Table S1: Data availability (separate file)
The assemblies of CHM13 and HG002-ChrX generated by this study are available from NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/). All raw sequencing data used by this project was generated by previous studies and is listed by NCBI SRA accession (https://www.ncbi.nlm.nih.gov/sra/). References to the respective studies are provided in the main text.

Table S2: ddPCR array assays (separate file)
Primers (fwd/rev), the restriction enzyme, gDNA concentration, and normalization gene used to validate copy count of repeat arrays in CHM13. Details for the single-copy genes used for normalization are in Table S3.

Table S3: ddPCR gene assays (separate file):
Primers (fwd/rev), the restriction enzyme, gDNA concentration, and the location (in GRCh38) of the single-copy genes used to normalize ddPCR copy number values.

Table S4: Summary statistics for reads recruited across different satellite classes (separate file):
For all types of data, except for HiFi reads, the statistics are additionally broken out by strand (orientation with respect to the canonical unit). The 'Total length' is the total size of all 1 kbp windows assigned to a particular satellite class. 'Expected length' was computed by multiplying the total length of reads within a particular dataset by the fraction of the assembly assigned to a particular satellite class (more precisely, fraction of the assembly bases within the 1 kbp windows that were assigned to a particular satellite class based on k-mer analysis). Assigned assembly fraction values were 0.44% for HSat1, 0.94% for HSat2, 1.56% for HSat3, and 2.79% for AlphaSat. Enrichment was computed based on the 'total' vs 'expected' length values, as ('total' - 'expected' / 'expected'). 'Avg. read length' provides average length across assigned reads/subreads.

Table S5: Coordinates of alpha and human satellite arrays in v1.0 assembly (separate file):
Coordinates of alpha and human satellite arrays in v1.0 assembly of CHM13 that were validated by TandemTools. The rDNA arrays on chromosomes 13, 14, 15, 21, and 22 were excluded from this validation. Corresponding TandemTools plots are shown in Fig. S27.

Table S6: CHM13v1.1 Annotation gene summary (separate file):
The table lists counts of genes by category and biotype for genes identified by the Liftoff + CAT annotation pipeline. The missing gene column details gene counts found in GENCODE but missing in the CHM13 annotation.

Table S7: CHM13v1.1 Annotation transcript summary (separate file):

The table lists counts of transcripts by category and biotype for transcripts identified by the Liftoff + CAT annotation pipeline. The missing transcript column details gene counts found in GENCODE but missing in the CHM13 annotation.

Table S8: Frameshits in genes and transcripts (separate file):
The table shows the coordinates, IDs, and source name of genes and transcripts which have a frameshift versus the GENCODE version in CHM13v1.1. The frameshift may be a true difference in the CHM13 genome or an error in the assembly. The last two columns show whether any frameshifts intersect known issues in the assembly or are on the medically relevant list of genes from GIAB.

Table S9: Missing GENCODE genes (separate file):
The table lists the full names of GENCODE genes (ID and name) missing in the CHM13 annotation but present in GRCh38. The genes are annotated with their biotype along with a reason why they could not be mapped to CHM13. The repeat type annotation indicates if the gene intersects a repeat region, which may indicate the gene is copy-number variable. Lastly, the medically relevant column specifies if the missing gene name is on the GIAB list of medically relevant genes.

Table S10: Missing GENCODE transcripts (separate file):
As in Table S9, the table lists the full names of GENCODE transcripts (ID and name) missing in the CHM13 annotation. The transcripts are annotated with their biotype.

Table S11: Genes found in CHM13 not present in GRCh38 (separate file):
The table lists the closest GENCODE ID, gene name, and biotype where available. In cases when a gene was identified using Iso-Seq, it has the biotype StringTie and no associated GENCODE information. The novel region column indicates whether the gene falls into a region on CHM13 that had no primary alignments from GRCh38. The GRCh38 issue column indicates if the gene falls into a region matching a known GRCh38, lifted over to CHM13 coordinates.The next column indicates if the gene is either in the novel region or in a GRCh38 known issue. Lastly, the medically relevant column indicates if the closest GENCODE name is in the GIAB medically relevant gene list.

Table S12: Transcripts found in CHM13 not present in GRCh38 (separate file):
As in Table S11, the closest GENCODE ID, transcript name, and biotype are given where available. Novel region and GRCh38 known issues as in Table S11.

Table S13: Identity of protein-coding genes in CHM13 not in GRCh38 to their closest GENCODE match (separate file):

For each protein-coding gene exclusive to CHM13, the table reports the closest paralog identified in GENCODE and both the nucleotide and amino acid identity and extent of the match. The assembly issue column indicates if the CHM13 gene falls in a region with a known assembly issue. The frameshift column indicates if the CHM13 gene has a frameshift versus its GENCODE match (see Table S8 for the full frameshift list).

Table S14: Total length of unmappable regions across assemblies (separate file):
Total length of 'unmappable' regions (Mbp) and the fraction of the assembly they represent (using 3.1 Gbp as genome size). A region is considered unmappable by a particular technology if less than 1 kbp or 33% of the read length (whichever is smaller) of sequence is represented by unique k-mers (where k equals the technology's characteristic perfect run size). Note that regions with gaps can be classified as mappable given a unique sequence at their boundaries, allowing reads to extend into the gap. To estimate the potential effects of consensus quality on mappability, we added random errors to GRCh38 at the frequency of 0.01% (Q40) (with 45% of the errors being insertion, 45% -- deletions, and 10% -- substitutions). We observe that CHM13 v1.0 assembly is expected to be more mappable with long read technologies than GRCh38, despite having a higher repeat content. The increased mappability is not due to consensus quality as GRCh38 with added errors is still less mappable than CHM13 v1.0.

Table S15: Identified FRG1 SD paralogs in CHM13 v1.0, CHM13v1.1 and GRCh38 (separate file):
Annotated segmental duplications of the ancestral FRG1 locus (chr4:189940872-189963192, GRCh38, "Segmental Dups" track; chr4:193304806-193328432, CHM13v1.0, SEDEF track) were intersected with annotated genes (GENCODEv36 for GRCh38 and CAT+Liftoff v4 for CHM13v1.0) to identify FRG1 paralogs.

Table S16: *FRG1* divergence (separate file):
Pairwise identity for all *FRG1* paralogs in the CHM13 assembly (see Table S15 for gene locations). The number of base substitutions per site from between sequences was calculated and percent identity calculated by taking one minus this value. Analyses were conducted using the Kimura 2-parameter model and MEGA X as for the tree in Fig. S39A. Similarity between FRG1DP and FRG1BP4~10 is highlighted in blue.

Table S17: *FRG1* Expression (separate file):
The table lists transcripts per million (TPM), colored from white to red by abundance. 'Gene': The gene name in Fig. 4. 'Cen/Acro', 'Chr', 'Start', 'End': the location of the gene (centromeric satellite or acrocentric) along with the chromosome and coordinates in the CHM13v1.0 and CHM13v1.1 assemblies. 'Gene (GENCODEv35)': The gene name of the closest GENCODEv35 gene. 'Ensembl ID': closest ID in Ensembl. 'T2T Annotation ID': the ID of the annotated gene

in the CHM13 assembly. The remaining columns show abundance estimated from Salmon and short-reads aligned with unique 21 and 51-mer markers along with IsoSeq reads.