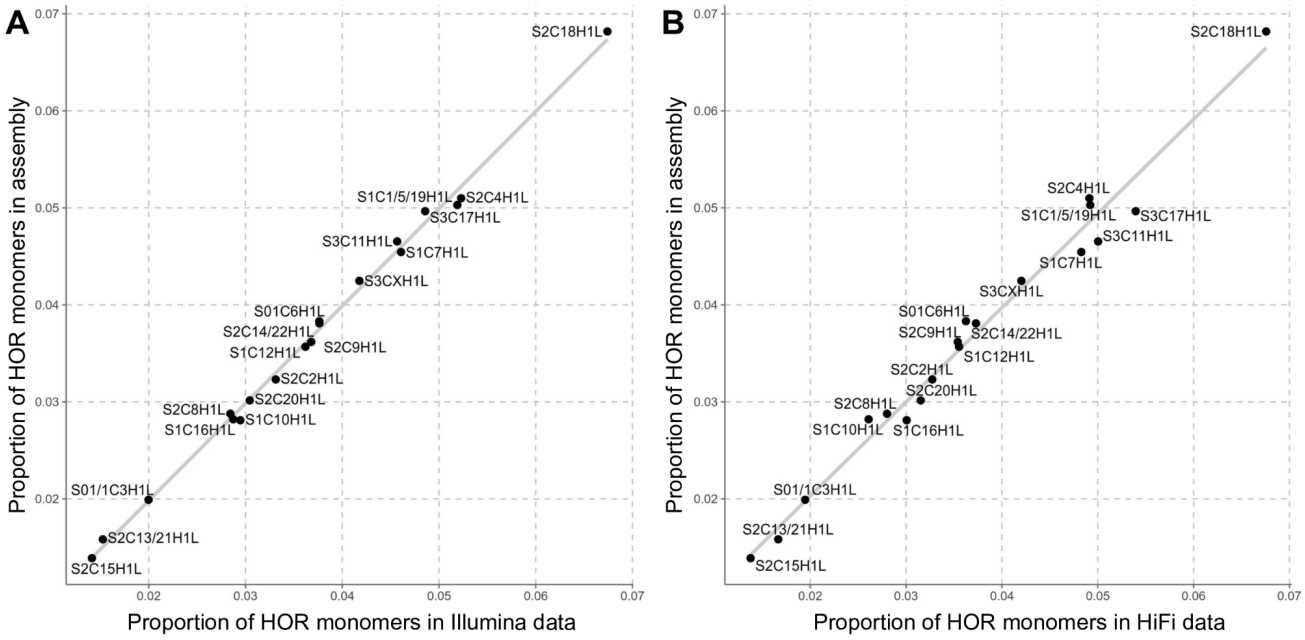


Fig. S28: Proportions of HOR monomers (160bp or longer) in the CHM13v1.0 assembly vs PCR-free Illumina data (A) and HiFi data (B). The line shown in the panels was calculated using a linear regression model in R (lm function) and geom_smooth function from the ggplot package. Note that the wider distribution of dots around the line in the HiFi sample may be due to more frequent sequencing errors in the HiFi reads, as compared to the Illumina reads.

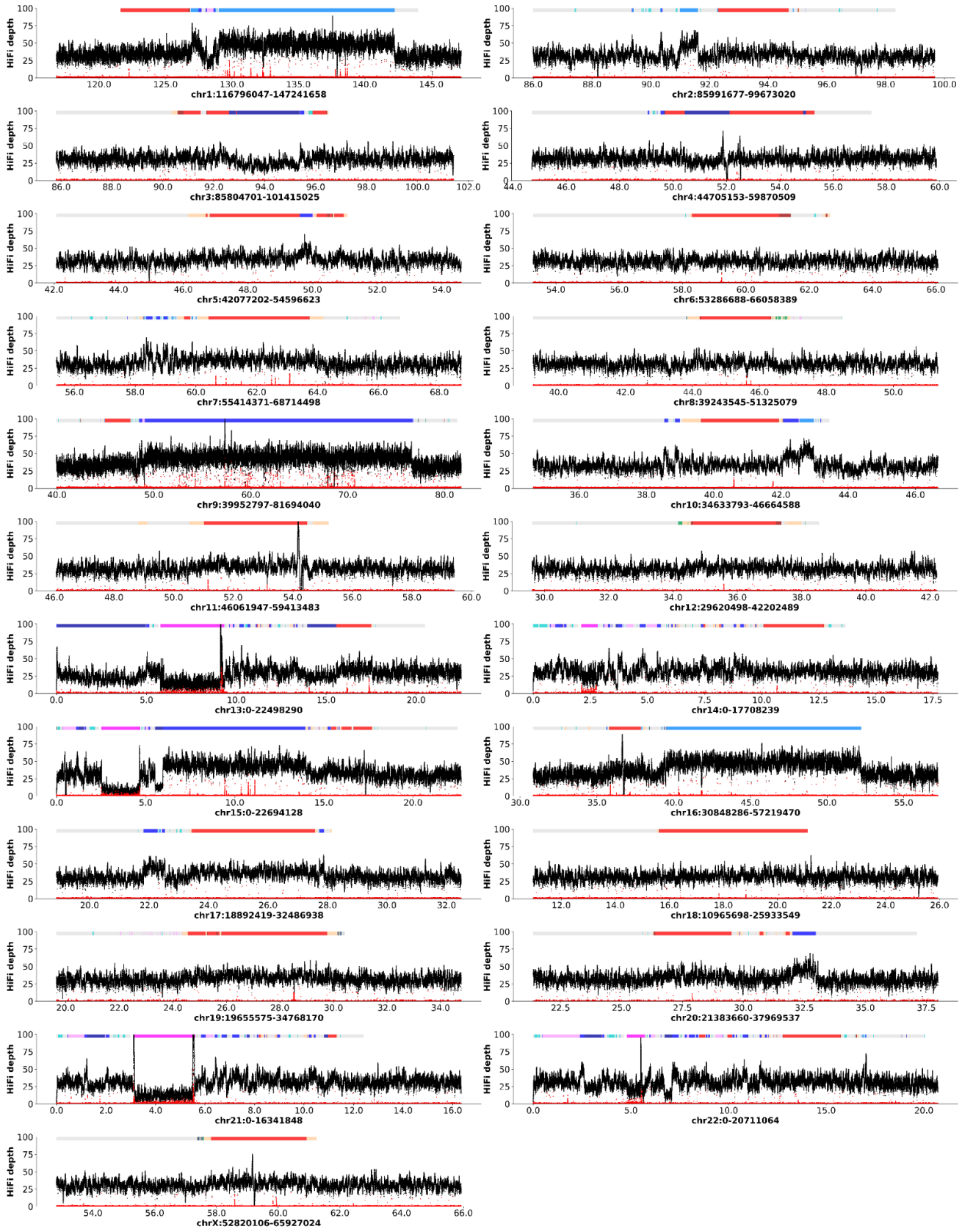


Fig. S29: NucFreq plot of centromeric and satellite regions in CHM13v1.1. Chromosomes 1-22 and chromosome X. HiFi coverage depth (black) along with secondary allele frequency (red) for all centromeres and surrounding regions. Top bar shows annotations (*30*) with AlphaSat (centromeric HORs) in red, HSat1 and HSat3 arrays in dark blue, HSat2 arrays in light blue, and rDNA in pink.

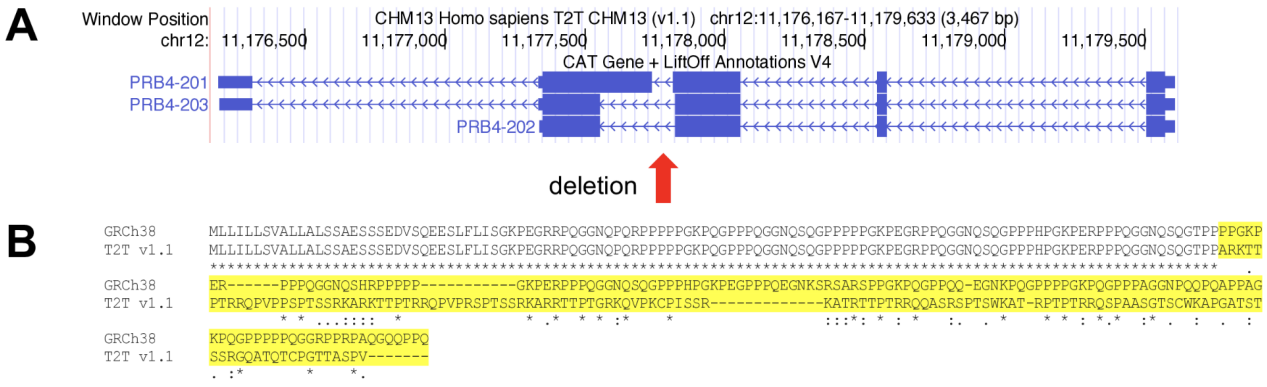


Fig. S30: Frameshift in PRB4. (A) Browser screenshot of the region. (B) Alignment of the predicted protein sequences.

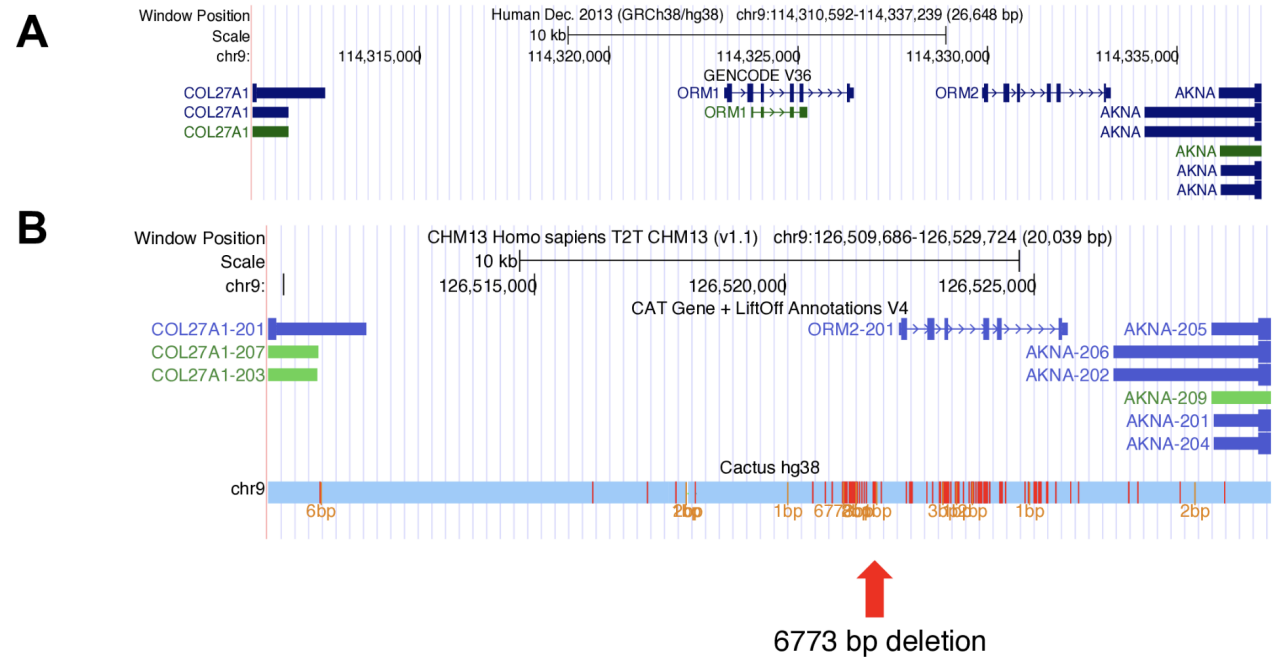


Fig. S31: ORM1 gene deletion. (A) A browser screenshot of GRCh38 showing a region on Chr9, with annotations for *ORM1* and *ORM2* genes. Upstream and downstream genes (*COL27A1* and *AKNA*) are included for context. (B) Browser screenshot of the corresponding region on CHM13 v1.1 assembly. *ORM2* is annotated; ORM1 falls within the 6,773 bp deletion, shown in the alignment track, and is thus missing.

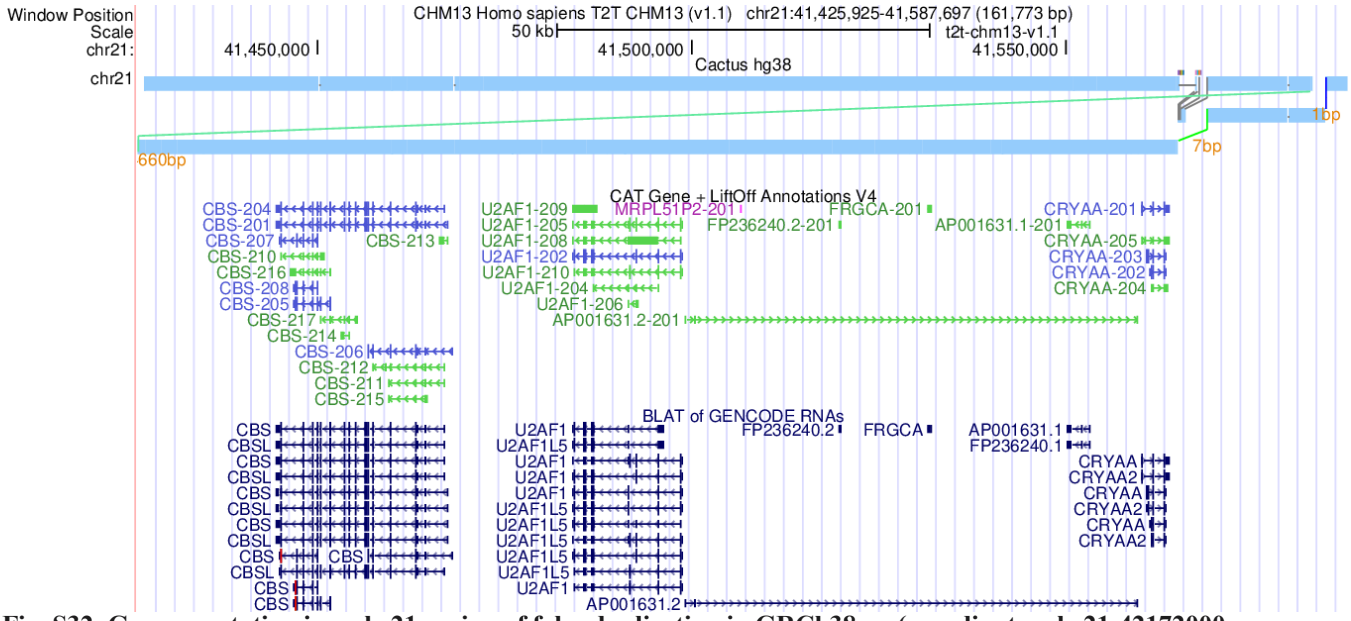


Fig. S32: **Gene annotation in a chr21 region of false duplication in GRCh38 on (coordinates chr21:42172000-42320000)**. The Cactus GRCh38 tracks show the alignment of the duplicated regions in GRCh38. The GENCODE BLAT track shows the multiple copies of the genes that all map to the same region in CHM13. The CAT+Liftoff annotation track shows how the annotation is resolved so that only one copy of each gene is included in the final annotation.

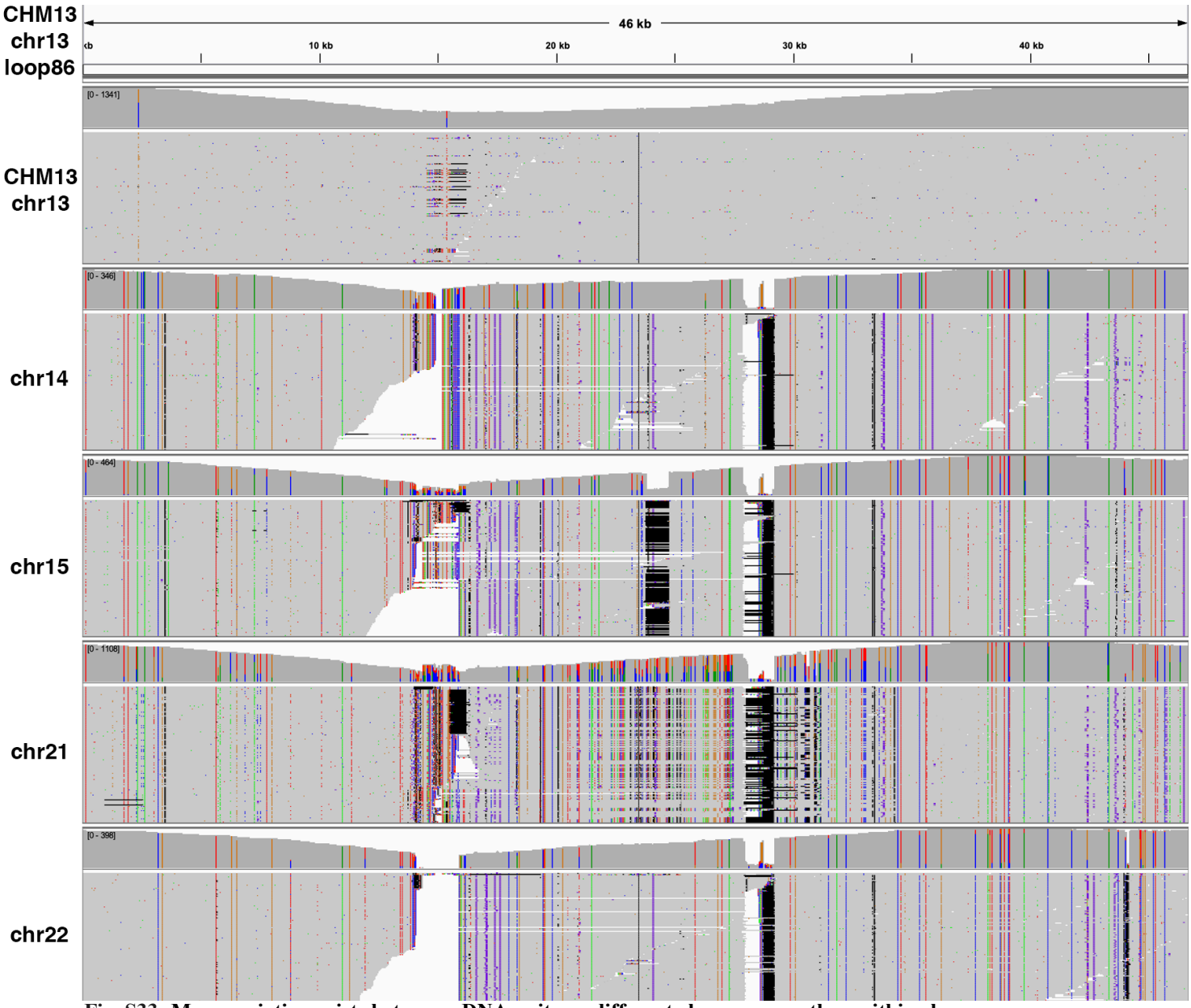


Fig. S33: More variation exists between rDNA units on different chromosomes than within chromosomes. CHM13 HiFi reads were binned by chromosome using the method described above in the section "Resolution of the rDNA arrays". Chromosome-specific HiFi reads were then aligned to the major morph on chromosome 13 ("loop86") using minimap2 with the `-ax map-pb` option and visualized in IGV. The alignments of chr13 reads versus the chr13 major morph are much cleaner than the alignments of reads from other chromosomes to chr13, demonstrating that the individual rDNA units on chr13 are more similar to one another than to those on other chromosomes. The major morphs on chromosomes 14, 15, 21, 22 show a similar pattern when this experiment is repeated.

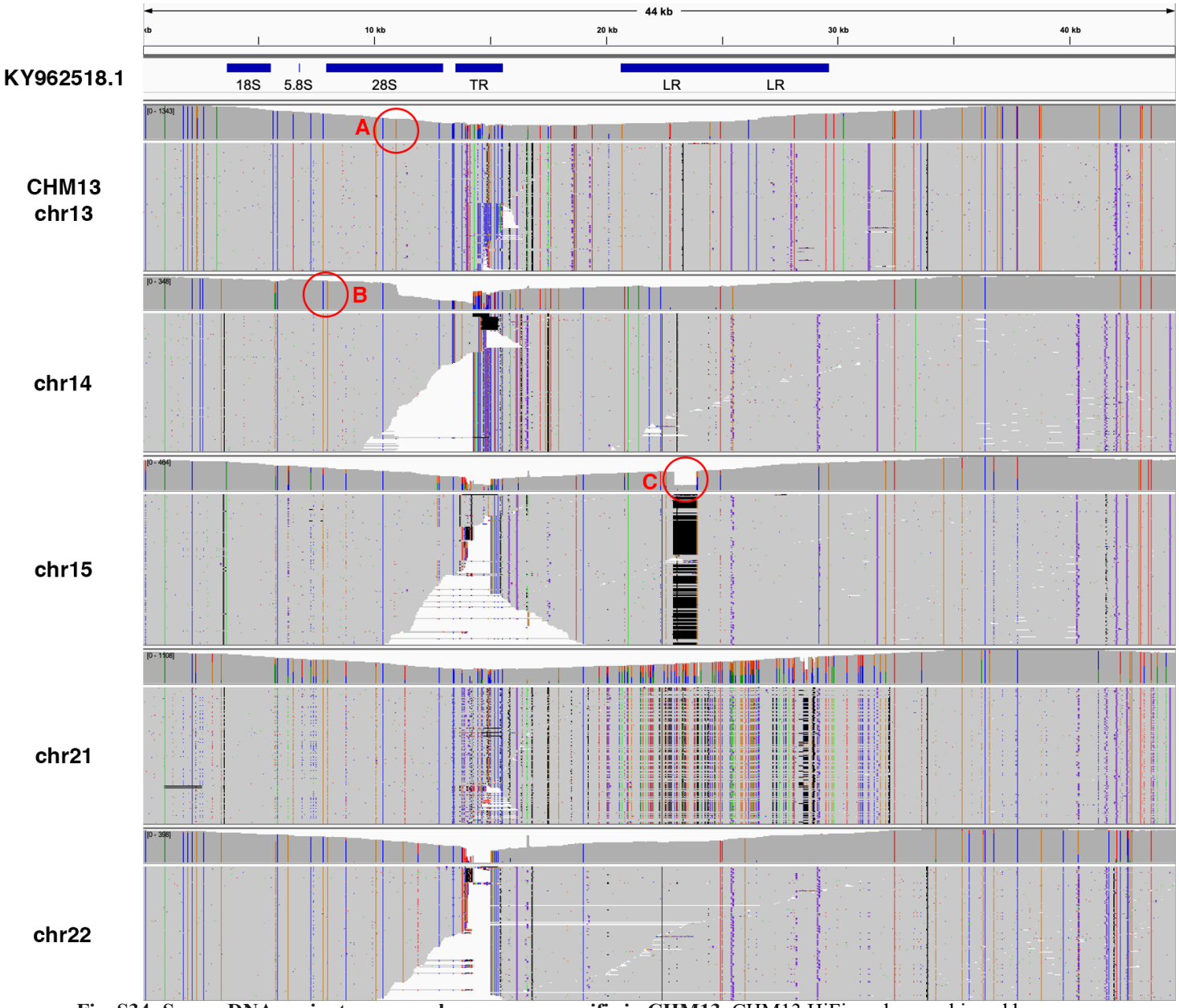


Fig. S34: Some rDNA variants appear chromosome specific in CHM13. CHM13 HiFi reads were binned by chromosome using the method described above in the section "Resolution of the rDNA arrays". Chromosome-specific HiFi reads were then aligned to a reference rDNA unit from GenBank (KY962518.1) using minimap2 with the `-ax map-pb` option and visualized in IGV. Key features of the rDNA unit, including the 18S, 5.8S, and 28S genes and the "TR" and "LR" repeats are annotated at the top. **(A)** An example SNP within the 28S is unique to CHM13 chr13 (brown variant). **(B)** Another example SNP is shared by chr14 and chr22 in CHM13, but absent from the reference and other CHM13 chromosomes. **(C)** A deletion within the LR is unique to the major morph on chr15 in CHM13, but the reference variant is also present at lower frequency in a minor morph.

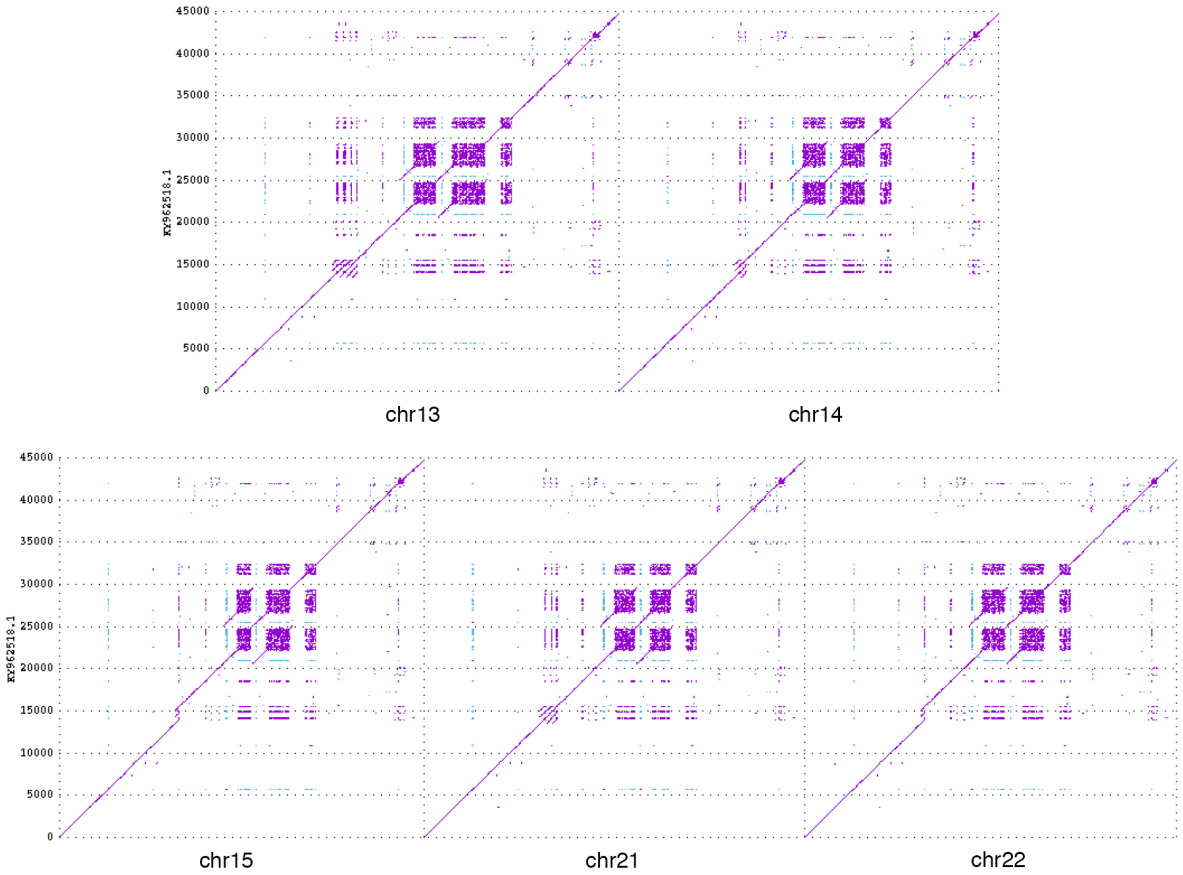


Fig. S35: The major rDNA morphs on each chromosome are structurally unique. MUMmer doplots are shown for exact matches of length 20 (forward matches in purple, reverse complement in blue). Chromosome morphs are on the x-axis and the KY962518.1 reference is on the y-axis. The "TR" tandem repeat at position 15 kbp is frequently variable between morphs, with a copy number of 4, 2, 1, 3, and 1 for the major morph on chromosomes 13, 14, 15, 21, and 22, respectively (the reference has 3 copies). The "LR" long repeat between positions 20 kbp and 30 kbp is another frequent source of variation between morphs, with the second copy expanded in chromosome 13 and the first copy collapsed in chromosome 15, relative to the reference. Most morphs can be distinguished based on the size of their TR and LR repeats.

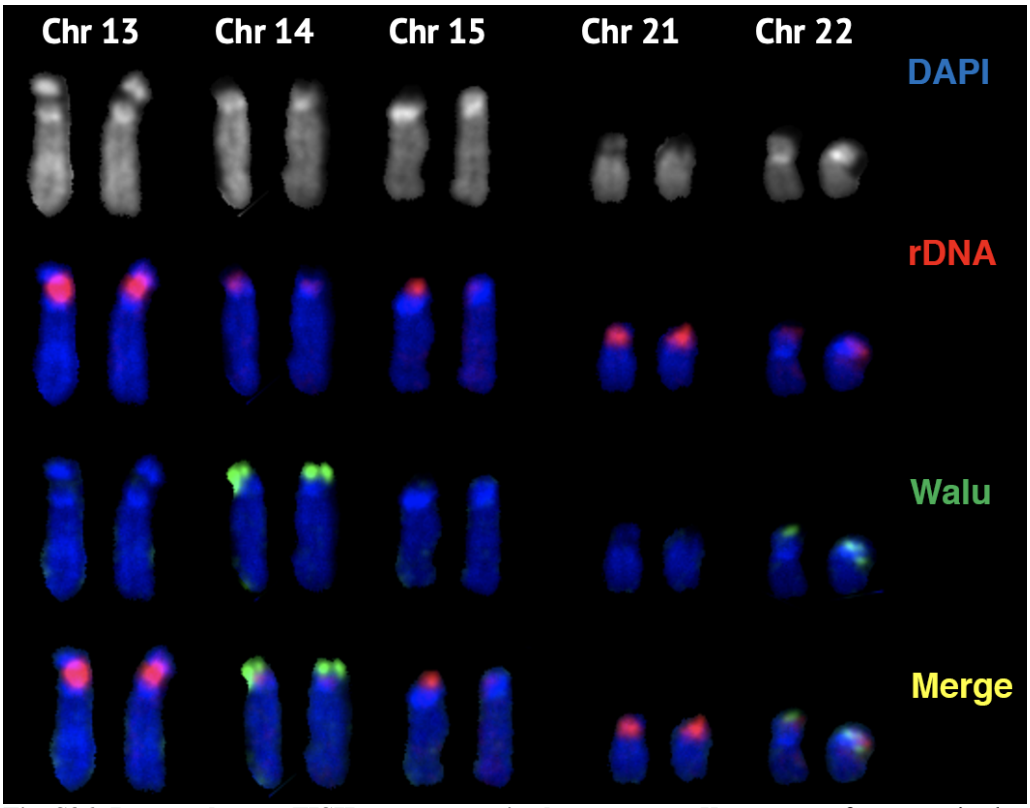


Fig. S36: Repeat element FISH on acrocentric chromosomes. Karyogram of acrocentric chromosomes from a CHM13 chromosome spread labeled by FISH with rDNA probe (orange) and the WaluSat probe (green). All acrocentric chromosomes were identified by morphology. WaluSat signal was detected on chromosomes 14 and 22, as predicted by the assembly.

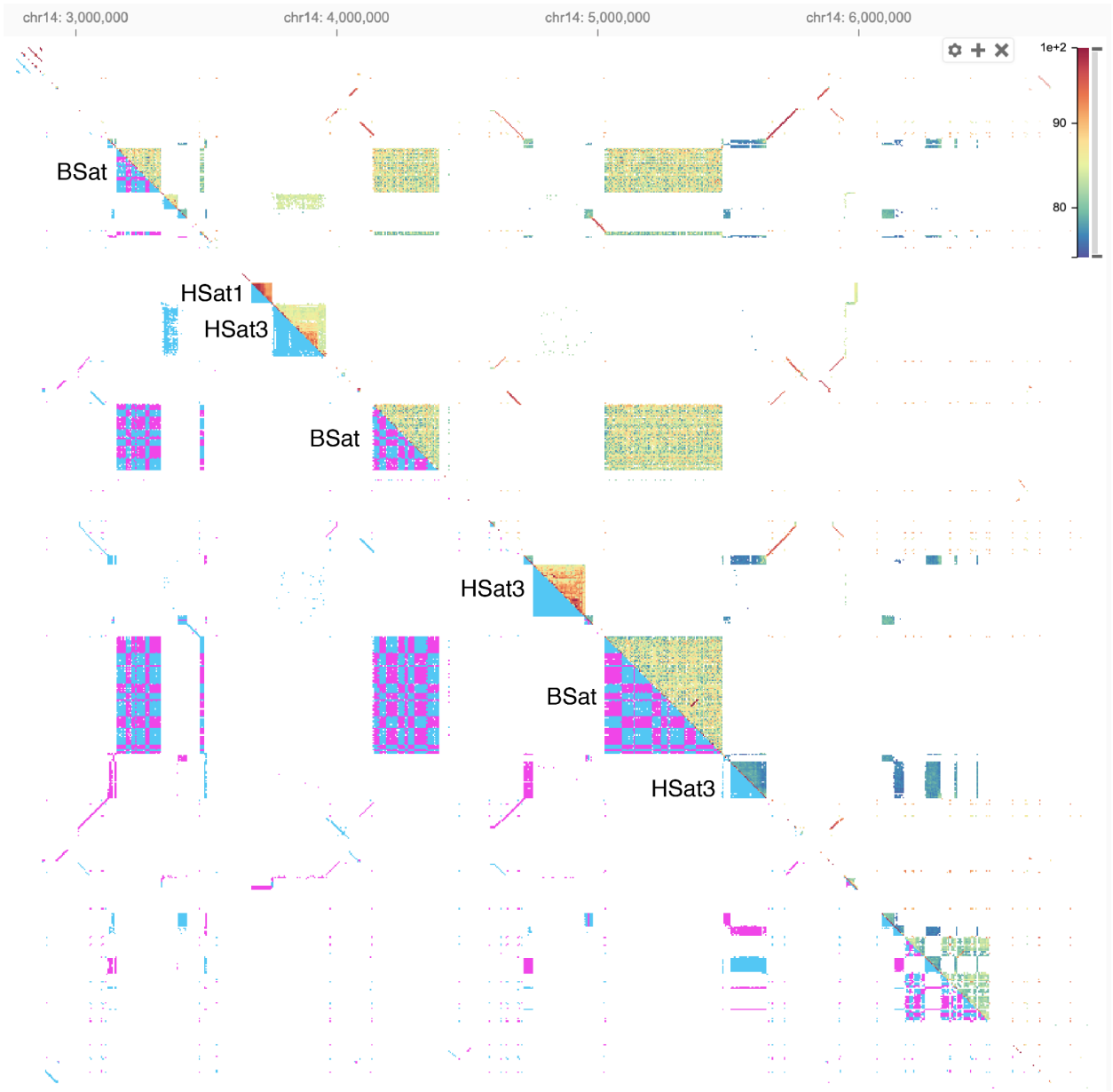


Fig. S37: Beta satellite arrays on the acrocentric proximal short arms show a mosaic inversion structure. A self-alignment dotplot of CHM13 Chromosome 14 (position 3–7 Mbp) is shown with the upper triangle displaying alignment similarity (as noted by the color scale) and the lower triangle displaying alignment orientation (forward in purple, reverse complement in blue). While the HSat1 and HSat3 tandem units only align in a forward orientation, the BSat arrays show a checkerboard pattern that is indicative of large inverted blocks of satellite sequence. Inversions are also evident within the alpha satellite monomer repeats at the bottom right of the plot. A similar pattern appears on the other acrocentric chromosomes, but only within the proximal satellite arrays and not the distal arrays.

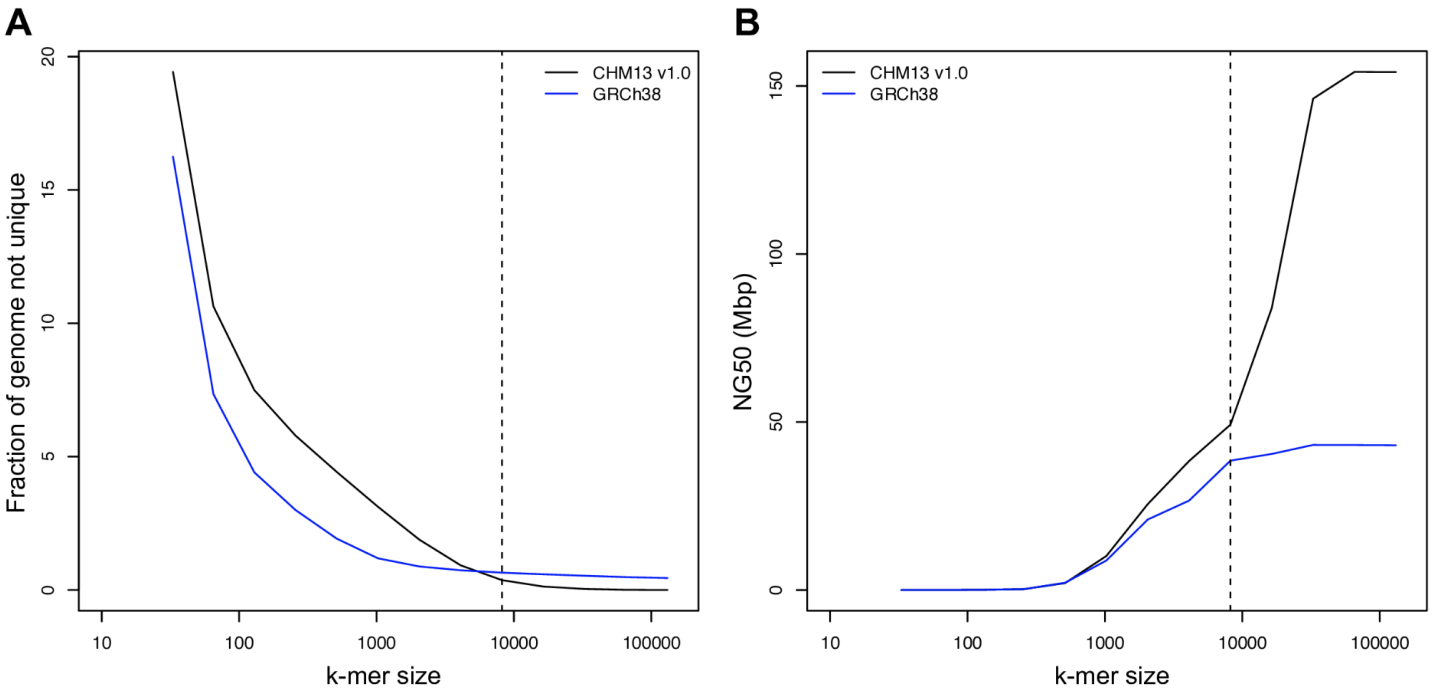


Fig. S38: *k*-mer analysis of CHM13v1.0 and GRCh38. (A) Fraction of genome (in percent) spanned by multi-copy *k*-mers (y axis) vs. *k*-mer size (x axis). CHM13v1.0 has higher repeat content at lower *k*-mer sizes, but as *k* exceeds 8 kbp (vertical line) GRCh38 overtakes CHM13v1.0 in repetitive sequence fraction. This indicates that individual instances of large genomic repeats might not have been accurately resolved within GRCh38, but were rather copy/pasted or compiled from model sequence, effectively homogenizing repeat units. (B) NG50 of genomic fragments after breaking at the start/end of every stretch of non-unique *k*-mers (y axis) vs. *k*-mer size (x axis). Non-unique stretches formed their own contigs. Note that due to the co-localization of the repeats within the genome, CHM13v1.0 is more "assemblable" than GRCh38 at all *k*-mer sizes over 1 kbp despite a larger fraction of it being repetitive. Substantial difference between CHM13v1.0 and GRCh38 statistics at higher values of *k* (over 8 kbp) indicates that a haploid human genome is more easily assembled by highly accurate sequencing reads (of 10 kbp and longer) than one would presume from GRCh38 reference analysis. This is consistent with the observation that projections relying on GRCh38 substantially under-estimated continuity of recent HiFi-based assemblies (*16*, *128*).

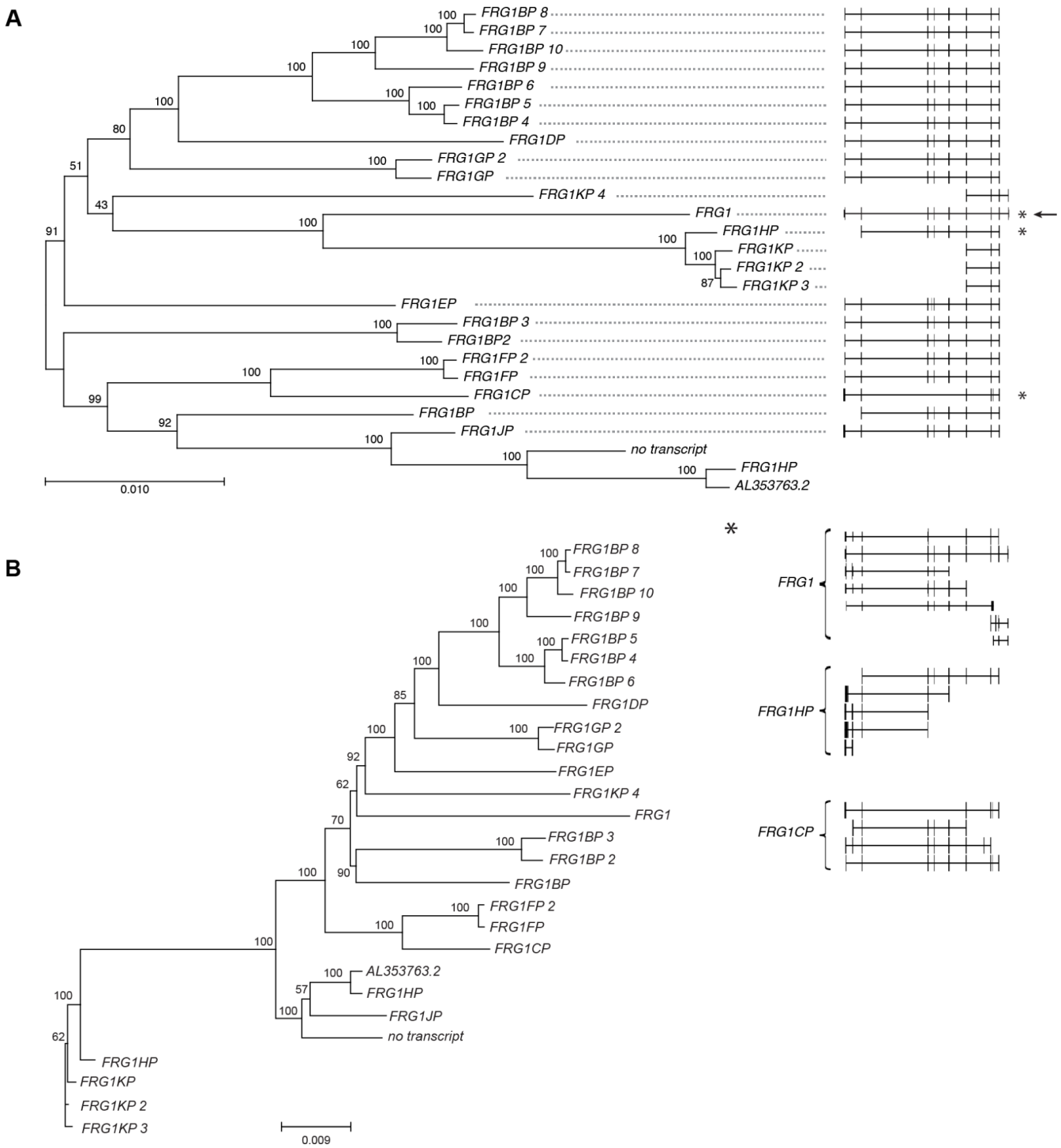


Fig. S39. Phylogenetic tree of the *FRG1* paralogs. (A) Tree inferred from the Neighbor-joining method. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. Units are the number of base substitutions per site. Numbers shown next to the branches are the percentage of replicate trees clustered together in the bootstrap test. Genes annotated with multiple transcripts are marked with *. (B) Tree inferred from the maximum likelihood comparison method with 100 bootstrap replicates. Both trees share similar insights on the branches close to *FRG1DP*.

References and Notes

1. V. A. Schneider, T. Graves-Lindsay, K. Howe, N. Bouk, H.-C. Chen, P. A. Kitts, T. D. Murphy, K. D. Pruitt, F. Thibaud-Nissen, D. Albracht, R. S. Fulton, M. Kremitzki, V. Magrini, C. Markovic, S. McGrath, K. M. Steinberg, K. Auger, W. Chow, J. Collins, G. Harden, T. Hubbard, S. Pelan, J. T. Simpson, G. Threadgold, J. Torrance, J. M. Wood, L. Clarke, S. Koren, M. Boitano, P. Peluso, H. Li, C.-S. Chin, A. M. Phillippy, R. Durbin, R. K. Wilson, P. Flicek, E. E. Eichler, D. M. Church, Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017). doi:10.1101/gr.213611.116 Medline

2. International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001). doi:10.1038/35057062 Medline

3. J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R.-R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.-H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.-H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N.

Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, X. Zhu, The sequence of the human genome. *Science* **291**, 1304–1351 (2001). doi:10.1126/science.1058040 Medline

4. E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. J. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H.-H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, J. C. Venter, A whole-genome assembly of *Drosophila*. *Science* **287**, 2196–2204 (2000). doi:10.1126/science.287.5461.2196 Medline

5. E. E. Eichler, R. A. Clark, X. She, An assessment of the sequence gaps: Unfinished business in a finished human genome. *Nat. Rev. Genet.* **5**, 345–354 (2004). doi:10.1038/nrg1322 Medline

6. K. H. Miga, Y. Newton, M. Jain, N. Altemose, H. F. Willard, W. J. Kent, Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res.* **24**, 697–707 (2014). doi:10.1101/gr.159624.113 Medline

7. M. Gupta, A. R. Dhanasekaran, K. J. Gardiner, Mouse models of Down syndrome: Gene content and consequences. *Mamm. Genome* **27**, 538–555 (2016). doi:10.1007/s00335-016-9661-8 Medline

8. M. J. P. Chaisson, J. Huddleston, M. Y. Dennis, P. H. Sudmant, M. Malig, F. Hormozdiari, F. Antonacci, U. Surti, R. Sandstrom, M. Boitano, J. M. Landolin, J. A. Stamatoyannopoulos, M. W. Hunkapiller, J. Korlach, E. E. Eichler, Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015). doi:10.1038/nature13907 Medline

9. International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004). doi:10.1038/nature03001 Medline

10. J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, S. Turner, Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009). doi:10.1126/science.1162986 Medline

11. K. Berlin, S. Koren, C.-S. Chin, J. P. Drake, J. M. Landolin, A. M. Phillippy, Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33**, 623–630 (2015). doi:10.1038/nbt.3238 Medline

12. M. Jain, S. Koren, K. H. Miga, J. Quick, A. C. Rand, T. A. Sasani, J. R. Tyson, A. D. Beggs, A. T. Dilthey, I. T. Fiddes, S. Malla, H. Marriott, T. Nieto, J. O'Grady, H. E. Olsen, B. S. Pedersen, A. Rhie, H. Richardson, A. R. Quinlan, T. P. Snutch, L. Tee, B. Paten, A. M. Phillippy, J. T. Simpson, N. J. Loman, M. Loose, Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).

doi:10.1038/nbt.4060 Medline

13. M. Jain, H. E. Olsen, D. J. Turner, D. Stoddart, K. V. Bulazel, B. Paten, D. Haussler, H. F. Willard, M. Akeson, K. H. Miga, Linear assembly of a human centromere on the Y chromosome. *Nat. Biotechnol.* **36**, 321–323 (2018). doi:10.1038/nbt.4109 Medline

14. K. H. Miga, S. Koren, A. Rhie, M. R. Vollger, A. Gershman, A. Bzikadze, S. Brooks, E. Howe, D. Porubsky, G. A. Logsdon, V. A. Schneider, T. Potapova, J. Wood, W. Chow, J. Armstrong, J. Fredrickson, E. Pak, K. Tigyi, M. Kremitzki, C. Markovic, V. Maduro, A. Dutra, G. G. Bouffard, A. M. Chang, N. F. Hansen, A. B. Wilfert, F. Thibaud-Nissen, A. D. Schmitt, J.-M. Belton, S. Selvaraj, M. Y. Dennis, D. C. Soto, R. Sahasrabudhe, G. Kaya, J. Quick, N. J. Loman, N. Holmes, M. Loose, U. Surti, R. A. Risques, T. A. Graves Lindsay, R. Fulton, I. Hall, B. Paten, K. Howe, W. Timp, A. Young, J. C. Mullikin, P. A. Pevzner, J. L. Gerton, B. A. Sullivan, E. E. Eichler, A. M. Phillippy, Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020). doi:10.1038/s41586-020-2547-7 Medline

15. A. M. Wenger, P. Peluso, W. J. Rowell, P.-C. Chang, R. J. Hall, G. T. Concepcion, J. Ebler, A. Fungtammasan, A. Kolesnikov, N. D. Olson, A. Töpfer, M. Alonge, M. Mahmoud, Y. Qian, C.-S. Chin, A. M. Phillippy, M. C. Schatz, G. Myers, M. A. DePristo, J. Ruan, T. Marschall, F. J. Sedlazeck, J. M. Zook, H. Li, S. Koren, A. Carroll, D. R. Rank, M. W. Hunkapiller, Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019). doi:10.1038/s41587-019-0217-9 Medline

16. S. Nurk, B. P. Walenz, A. Rhie, M. R. Vollger, G. A. Logsdon, R. Grothe, K. H. Miga, E. E. Eichler, A. M. Phillippy, S. Koren, HiCanu: Accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020). doi:10.1101/gr.263566.120 Medline

17. J. Huddleston, M. J. P. Chaisson, K. M. Steinberg, W. Warren, K. Hoekzema, D. Gordon, T. A. Graves-Lindsay, K. M. Munson, Z. N. Kronenberg, L. Vives, P. Peluso, M. Boitano, C.-S. Chin, J. Korlach, R. K. Wilson, E. E. Eichler, Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–685 (2017). doi:10.1101/gr.214007.116 Medline

18. E. E. Eichler, U. Surti, R. Ophoff, Proposal for construction a human haploid BAC library from hydatidiform mole source material (2002); www.genome.gov/Pages/Research/Sequencing/BACLibrary/HydatidiformMoleBAC021203.pdf.

19. K. M. Steinberg, V. A. Schneider, T. A. Graves-Lindsay, R. S. Fulton, R. Agarwala, J. Huddleston, S. A. Shiryev, A. Morgulis, U. Surti, W. C. Warren, D. M. Church, E. E. Eichler, R. K. Wilson, Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Res.* **24**, 2066–2076 (2014). doi:10.1101/gr.180893.114 Medline

20. M. R. Vollger, G. A. Logsdon, P. A. Audano, A. Sulovari, D. Porubsky, P. Peluso, A. M. Wenger, G. T. Concepcion, Z. N. Kronenberg, K. M. Munson, C. Baker, A. D. Sanders, D. C. J. Spierings, P. M. Lansdorp, U. Surti, M. W. Hunkapiller, E. E. Eichler, Improved assembly and variant detection of a haploid human genome using single-molecule, high-

fidelity long reads. *Ann. Hum. Genet.* **84**, 125–140 (2020). doi:10.1111/ahg.12364 Medline

21. G. A. Logsdon, M. R. Vollger, P. Hsieh, Y. Mao, M. A. Liskovykh, S. Koren, S. Nurk, L. Mercuri, P. C. Dishuck, A. Rhie, L. G. de Lima, T. Dvorkina, D. Porubsky, W. T. Harvey, A. Mikheenko, A. V. Bzikadze, M. Kremitzki, T. A. Graves-Lindsay, C. Jain, K. Hoekzema, S. C. Murali, K. M. Munson, C. Baker, M. Sorensen, A. M. Lewis, U. Surti, J. L. Gerton, V. Larionov, M. Ventura, K. H. Miga, A. M. Phillippy, E. E. Eichler, The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021). doi:10.1038/s41586-021-03420-7 Medline

22. J.-B. Fan, U. Surti, P. Taillon-Miller, L. Hsie, G. C. Kennedy, L. Hoffner, T. Ryder, D. G. Mutch, P.-Y. Kwok, Paternal origins of complete hydatidiform moles proven by whole genome single-nucleotide polymorphism haplotyping. *Genomics* **79**, 58–62 (2002). doi:10.1006/geno.2001.6676 Medline

23. See supplementary materials.

24. 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). doi:10.1038/nature15393 Medline

25. S. Aganezov, S. M. Yan, D. C. Soto, M. Kirsche, S. Zarate, P. Avdeyev, D. J. Taylor, K. Shafin, A. Shumate, C. Xiao, J. Wagner, J. McDaniel, N. D. Olson, M. E. G. Sauria, M. R. Vollger, A. Rhie, M. Meredith, S. Martin, J. Lee, S. Koren, J. A. Rosenfeld, B. Paten, R. Layer, C.-S. Chin, F. J. Sedlazeck, N. F. Hansen, D. E. Miller, A. M. Phillippy, K. H. Miga, R. C. McCoy, M. Y. Dennis, J. M. Zook, M. C. Schatz, A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022). doi:10.1126/science.abl3533

26. E. W. Myers, The fragment assembly string graph. *Bioinformatics* **21**, ii79–ii85 (2005). doi:10.1093/bioinformatics/bti1114 Medline

27. H. Li, Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016). doi:10.1093/bioinformatics/btw152 Medline

28. M. Rautiainen, T. Marschall, GraphAligner: Rapid and versatile sequence-to-graph alignment. *Genome Biol.* **21**, 253 (2020). doi:10.1186/s13059-020-02157-2 Medline

29. A. Gershman, M. E. G. Sauria, X. Guitart, M. R. Vollger, P. W. Hook, S. J. Hoyt, M. Jain, A. Shumate, R. Razaghi, S. Koren, N. Altemose, G. V. Caldas, G. A. Logsdon, A. Rhie, E. E. Eichler, M. C. Schatz, R. J. O'Neill, A. M. Phillippy, K. H. Miga, W. Timp, Epigenetic patterns in a complete human genome. *Science* **376**, eabj5089 (2022). doi:10.1126/science.abj5089

30. N. Altemose, G. A. Logsdon, A. V. Bzikadze, P. Sidhwani, S. A. Langley, G. V. Caldas, S. J. Hoyt, L. Uralsky, F. D. Ryabov, C. J. Shew, M. E. G. Sauria, M. Borchers, A. Gershman, A. Mikheenko, V. A. Shepelev, T. Dvorkina, O. Kunyavskaya, M. R. Vollger, A. Rhie, A. M. McCartney, M. Asri, R. Lorig-Roach, K. Shafin, J. K. Lucas, S. Aganezov, D. Olson, L. Gomes de Lima, T. Potapova, G. A. Hartley, M. Haukness, P. Kerpedjiev, F. Gusev, K. Tigyi, S. Brooks, A. Young, S. Nurk, S. Koren, S. R. Salama, B. Paten, E. I. Rogaev, A. Streets, G. H. Karpen, A. F. Dernburg, B. A. Sullivan, A. F. Straight, T. J. Wheeler, J. L. Gerton, E. E. Eichler, A. M. Phillippy, W. Timp, M. Y. Dennis, R. J.

O'Neill, J. M. Zook, M. C. Schatz, P. A. Pevzner, M. Diekhans, C. H. Langley, I. A. Alexandrov, K. H. Miga, Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022). doi:10.1126/science.abl4178

31. M. M. Parks, C. M. Kurylo, R. A. Dass, L. Bojmar, D. Lyden, C. T. Vincent, S. C. Blanchard, Variant ribosomal RNA alleles are conserved and exhibit tissue-specific expression. *Sci. Adv.* **4**, eaao0665 (2018). doi:10.1126/sciadv.aao0665 Medline

32. J. O. Nelson, G. J. Watase, N. Warsinger-Pepe, Y. M. Yamashita, Mechanisms of rDNA copy number maintenance. *Trends Genet.* **35**, 734–742 (2019). doi:10.1016/j.tig.2019.07.006 Medline

33. M. Rautiainen, T. Marschall, MBG: Minimizer-based sparse de Bruijn graph construction. *Bioinformatics* **37**, 2476–2478 (2021). doi:10.1093/bioinformatics/btab004 Medline

34. A. M. McCartney, K. Shafin, M. Alonge, A. V. Bzikadze, G. Formenti, A. Fungtammasan, K. Howe, C. Jain, S. Koren, G. A. Logsdon, K. H. Miga, A. Mikheenko, B. Paten, A. Shumate, D. C. Soto, I. Sović, J. M. D. Wood, J. M. Zook, A. M. Phillippy, A. Rhie, Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nat. Methods* 10.1038/s41592-022-01440-3 (2022). doi:10.1038/s41592-022-01440-3

35. J. M. Flynn, M. Long, R. A. Wing, A. G. Clark, Evolutionary dynamics of abundant 7-bp satellites in the genome of *Drosophila virilis*. *Mol. Biol. Evol.* **37**, 1362–1375 (2020). doi:10.1093/molbev/msaa010 Medline

36. W. M. Guiblet, M. A. Cremona, M. Cechova, R. S. Harris, I. Kejnovská, E. Kejnovsky, K. Eckert, F. Chiaromonte, K. D. Makova, Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate. *Genome Res.* **28**, 1767–1778 (2018). doi:10.1101/gr.241257.118 Medline

37. A. Mikheenko, A. V. Bzikadze, A. Gurevich, K. H. Miga, P. A. Pevzner, TandemTools: Mapping long reads and assessing/improving assembly quality in extra-long tandem repeats. *Bioinformatics* **36**, i75–i83 (2020). doi:10.1093/bioinformatics/btaa440 Medline

38. M. R. Vollger, P. C. Dishuck, M. Sorensen, A. E. Welch, V. Dang, M. L. Dougherty, T. A. Graves-Lindsay, R. K. Wilson, M. J. P. Chaisson, E. E. Eichler, Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019). doi:10.1038/s41592-018-0236-3 Medline

39. I. T. Fiddes, J. Armstrong, M. Diekhans, S. Nachtweide, Z. N. Kronenberg, J. G. Underwood, D. Gordon, D. Earl, T. Keane, E. E. Eichler, D. Haussler, M. Stanke, B. Paten, Comparative Annotation Toolkit (CAT)-simultaneous clade and personal genome annotation. *Genome Res.* **28**, 1029–1038 (2018). doi:10.1101/gr.233460.117 Medline

40. A. Shumate, S. L. Salzberg, Liftoff: Accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021). doi:10.1093/bioinformatics/btaa1016 Medline

41. A. Frankish, M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, C. Sisu, J. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, S. Carbonell Sala, J. Chrast, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. García Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, T. Hunt, O. G. Izuogu, J. Lagarde, F. J. Martin, L. Martínez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo,

M. Ruffier, B. M. Schmitt, E. Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, J. Xu, A. Yates, D. Zerbino, Y. Zhang, B. Aken, J. S. Choudhary, M. Gerstein, R. Guigó, T. J. P. Hubbard, M. Kellis, B. Paten, A. Reymond, M. L. Tress, P. Flicek, GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019). doi:10.1093/nar/gky955 Medline

42. M. R. Vollger, X. Guitart, P. C. Dishuck, L. Mercuri, W. T. Harvey, A. Gershman, M. Diekhans, A. Sulovari, K. M. Munson, A. P. Lewis, K. Hoekzema, D. Porubsky, R. Li, S. Nurk, S. Koren, K. H. Miga, A. M. Phillippy, W. Timp, M. Ventura, E. E. Eichler, Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022). doi:10.1126/science.abj6965

43. S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J. P. Spence, Y. S. Song, G. Poletti, F. Balloux, G. van Driem, P. de Knijff, I. G. Romero, A. R. Jha, D. M. Behar, C. M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O. L. Posukh, E. Balanovska, O. Balanovsky, S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M. S. Abdullah, A. Ruiz-Linares, C. M. Beall, A. Di Rienzo, C. Jeong, E. B. Starikovskaya, E. Metspalu, J. Parik, R. Villems, B. M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J. T. S. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M. F. Hammer, T. Kivisild, W. Klitz, C. A. Winkler, D. Labuda, M. Bamshad, L. B. Jorde, S. A. Tishkoff, W. S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Pääbo, J. Kelso, N. Patterson, D. Reich, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016). doi:10.1038/nature18964 Medline

44. M. S. Lindström, D. Jurada, S. Bursac, I. Orsolic, J. Bartek, S. Volarevic, Nucleolus as an emerging hub in maintenance of genome stability and cancer pathogenesis. *Oncogene* **37**, 2351–2366 (2018). doi:10.1038/s41388-017-0121-z Medline

45. R. Lyle, P. Prandini, K. Osoegawa, B. ten Hallers, S. Humphray, B. Zhu, E. Eyras, R. Castelo, C. P. Bird, S. Gagos, C. Scott, A. Cox, S. Deutsch, C. Ucla, M. Cruts, S. Dahoun, X. She, F. Bena, S.-Y. Wang, C. Van Broeckhoven, E. E. Eichler, R. Guigo, J. Rogers, P. J. de Jong, A. Reymond, S. E. Antonarakis, Islands of euchromatin-like sequence and expressed polymorphic sequences within the short arm of human chromosome 21. *Genome Res.* **17**, 1690–1696 (2007). doi:10.1101/gr.6675307 Medline

46. I. Floutsakou, S. Agrawal, T. T. Nguyen, C. Seoighe, A. R. D. Ganley, B. McStay, The shared genomic architecture of human nucleolar organizer regions. *Genome Res.* **23**, 2003–2012 (2013). doi:10.1101/gr.157941.113 Medline

47. J.-H. Kim, A. T. Dilthey, R. Nagaraja, H.-S. Lee, S. Koren, D. Dudekula, W. H. Wood Iii, Y. Piao, A. Y. Ogurtsov, K. Utani, V. N. Noskov, S. A. Shabalina, D. Schlessinger, A. M. Phillippy, V. Larionov, Variation in human chromosome 21 ribosomal RNA genes characterized by TAR cloning and long-read sequencing. *Nucleic Acids Res.* **46**, 6712–6725 (2018). doi:10.1093/nar/gky442 Medline

48. S. Caburet, C. Conti, C. Schurra, R. Lebofsky, S. J. Edelstein, A. Bensimon, Human

ribosomal RNA gene arrays display a broad range of palindromic structures. *Genome Res.* **15**, 1079–1085 (2005). doi:10.1101/gr.3970105 Medline

49. S. J. Hoyt, J. M. Storer, G. A. Hartley, P. G. S. Grady, A. Gershman, L. G. de Lima, C. Limouse, R. Halabian, L. Wojenski, M. Rodriguez, N. Altemose, A. Rhie, L. J. Core, J. L. Gerton, W. Makalowski, D. Olson, J. Rosen, A. F. A. Smit, A. F. Straight, M. R. Vollger, T. J. Wheeler, M. C. Schatz, E. E. Eichler, A. M. Phillippy, W. Timp, K. H. Miga, R. J. O'Neill, From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science* **376**, eabk3112 (2022). doi:10.1126/science.abk3112

50. M. van Sluis, M. Ó. Gailín, J. G. W. McCarter, H. Mangan, A. Grob, B. McStay, Human NORs, comprising rDNA arrays and functionally conserved distal elements, are located within dynamic chromosomal regions. *Genes Dev.* **33**, 1688–1701 (2019). doi:10.1101/gad.331892.119 Medline

51. B. A. Sullivan, L. S. Jenkins, E. M. Karson, J. Leana-Cox, S. Schwartz, Evidence for structural heterogeneity from molecular cytogenetic analysis of dicentric Robertsonian translocations. *Am. J. Hum. Genet.* **59**, 167–175 (1996). Medline

52. G. M. Greig, P. E. Warburton, H. F. Willard, Organization and evolution of an alpha satellite DNA subset shared by human chromosomes 13 and 21. *J. Mol. Evol.* **37**, 464–475 (1993). doi:10.1007/BF00160427 Medline

53. A. L. Jørgensen, S. Kølvraa, C. Jones, A. L. Bak, A subfamily of alphoid repetitive DNA shared by the NOR-bearing human chromosomes 14 and 22. *Genomics* **3**, 100–109 (1988). doi:10.1016/0888-7543(88)90139-5 Medline

54. W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, D. Haussler, The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002). doi:10.1101/gr.229102 Medline

55. C. Wijmenga, J. E. Hewitt, L. A. Sandkuijl, L. N. Clark, T. J. Wright, H. G. Dauwerse, A.-M. Gruter, M. H. Hofker, P. Moerer, R. Williamson, G.-J. B. van Ommen, G. W. Padberg, R. R. Frants, Chromosome 4q DNA rearrangements associated with facioscapulohumeral muscular dystrophy. *Nat. Genet.* **2**, 26–30 (1992). doi:10.1038/ng0992-26 Medline

56. R. J. L. F. Lemmers, P. J. van der Vliet, R. Klooster, S. Sacconi, P. Camaño, J. G. Dauwerse, L. Snider, K. R. Straasheijm, G. J. van Ommen, G. W. Padberg, D. G. Miller, S. J. Tapscott, R. Tawil, R. R. Frants, S. M. van der Maarel, A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science* **329**, 1650–1653 (2010). doi:10.1126/science.1189044 Medline

57. P. K. Grewal, M. van Geel, R. R. Frants, P. de Jong, J. E. Hewitt, Recent amplification of the human *FRG1* gene during primate evolution. *Gene* **227**, 79–88 (1999). doi:10.1016/S0378-1119(98)00587-3 Medline

58. J. C. van Deutekom, R. J. Lemmers, P. K. Grewal, M. van Geel, S. Romberg, H. G. Dauwerse, T. J. Wright, G. W. Padberg, M. H. Hofker, J. E. Hewitt, R. R. Frants, Identification of the first gene (*FRG1*) from the FSHD region on human chromosome 4q35. *Hum. Mol. Genet.* **5**, 581–590 (1996). doi:10.1093/hmg/5.5.581 Medline

59. K. H. Miga, T. Wang, The need for a human pangenome reference sequence. *Annu. Rev. Genomics Hum. Genet.* **22**, 81–102 (2021). doi:10.1146/annurev-genom-120120-081921

Medline

60. R. R. Wick, M. B. Schultz, J. Zobel, K. E. Holt, Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015). doi:10.1093/bioinformatics/btv383 Medline

61. S. Nurk, S. Koren, A. Rhie, M. Rautiainen, T2T-CHM13 supplemental code and data. Zenodo (2021); https://doi.org/10.5281/zenodo.5598253.

62. B. K. Maples, S. Gravel, E. E. Kenny, C. D. Bustamante, RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013). doi:10.1016/j.ajhg.2013.06.020 Medline

63. E. Lowy-Gallego, S. Fairley, X. Zheng-Bradley, M. Ruffier, L. Clarke, P. Flicek, Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project. *Wellcome Open Res.* **4**, 50 (2019). doi:10.12688/wellcomeopenres.15126.2 Medline

64. International HapMap Consortium, A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007). doi:10.1038/nature06258 Medline

65. H. Li, J. M. Bloom, Y. Farjoun, M. Fleharty, L. Gauthier, B. Neale, D. MacArthur, A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat. Methods* **15**, 595–597 (2018). doi:10.1038/s41592-018-0054-7 Medline

66. L. Chen, A. B. Wolf, W. Fu, L. Li, J. M. Akey, Identifying and interpreting apparent Neanderthal ancestry in African individuals. *Cell* **180**, 677–687.e16 (2020). doi:10.1016/j.cell.2020.01.012 Medline

67. K. Prüfer, C. de Filippo, S. Grote, F. Mafessoni, P. Korlević, M. Hajdinjak, B. Vernot, L. Skov, P. Hsieh, S. Peyrégne, D. Reher, C. Hopfe, S. Nagel, T. Maricic, Q. Fu, C. Theunert, R. Rogers, P. Skoglund, M. Chintalapati, M. Dannemann, B. J. Nelson, F. M. Key, P. Rudan, Ž. Kućan, I. Gušić, L. V. Golovanova, V. B. Doronichev, N. Patterson, D. Reich, E. E. Eichler, M. Slatkin, M. H. Schierup, A. M. Andrés, J. Kelso, M. Meyer, S. Pääbo, A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655–658 (2017). doi:10.1126/science.aao1887 Medline

68. M. Byrska-Bishop, U. S. Evani, X. Zhao, A. O. Basile, H. J. Abel, A. A. Regier, A. Corvelo, W. E. Clarke, R. Musunuri, K. Nagulapalli, S. Fairley, A. Runnels, L. Winterkorn, E. Lowy-Gallego, P. Flicek, S. Germer, H. Brand, I. M. Hall, M. E. Talkowski, G. Narzisi, M. C. Zody, High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. bioRxiv 2021.02.06.430068 [Preprint] (2021); https://doi.org/10.1101/2021.02.06.430068.

69. G. A. Logsdon, M. R. Vollger, E. E. Eichler, Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020). doi:10.1038/s41576-020-0236-x Medline

70. J. R. Miller, A. L. Delcher, S. Koren, E. Venter, B. P. Walenz, A. Brownley, J. Johnson, K. Li, C. Mobarry, G. Sutton, Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**, 2818–2824 (2008). doi:10.1093/bioinformatics/btn548 Medline

71. S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, A. M. Phillippy, Canu:

Scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017). doi:10.1101/gr.215087.116 Medline

72. E. T. Dawson, R. Durbin, GFAKluge: A C++ library and command line utilities for the Graphical Fragment Assembly formats. *J. Open Source Softw.* **4**, 1083 (2019). doi:10.21105/joss.01083 Medline

73. G. Gonnella, S. Kurtz, GfaPy: A flexible and extensible software library for handling sequence graphs in Python. *Bioinformatics* **33**, 3094–3095 (2017). doi:10.1093/bioinformatics/btx398 Medline

74. H. Li, X. Feng, C. Chu, The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* **21**, 265 (2020). doi:10.1186/s13059-020-02168-z Medline

75. Y. Peng, H. C. M. Leung, S. M. Yiu, F. Y. L. Chin, "IDBA – A practical iterative de Bruijn graph de novo assembler" in *Research in Computational Molecular Biology. RECOMB 2010*, B. Berger, Ed. (Lecture Notes in Computer Science Series, vol. 6044, Springer, 2010), pp. 426–440.

76. C. Jain, S. Koren, A. Dilthey, A. M. Phillippy, S. Aluru, A fast adaptive algorithm for computing whole-genome homology maps. *Bioinformatics* **34**, i748–i756 (2018). doi:10.1093/bioinformatics/bty597 Medline

77. A. Šípek Jr., R. Mihalová, A. Panczak, L. Hrčková, M. Janashia, N. Kaspříková, M. Kohoutová, Heterochromatin variants in human karyotypes: A possible association with reproductive failure. *Reprod. Biomed. Online* **29**, 245–250 (2014). doi:10.1016/j.rbmo.2014.04.021 Medline

78. C. Jain, A. Rhie, H. Zhang, C. Chu, B. P. Walenz, S. Koren, A. M. Phillippy, Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36**, i111–i118 (2020). doi:10.1093/bioinformatics/btaa435 Medline

79. C. Jain, A. Rhie, N. Hansen, S. Koren, A. M. Phillippy, Long read mapping to repetitive reference sequences using Winnowmap2. *Nat. Methods* 10.1038/s41592-022-01457-8 (2022). doi:10.1038/s41592-022-01457-8

80. A. V. Bzikadze, P. A. Pevzner, Automated assembly of centromeres from ultra-long error-prone reads. *Nat. Biotechnol.* **38**, 1309–1316 (2020). doi:10.1038/s41587-020-0582-4 Medline

81. A. Rhie, B. P. Walenz, S. Koren, A. M. Phillippy, Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020). doi:10.1186/s13059-020-02134-9 Medline

82. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018). doi:10.1093/bioinformatics/bty191 Medline

83. R. Vaser, I. Sović, N. Nagarajan, M. Šikić, Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017). doi:10.1101/gr.214270.116 Medline

84. A. Rhie, S. A. McCarthy, O. Fedrigo, J. Damas, G. Formenti, S. Koren, M. Uliano-Silva, W. Chow, A. Fungtammasan, J. Kim, C. Lee, B. J. Ko, M. Chaisson, G. L. Gedman, L. J. Cantin, F. Thibaud-Nissen, L. Haggerty, I. Bista, M. Smith, B. Haase, J. Mountcastle, S.

Winkler, S. Paez, J. Howard, S. C. Vernes, T. M. Lama, F. Grutzner, W. C. Warren, C. N. Balakrishnan, D. Burt, J. M. George, M. T. Biegler, D. Iorns, A. Digby, D. Eason, B. Robertson, T. Edwards, M. Wilkinson, G. Turner, A. Meyer, A. F. Kautt, P. Franchini, H. W. Detrich III, H. Svardal, M. Wagner, G. J. P. Naylor, M. Pippel, M. Malinsky, M. Mooney, M. Simbirsky, B. T. Hannigan, T. Pesout, M. Houck, A. Misuraca, S. B. Kingan, R. Hall, Z. Kronenberg, I. Sović, C. Dunn, Z. Ning, A. Hastie, J. Lee, S. Selvaraj, R. E. Green, N. H. Putnam, I. Gut, J. Ghurye, E. Garrison, Y. Sims, J. Collins, S. Pelan, J. Torrance, A. Tracey, J. Wood, R. E. Dagnew, D. Guan, S. E. London, D. F. Clayton, C. V. Mello, S. R. Friedrich, P. V. Lovell, E. Osipova, F. O. Al-Ajli, S. Secomandi, H. Kim, C. Theofanopoulou, M. Hiller, Y. Zhou, R. S. Harris, K. D. Makova, P. Medvedev, J. Hoffman, P. Masterson, K. Clark, F. Martin, K. Howe, P. Flicek, B. P. Walenz, W. Kwak, H. Clawson, M. Diekhans, L. Nassar, B. Paten, R. H. S. Kraus, A. J. Crawford, M. T. P. Gilbert, G. Zhang, B. Venkatesh, R. W. Murphy, K.-P. Koepfli, B. Shapiro, W. E. Johnson, F. Di Palma, T. Marques-Bonet, E. C. Teeling, T. Warnow, J. M. Graves, O. A. Ryder, D. Haussler, S. J. O'Brien, J. Korlach, H. A. Lewin, K. Howe, E. W. Myers, R. Durbin, A. M. Phillippy, E. D. Jarvis, Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021). doi:10.1038/s41586-021-03451-0 Medline

85. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997 [q-bio.GN] (2013).

86. R. Poplin, P.-C. Chang, D. Alexander, S. Schwartz, T. Colthurst, A. Ku, D. Newburger, J. Dijamco, N. Nguyen, P. T. Afshar, S. S. Gross, L. Dorfman, C. Y. McLean, M. A. DePristo, A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018). doi:10.1038/nbt.4235 Medline

87. K. Shafin, T. Pesout, P.-C. Chang, M. Nattestad, A. Kolesnikov, S. Goel, G. Baid, J. M. Eizenga, K. H. Miga, P. Carnevali, M. Jain, A. Carroll, B. Paten, Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat. Methods* **18**, 1322–1332 (2021). doi:10.1038/s41592-021-01299-w Medline

88. G. Formenti, A. Rhie, B. P. Walenz, F. Thibaud-Nissen, K. Shafin, S. Koren, E. W. Myers, E. D. Jarvis, A. M. Phillippy, Merfin: improved variant filtering and polishing via k-mer validation. *Nat. Methods* 10.1038/s41592-022-01445-y (2022). doi:10.1038/s41592-022-01445-y

89. S. Zarate, A. Carroll, O. Krashenina, F. J. Sedlazeck, G. Jun, W. Salerno, E. Boerwinkle, R. Gibbs, Parliament2: Fast structural variant calling using optimized combinations of callers. bioRxiv 424267 [Preprint] (2018); https://doi.org/10.1101/424267.

90. F. J. Sedlazeck, P. Rescheneder, M. Smolka, H. Fang, M. Nattestad, A. von Haeseler, M. C. Schatz, Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018). doi:10.1038/s41592-018-0001-7 Medline

91. M. Kirsche, G. Prabhu, R. Sherman, B. Ni, S. Aganezov, M. C. Schatz, Jasmine: Population-scale structural variant comparison and analysis. bioRxiv 2021.05.27.445886 [Preprint] (2021); https://doi.org/10.1101/2021.05.27.445886.

92. J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, J. P.

Mesirov, Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011). doi:10.1038/nbt.1754 Medline

93. J. Schmutz, J. Wheeler, J. Grimwood, M. Dickson, J. Yang, C. Caoile, E. Bajorek, S. Black, Y. M. Chan, M. Denys, J. Escobar, D. Flowers, D. Fotopulos, C. Garcia, M. Gomez, E. Gonzales, L. Haydu, F. Lopez, L. Ramirez, J. Retterer, A. Rodriguez, S. Rogers, A. Salazar, M. Tsai, R. M. Myers, Quality assessment of the human genome sequence. *Nature* **429**, 365–368 (2004). doi:10.1038/nature02390 Medline

94. J. M. Zook, D. Catoe, J. McDaniel, L. Vang, N. Spies, A. Sidow, Z. Weng, Y. Liu, C. E. Mason, N. Alexander, E. Henaff, A. B. R. McIntyre, D. Chandramohan, F. Chen, E. Jaeger, A. Moshrefi, K. Pham, W. Stedman, T. Liang, M. Saghbini, Z. Dzakula, A. Hastie, H. Cao, G. Deikus, E. Schadt, R. Sebra, A. Bashir, R. M. Truty, C. C. Chang, N. Gulbahce, K. Zhao, S. Ghosh, F. Hyland, Y. Fu, M. Chaisson, C. Xiao, J. Trow, S. T. Sherry, A. W. Zaranek, M. Ball, J. Bobe, P. Estep, G. M. Church, P. Marks, S. Kyriazopoulou-Panagiotopoulou, G. X. Y. Zheng, M. Schnall-Levin, H. S. Ordonez, P. A. Mudivarti, K. Giorda, Y. Sheng, K. B. Rypdal, M. Salit, Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 160025 (2016). doi:10.1038/sdata.2016.25 Medline

95. K. Shafin, T. Pesout, R. Lorig-Roach, M. Haukness, H. E. Olsen, C. Bosworth, J. Armstrong, K. Tigyi, N. Maurer, S. Koren, F. J. Sedlazeck, T. Marschall, S. Mayes, V. Costa, J. M. Zook, K. J. Liu, D. Kilburn, M. Sorensen, K. M. Munson, M. R. Vollger, J. Monlong, E. Garrison, E. E. Eichler, S. Salama, D. Haussler, R. E. Green, M. Akeson, A. Phillippy, K. H. Miga, P. Carnevali, M. Jain, B. Paten, Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat. Biotechnol.* **38**, 1044–1053 (2020). doi:10.1038/s41587-020-0503-6 Medline

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

97. S. Koren, A. Rhie, B. P. Walenz, A. T. Dilthey, D. M. Bickhart, S. B. Kingan, S. Hiendleder, J. L. Williams, T. P. L. Smith, A. M. Phillippy, De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **36**, 1174–1182 (2018). doi:10.1038/nbt.4277 Medline

98. L. L. Sullivan, C. D. Boivin, B. Mravinac, I. Y. Song, B. A. Sullivan, Genomic size of CENP-A domain is proportional to total alpha satellite array size at human centromeres and expands in cancer cells. *Chromosome Res.* **19**, 457–470 (2011). doi:10.1007/s10577-011-9208-5 Medline

99. B. Mravinac, L. L. Sullivan, J. W. Reeves, C. M. Yan, K. S. Kopf, C. J. Farr, M. G. Schueler, B. A. Sullivan, Histone modifications within the human X centromere region. *PLOS ONE* **4**, e6602 (2009). doi:10.1371/journal.pone.0006602 Medline

100. A. M. Phillippy, M. C. Schatz, M. Pop, Genome assembly forensics: Finding the elusive mis-assembly. *Genome Biol.* **9**, R55–R55 (2008). doi:10.1186/gb-2008-9-3-r55 Medline

101. G. Tischler, S. Leonard, biobambam: Tools for read pair collation based algorithms on BAM files. *Source Code Biol. Med.* **9**, 13 (2014). doi:10.1186/1751-0473-9-13

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

103. E. Falconer, M. Hills, U. Naumann, S. S. S. Poon, E. A. Chavez, A. D. Sanders, Y. Zhao, M. Hirst, P. M. Lansdorp, DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat. Methods* **9**, 1107–1112 (2012). doi:10.1038/nmeth.2206 Medline

104. A. D. Sanders, E. Falconer, M. Hills, D. C. J. Spierings, P. M. Lansdorp, Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nat. Protoc.* **12**, 1151–1176 (2017). doi:10.1038/nprot.2017.029 Medline

105. D. Porubsky, A. D. Sanders, A. Taudt, M. Colomé-Tatché, P. M. Lansdorp, V. Guryev, breakpointR: An R/Bioconductor package to localize strand state changes in Strand-seq data. *Bioinformatics* **36**, 1260–1261 (2020). Medline

106. D. Porubsky, P. Ebert, P. A. Audano, M. R. Vollger, W. T. Harvey, P. Marijon, J. Ebler, K. M. Munson, M. Sorensen, A. Sulovari, M. Haukness, M. Ghareghani, P. M. Lansdorp, B. Paten, S. E. Devine, A. D. Sanders, C. Lee, M. J. P. Chaisson, J. O. Korbel, E. E. Eichler, T. Marschall, Human Genome Structural Variation Consortium, Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat. Biotechnol.* **39**, 302–308 (2021). doi:10.1038/s41587-020-0719-5 Medline

107. P. Kerpedjiev, N. Abdennur, F. Lekschas, C. McCallum, K. Dinkla, H. Strobelt, J. M. Luber, S. B. Ouellette, A. Azhir, N. Kumar, J. Hwang, S. Lee, B. H. Alver, H. Pfister, L. A. Mirny, P. J. Park, N. Gehlenborg, HiGlass: Web-based visual exploration and analysis of genome interaction maps. *Genome Biol.* **19**, 125 (2018). doi:10.1186/s13059-018-1486-1 Medline

108. R. Krishnakumar, A. Sinha, S. W. Bird, H. Jayamohan, H. S. Edwards, J. S. Schoeniger, K. D. Patel, S. S. Branda, M. S. Bartsch, Systematic and stochastic influences on the performance of the MinION nanopore sequencer across a range of nucleotide bias. *Sci. Rep.* **8**, 3159 (2018). doi:10.1038/s41598-018-21484-w Medline

109. L. I. Uralsky, V. A. Shepelev, A. A. Alexandrov, Y. B. Yurov, E. I. Rogaev, I. A. Alexandrov, Classification and monomer-by-monomer annotation dataset of suprachromosomal family 1 alpha satellite higher-order repeats in hg38 human genome assembly. *Data Brief* **24**, 103708 (2019). doi:10.1016/j.dib.2019.103708 Medline

110. T. J. Wheeler, S. R. Eddy, nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**, 2487–2489 (2013). doi:10.1093/bioinformatics/btt403 Medline

111. B. Gel, E. Serra, karyoploteR: An R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017). doi:10.1093/bioinformatics/btx346 Medline

112. R. S. Harris, thesis, Pennsylvania State University (2007).

113. B. Paten, D. Earl, N. Nguyen, M. Diekhans, D. Zerbino, D. Haussler, Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.* **21**, 1512–1528 (2011). doi:10.1101/gr.123356.111 Medline

114. S. Kovaka, A. V. Zimin, G. M. Pertea, R. Razaghi, S. L. Salzberg, M. Pertea, Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019). doi:10.1186/s13059-019-1910-1 Medline

115. M. Stanke, M. Diekhans, R. Baertsch, D. Haussler, Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008). doi:10.1093/bioinformatics/btn013 Medline

116. C. A. Miller, J. R. Walker, T. L. Jensen, W. F. Hooper, R. S. Fulton, J. S. Painter, M. A. Sekeres, T. J. Ley, D. H. Spencer, J. B. Goll, M. J. Walter, Failure to detect mutations in *U2AF1* due to changes in the GRCh38 reference sequence. *J. Mol. Diagn.* **24**, 219–223 (2022). doi:10.1016/j.jmoldx.2021.10.013 Medline

117. C. Pockrandt, M. Alzamel, C. S. Iliopoulos, K. Reinert, GenMap: Ultra-fast computation of genome mappability. *Bioinformatics* **36**, 3687–3692 (2020). doi:10.1093/bioinformatics/btaa222 Medline

118. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013). doi:10.1093/molbev/mst010 Medline

119. S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018). doi:10.1093/molbev/msy096 Medline

120. N. Saitou, M. Nei, The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987). Medline

121. M. Kimura, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120 (1980). doi:10.1007/BF01731581 Medline

122. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014). doi:10.1093/bioinformatics/btu033 Medline

123. M. Nei, S. Kumar, *Molecular Evolution and Phylogenetics* (Oxford Univ. Press, 2000).

124. R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017). doi:10.1038/nmeth.4197 Medline

125. C. Soneson, M. Love, M. Robinson, Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res.* **4**, 1521 (2016). doi:10.12688/f1000research.7563.2 Medline.

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

127. M. Ghareghani, D. Porubskỳ, A. D. Sanders, S. Meiers, E. E. Eichler, J. O. Korbel, T. Marschall, Strand-seq enables reliable separation of long reads by chromosome via expectation maximization. *Bioinformatics* **34**, i115–i123 (2018). doi:10.1093/bioinformatics/bty290 Medline

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