Title

The complete sequence of a human genome

Journal

Science, 376(6588)

ISSN

0036-8075

Authors

Nurk, Sergey

Koren, Sergey

Rhie, Arang

et al.

Publication Date

2022-04-01

DOI

10.1126/science.abj6987

Peer reviewed

HHS Public Access

Author manuscript

Science. Author manuscript; available in PMC 2022 June 10.

Published in final edited form as:

Science. 2022 April ; 376(6588): 44–53. doi:10.1126/science.abj6987.

The complete sequence of a human genome*

A full list of authors and affiliations appears at the end of the article.

Abstract

Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished. Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion base pair (bp) sequence of a human genome, T2T-CHM13, that includes gapless assemblies for all chromosomes except Y, corrects errors in the prior references, and introduces nearly 200 million bp of sequence containing 1,956 gene predictions, 99 of which are predicted to be protein coding. The completed regions include all centromeric satellite arrays, recent segmental duplications, and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies.

One-Sentence Summary:

Twenty years after the initial drafts, a truly complete sequence of a human genome reveals what has been missing.

The current human reference genome was released by the Genome Reference Consortium (GRC) in 2013 and most recently patched in 2019 (GRCh38.p13) (1). This reference traces

---

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

†These authors contributed equally to this work

‡Present address: Oxford Nanopore Technologies Inc.; Lexington, MA, USA

Author contributions: *Analysis teams (leads*): Assembly:* SN*, SK*, MR*, MA, HC, CSC, RD, EG, MKi, MKo, HL, TM, EWM, IS, BPW, AW, AMP. *Acrocentrics:* AMP*, JLG*, MR, SEA, MB, RD, LGL, TP. *Validation:* AR*, AVB*, AM*, MA*, AMM*, KS*, WC, LGL, TD, GF, AF, KH, CJ, EDJ, DP, VAS, YS, BAS, FTN, JT, JMDW, AMP. *Segmental duplications:* MRV*, EEE*, SN, SK, MD, PCD, AG, GAL, DP, CJS, DCS, MYD, WT, KHM, AMP. *Satellite annotation:* NA*, IAA*, KHM*, AVB, LU, TD, LGL, PAP, EIR, ASt, BAS, AMP. *Epigenetics:* AG*, WT*, SK, AR, MRV, NA, SJH, GAL, GVC, MCS, RJO, EEE, KHM, AMP. *Variants:* SA*, DCS*, SMY*, SZ*, RCM*, MYD*, JMZ*, MCS*, NFH, MKi, JM, DEM, NDO, JAR, FJS, KS, ASh, JW, CX, AMP. *Repeat annotation:* SJH*, RJO*, AG, PGSG, GAH, LGL, AFAS, JMS. *Gene annotation:* MD*, MH*, ASh*, SN, SK, PCD, ITF, SLS, FTN, AMP. *Browsers:* MD*, NCC, PK. *Data generation:* SJH, GGB, SYB, GVC, RSF, TAGL, IMH, MWH, MJ, JK, VVM, JCM, BP, PP, ACY, US, MYD, JLG, RJO, WT, EEE, KHM, AMP. *Computational resources:* CSC, AF, RJO, MCS, KHM, AMP. *Manuscript draft:* AMP. *Figures:* SK, SN, AMP, AR. *Editing:* AMP, SN, SK, AR, EEE, KHM, with the assistance of all authors. *Supplement:* SN, SK, with the assistance of the working groups. *Supervision:* RCM, MYD, IAA, JLG, RJO, WT, JMZ, MCS, EEE, KHM, AMP. *Conceptualization:* EEE, KHM, AMP.

Data and materials availability: The T2T-CHM13 and T2T-HG002-ChrX assemblies generated by this study are archived under NCBI GenBank accessions GCA_009914755 and CP086568, respectively. The raw sequencing data was described in prior studies and is summarized in table S1. For convenience, links to the sequence data and genome browsers are also available from https://github.com/marbl/CHM13. Supplementary data for fig. S39 and the string graph construction code are archived at Zenodo (61) and also https://github.com/snurk/sg_sandbox. CHM13hTERT cells were obtained for research use via a material transfer agreement with U. Surti and the University of Pittsburgh.

its origin to the publicly funded Human Genome Project (2) and has been continually improved over the past two decades. Unlike the competing Celera effort (3) and most modern sequencing projects based on “shotgun” sequence assembly (4), the GRC assembly was constructed from sequenced bacterial artificial chromosomes (BACs) that were ordered and oriented along the human genome via radiation hybrid, genetic linkage, and fingerprint maps. However, limitations of BAC cloning led to an underrepresentation of repetitive sequences, and the opportunistic assembly of BACs derived from multiple individuals resulted in a mosaic of haplotypes. As a result, several GRC assembly gaps are unsolvable due to incompatible structural polymorphisms on their flanks, and many other repetitive and polymorphic regions were left unfinished or incorrectly assembled (5).

The GRCh38 reference assembly contains 151 Mbp of unknown sequence distributed throughout the genome, including pericentromeric and subtelomeric regions, recent segmental duplications, ampliconic gene arrays, and ribosomal DNA (rDNA) arrays, all of which are necessary for fundamental cellular processes (Fig. 1A). Some of the largest reference gaps include human satellite (HSat) repeat arrays and the short arms of all five acrocentric chromosomes, which are represented in GRCh38 as multi-megabase stretches of unknown bases (Figs. 1B and 1C). In addition to these apparent gaps, other regions of GRCh38 are artificial or are otherwise incorrect. For example, the centromeric alpha satellite arrays are represented as computationally generated models of alpha satellite monomers to serve as decoys for resequencing analyses (6), while sequence assigned to the short arm of Chromosome 21 appears falsely duplicated and poorly assembled (7). When compared to other human genomes, GRCh38 also shows a genome-wide deletion bias that is indicative of incomplete assembly (8). Despite finishing efforts from both the Human Genome Project (9) and GRC (1) that improved the quality of the reference, there was limited progress towards closing the remaining gaps in the years that followed (Fig. 1D).

Long-read shotgun sequencing overcomes the limitations of BAC-based assembly and bypasses the challenges of structural polymorphism between genomes. PacBio’s multi-kilobase, single-molecule reads (10) proved capable of resolving complex structural variation and gaps in GRCh38 (8, 11), while Oxford Nanopore’s >100 kbp “ultra-long” reads (12), enabled complete assemblies of a human centromere (ChrY) (13) and, later, an entire chromosome (ChrX) (14). However, the high error rate (>5%) of these technologies posed challenges for the assembly of long, near-identical repeat arrays. PacBio’s most recent “HiFi” circular consensus sequencing offers a compromise of 20 kbp read lengths with an error rate of 0.1% (15). Whereas ultra-long reads are useful for spanning repeats, HiFi reads excel at differentiating subtly diverged repeat copies or haplotypes (16).

To finish the last remaining regions of the genome, we leveraged the complementary aspects of PacBio HiFi and Oxford Nanopore ultra-long read sequencing to assemble the uniformly homozygous CHM13hTERT cell line (hereafter, CHM13) (17). The resulting T2T-CHM13 reference assembly removes a 20-year-old barrier that has hidden 8% of the genome from sequence-based analysis, including all centromeric regions and the entire short arms of five human chromosomes. Here we describe the construction, validation, and initial analysis of a truly complete human reference genome and discuss its potential impact on the field.

Cell line and sequencing

As with many prior reference genome improvement efforts (1, 8, 17–20), including the T2T assemblies of human chromosomes X (14) and 8 (21), we targeted a complete hydatidiform mole for sequencing. Most CHM genomes arise from the loss of the maternal complement and duplication of the paternal complement postfertilization and are, therefore, homozygous with a 46,XX karyotype (22). Sequencing of CHM13 confirmed nearly uniform homozygosity, with the exception of a few thousand heterozygous variants and a megabase-scale heterozygous deletion within the rDNA array on Chromosome 15 (23) (figs. S1 to S2). Local ancestry analysis shows the majority of the CHM13 genome is of European origin, including regions of Neanderthal introgression, with some predicted admixture (23) (Fig. 1A). Compared to diverse samples from the 1000 Genomes Project (1KGP) (24), CHM13 possesses no apparent excess of singleton alleles or loss-of-function variants (25).

We extensively sequenced CHM13 with multiple technologies (23), including 30× PacBio circular consensus sequencing (HiFi) (16, 20), 120× Oxford Nanopore ultra-long read sequencing (ONT) (14, 21), 100× Illumina PCR-Free sequencing (ILMN) (1), 70× Illumina / Arima Genomics Hi-C (Hi-C) (14), BioNano optical maps (14), and Strand-seq (20) (table S1). To enable assembly of the highly repetitive centromeric satellite arrays and closely related segmental duplications, we developed methods for assembly, polishing, and validation that better utilize these available datasets.

Genome assembly

The basis of the T2T-CHM13 assembly is a high-resolution assembly string graph (26) built directly from HiFi reads. In a bidirected string graph, nodes represent unambiguously assembled sequences and edges correspond to the overlaps between them, due to either repeats or true adjacencies in the underlying genome. The CHM13 graph was constructed using a purpose-built method that combines components from existing assemblers (16, 27) along with specialized graph processing (23). Most HiFi errors are small insertions or deletions within homopolymer runs and simple sequence repeats (16), so homopolymer runs were first “compressed” to a single nucleotide (e.g., $[A]_n$ becomes $[A]_1$ for $n > 1$). All compressed reads were then aligned to one another to identify and correct small errors, and differences within simple sequence repeats were masked. After compression, correction, and masking, only exact read overlaps were considered during graph construction, followed by iterative graph simplification (23).

In the resulting graph, most components originate from a single chromosome and have an almost linear structure (Fig. 2A), which suggests few perfect repeats greater than roughly 10 kbp exist between different chromosomes or distant loci. Two notable exceptions are the five acrocentric chromosomes, which form a single connected component in the graph, and a recent multi-megabase HSat3 duplication on Chromosome 9, consistent with the 9qh+ karyotype of CHM13 (fig. S3). Minor fragmentation of the chromosomes into multiple components resulted from a lack of HiFi sequencing coverage across GA-rich sequences (16). These gaps were later filled with a prior ONT-based assembly (CHM13v0.7) (14).

Ideally, the complete sequence for each chromosome should exist as a walk through the string graph where some nodes may be traversed multiple times (repeats) and some not at all (errors and heterozygous variants). To help identify the correct walks, we estimated coverage depth and multiplicity of the nodes (23), which allowed most tangles to be manually resolved as unique walks visiting each node the appropriate number of times (Figs. 2B and fig. S4). In the remaining cases, the correct path was ambiguous and required integration of ONT reads (Figs. 2C and 2D). Where possible, ONT reads were aligned to candidate traversals or directly to the HiFi graph (28) to guide the correct walk (fig. S5), but more elaborate strategies were required for recent satellite array duplications on chromosomes 6 and 9 (23). Only the five rDNA arrays, constituting approximately 10 Mbp of sequence, could not be resolved with the string graph and required a specialized approach (described below). An accurate consensus sequence for the selected graph walks was computed from the uncompressed HiFi reads (23), resulting in the CHM13v0.9 draft assembly.

For comparative genomics of the centromere (29, 30), we repeated this process on an additional X chromosome from the Coriell GM24385 cell line (NIST ID: HG002). The resulting T2T-HG002-ChrX assembly shows comparable accuracy to T2T-CHM13 (23) (figs. S6 to S8).

rDNA assembly

The most complex region of the CHM13 string graph involves the human ribosomal DNA arrays and their surrounding sequence (Fig. 2D). Human rDNAs are 45 kbp near-identical repeats that encode the 45S rRNA and are arranged in large, tandem repeat arrays embedded within the short arms of the acrocentric chromosomes. The length of these arrays varies between individuals (36), and even somatically, especially with aging and certain cancers (37). A typical diploid human genome has an average of 315 rDNA copies, with a standard deviation of 104 copies (36). We estimate that the diploid CHM13 genome contains approximately 400 rDNA copies based on ILMN depth of coverage (23) (fig. S9), or 409 ± 9 (mean ± s.d.) rDNA copies by ddPCR (fig. S10).

To assemble these highly dynamic regions of the genome, and overcome limitations of the string graph construction (23) (fig. S11), we constructed sparse de Bruijn graphs for each of the five rDNA arrays (38) (fig. S12). ONT reads were aligned to the graphs to identify a set of walks, which were converted to sequence, segmented into individual rDNA units, and clustered into “morphs” according to their sequence similarity. The copy number of each morph was estimated from the number of supporting ONT reads, and consensus sequences were polished with mapped HiFi reads. ONT reads spanning two or more rDNA units were used to build a morph graph representing the structure of each array (fig. S12).

The shorter arrays on Chromosomes 14 and 22 consist of a single primary morph arranged in a head-to-tail array, whereas the longer arrays on Chromosomes 13, 15, and 21 exhibit a more mosaic structure involving multiple, interspersed morphs. In these cases, the ONT reads were not long enough to fully resolve the ordering, and the primary morphs were artificially arranged in consecutive blocks reflecting their estimated copy number. These three arrays capture the chromosome-specific morphs but should be treated as model

sequences. The final T2T-CHM13 assembly contains 219 complete rDNA copies, totaling 9.9 Mbp of sequence.

Assembly validation and polishing

To evaluate concordance between the reads and the assembly we mapped all available primary data, including HiFi, ONT, ILMN, Strand-seq, and Hi-C, to the CHM13v0.9 draft assembly to identify both small and structural variants (see reference (31) for a complete description). Manual curation corrected 4 large and 993 small errors, resulting in the CHM13v1.0 assembly, and identified 44 large and 3,901 small heterozygous variants (31). Further telomere polishing and addition of the rDNA arrays (23) resulted in a complete, telomere-to-telomere assembly of a human genome, T2T-CHM13v1.1.

The T2T-CHM13 assembly is consistent with previously validated assemblies of chromosomes X (14) and 8 (21), and the sizes of assembled satellite arrays match ddPCR copy-number estimates for those tested (fig. S10 and tables S2 and S3). Mapped Strand-seq (figs. S13 and S14) and Hi-C (fig. S15) data show no signs of misorientations or other large-scale structural errors. The assembly correctly resolves 644 of 647 previously sequenced CHM13 BACs at >99.99% identity, with the 3 others reflecting errors in the BACs themselves (figs. S16 to S19).

Mapped sequencing read depth shows uniform coverage across all chromosomes (Fig. 3A), with 99.86% of the assembly within three standard deviations of the mean coverage for either HiFi or ONT (HiFi coverage 34.70 ± 7.03, ONT coverage 116.16 ± 16.96, excluding the mitochondrial genome). Ignoring the 10 Mbp of rDNA sequence, where most of the coverage deviation resides, 99.99% of the assembly is within three standard deviations (23). Alignment-free analysis of ILMN and HiFi copy number data also show concordance with the assembly (figs. S20 and S21). This is consistent with uniform coverage of the genome and confirms both the accuracy of the assembly and the absence of aneuploidy in the sequenced CHM13 cells.

Coverage increases or decreases were observed across multiple satellite arrays (Figs. 3B to 3D). However, given the uniformity of coverage across these arrays, association with specific satellite classes, and the sometimes opposite effect observed for HiFi and ONT, we hypothesize that these anomalies are related to biases introduced during sample preparation, sequencing, or basecalling, rather than assembly error (23) (figs. S22 to S26 and table S4). While the specific mechanisms require further investigation, prior studies have noted similar biases within certain satellite arrays and sequence contexts for both ONT and HiFi (32, 33).

Being the most difficult regions of the genome to assemble, we performed targeted validation of long tandem repeats to identify any errors missed by the genome-wide approach. The assembled rDNA morphs, being only 45 kbp each, were manually validated via inspection of the read alignments used for polishing. Alpha satellite higher-order repeats (HOR) were validated using a purpose-built method (34) (fig. S27 and table S5) and compared to independent ILMN-based HOR copy number estimates (fig. S28). All centromeric satellite arrays, including beta satellite (BSat) and human satellite (HSat)

repeats, were further validated by measuring the ratio of primary to secondary variants identified by HiFi reads (35) (fig. S29).

The consensus accuracy of the T2T-CHM13 assembly is estimated to be approximately 1 error per 10 Mbp (23, 31), which exceeds the historical standard of “finished” sequence by orders of magnitude. However, regions of low HiFi coverage were found to be associated with an enrichment of potential errors, as estimated from both HiFi and ILMN data (31). To guide future use of the assembly, we have cataloged all low-coverage, low-confidence, and known heterozygous sites identified by the above validation procedures (31). The total number of bases covered by potential issues in the T2T-CHM13 assembly is just 0.3% of the total assembly length compared to 8% for GRCh38 (Fig. 3A).

A truly complete genome

T2T-CHM13 includes gapless telomere-to-telomere assemblies for all 22 human autosomes and Chromosome X, comprising 3,054,815,472 bp of nuclear DNA, plus a 16,569 bp mitochondrial genome. This complete assembly adds or corrects 238 Mbp of sequence that does not co-linearly align to GRCh38 over a 1 Mbp interval (i.e., is non-syntenic), primarily comprising centromeric satellites (76%), non-satellite segmental duplications (19%), and rDNAs (4%) (Fig. 1C). 182 Mbp of sequence has no primary alignments to GRCh38 and is exclusive to T2T-CHM13. As a result, T2T-CHM13 increases the number of known genes and repeats in the human genome (Table 1).

To provide an initial annotation, we used both the Comparative Annotation Toolkit (CAT) (39) and Liftoff (40) to project the GENCODE v35 (41) reference annotation onto the T2T-CHM13 assembly. Additionally, CHM13 Iso-Seq transcriptome reads were assembled into transcripts and provided as complementary input to CAT. A comprehensive annotation was built by combining the CAT annotation with genes identified only by Liftoff (23).

The draft T2T-CHM13 annotation totals 63,494 genes and 233,615 transcripts, of which 19,969 genes (86,245 transcripts) are predicted to be protein coding, with 683 predicted frameshifts in 385 genes (469 transcripts) (Table 1, fig. S30, tables S6 to S8). Only 263 GENCODE genes (448 transcripts) are exclusive to GRCh38 and have no assigned ortholog in the CHM13 annotation (tables S9 and S10). Of these, 194 are due to a lower copy number in the CHM13 annotation (fig. S31), 46 do not align well to CHM13, and 23 correspond to known false-duplications in GRCh38 (25) (fig. S32). The majority of these genes are non-coding and associated with repetitive elements. Only 4 are annotated as being medically relevant (*CFHR1*, *CFHR3*, *OR51A2*, *UGT2B28*), all of which are due to lower copy number, and the only protein coding genes that align poorly are immunoglobulin and T-cell receptor genes, which are known to be highly diverse.

In comparison, a total of 3,604 genes (6,693 transcripts) are exclusive to CHM13 (tables S11 and S12). Most of these genes represent putative paralogs and localize to pericentromeric regions and the short arms of the acrocentrics, including 876 rRNA transcripts. Only 48 of the CHM13-exclusive genes (56 transcripts) were predicted solely from de novo assembled transcripts. Of all genes exclusive to CHM13, 140 are predicted to be protein coding based

on their GENCODE paralogs and have a mean of 99.5% nucleotide and 98.7% amino acid identity to their most similar GRCh38 copy (table S13). While some of these additional paralogs may be present (but unannotated) in GRCh38 (23), 1,956 of the genes exclusive to CHM13 (99 protein coding) are in regions with no primary alignment to GRCh38 (table S11). A broader set of 182 multi-exon protein coding genes fall within non-syntenic regions, 36% of which were confirmed to be expressed in CHM13 (42).

Compared to GRCh38, T2T-CHM13 is a more complete, accurate, and representative reference for both short- and long-read variant calling across human samples of all ancestries (25). Reanalysis of 3,202 short-read datasets from the 1KGP showed that T2T-CHM13 simultaneously reduces both false-negative and false-positive variant calls due to the addition of 182 Mbp of missing sequence and the exclusion of 1.2 Mbp of falsely duplicated sequence in GRCh38. These improvements, combined with a lower frequency of rare variants and errors in T2T-CHM13, eliminate tens of thousands of spurious variants per 1KGP sample (25). In addition, the T2T-CHM13 reference was found to be more representative of human copy number variation than GRCh38 when compared against 268 human genomes from the Simons Genome Diversity Project (SGDP) (42, 43). Specifically, within non-syntenic segmentally duplicated regions of the genome, T2T-CHM13 is nine times more predictive of SGDP copy number than GRCh38 (42). These results underscore both the quality of the assembly and the genomic stability of the cell line from which it was derived.

Acrocentric chromosomes

T2T-CHM13 uncovers the genomic structure of the short arms of the five acrocentric chromosomes, which, despite their importance for cellular function (44), have remained largely unsequenced to date. This omission has been due to their enrichment for satellite repeats and segmental duplications, which has prohibited sequence assembly and limited their characterization to cytogenetics, restriction mapping, and BAC sequencing (45–47). All five of CHM13's short arms follow a similar structure consisting of an rDNA array embedded within distal and proximal repeat arrays (Fig. 4). From telomere to centromere, the short arms vary in size from 10.1 Mbp (Chr14) to 16.7 Mbp (Chr15), with a combined length of 66.1 Mbp.

Compared to other human chromosomes, the short arms of the acrocentrics are unusually similar to one another. Specifically, we find that 5 kbp windows align with a median identity of 98.7% between the short arms, creating many opportunities for interchromosomal exchange (Fig. 4). This high degree of similarity is presumably due to recent non-allelic or ectopic recombination stemming from their colocalization in the nucleolus (46). Additionally, considering an 80% identity threshold, no 5 kbp window on the short arms is unique and 96% of the non-rDNA sequence can be found elsewhere in the genome, suggesting the acrocentrics are dynamic sources of segmental duplication.

CHM13's rDNA arrays vary in size from 0.7 Mbp (Chr14) to 3.6 Mbp (Chr13) and are in the expected arrangement, organized as head-to-tail tandem arrays with all 45S transcriptional units pointing towards the centromere. No inversions were noted within

the arrays and nearly all rDNA units are full length, in contrast to some prior studies that reported embedded inversions and other non-canonical structures (47, 48). Each array appears highly homogenized, and there is more variation between rDNA units on different chromosomes than within chromosomes (fig. S33), suggesting that intra-chromosomal exchange of rDNA units via non-allelic homologous recombination is more common than inter-chromosomal exchange.

Many 45S gene copies on the same chromosome are identical to one another, while the identity of the most frequent 45S morphs between chromosomes ranges from 99.4–99.7%. A Chromosome 15 rDNA morph shows the highest identity (98.9%) to the current KY962518.1 rDNA reference sequence, originally derived from a human Chromosome 21 BAC clone (47). As expected, the 13 kbp 45S is more conserved than the intergenic spacer (IGS), with all major 45S morphs aligning between 99.4–99.6% identity to KY962518.1. Certain rDNA variants appear chromosome-specific, including single-nucleotide variants within the 45S and its upstream promoter region (fig. S34). The most evident variants are repeat expansions and contractions within the tandem “R” repeat that immediately follows the 45S and the CT-rich “Long” repeat located in the middle of the IGS. The most frequent morph in each array can be uniquely distinguished by these two features (fig. S35).

From the telomere to the rDNA array, the structure of all five distal short arms follows a similar pattern involving a symmetric arrangement of inverted segmental duplications and ACRO, HSat3, BSat, and HSat1 repeats (Fig. 4); however, the sizes of these repeat arrays varies among chromosomes. Chromosome 13 is missing the distal half of the inverted duplication and has an expanded HSat1 array relative to the others. Despite their variability in size, all satellite arrays share a high degree of similarity (typically >90% identity) both within and between acrocentric chromosomes. Chromosomes 14 and 22 also feature the expansion of a 64-bp Alu-associated satellite repeat (“Walu”) within the distal inverted duplication (49), the location of which was confirmed via FISH (fig. S36). The distal junction (DJ) immediately prior to the rDNA array includes centromeric repeats (CER) and a highly conserved and actively transcribed 200 kbp palindromic repeat, which agrees with previous characterizations of the rDNA flanking sequences (46, 50).

Extending from the rDNA array to the centromere, the proximal short arms are larger in size and show a higher diversity of structures including shuffled segmental duplications (42), composite transposable element arrays (49), satellite arrays (including HSat3, BSat, HSat1, HSat5), and alpha satellite arrays (both monomeric and HORs) (30). Some proximal BSat arrays show a mosaic inversion structure that was also observed in HSat arrays elsewhere in the genome (30) (fig. S37). The proximal short arms of chromosomes 13, 14, and 21 appear to share the highest degree of similarity with a large region of segmental duplication including similar HOR subsets and a central and highly methylated SST1 array (Fig. 4). This coincides with these three chromosomes being most frequently involved in Robertsonian translocations (51). Alpha satellite HORs on chromosomes 13/21 and chromosomes 14/22 also share high similarity within each pair, but not between them (52, 53). Non-satellite sequences within these segmental duplications often exceed 99% identity and show evidence of transcription (29, 42, 49). Using the T2T-CHM13 reference as a basis, further study

of additional genomes is now needed to understand which of these features are conserved across the human population.

Analyses and resources

A number of companion studies were carried out to characterize the complete sequence of a human genome, including comprehensive analyses of centromeric satellites (30), segmental duplications (42), transcriptional (49) and epigenetic profiles (29), mobile elements (49), and variant calls (25). Up to 99% of the complete CHM13 genome can be confidently mapped with long-read sequencing, opening these regions of the genome to functional and variational analysis (23) (fig. S38 and table S14). We have produced a rich collection of annotations and omics datasets for CHM13, including RNA-Seq (30), Iso-Seq (21), PRO-Seq (49), CUT&RUN (30), and ONT methylation (29) experiments, and have made these datasets available via a centralized UCSC Assembly Hub genome browser (54).

To highlight the utility of these genetic and epigenetic resources mapped to a complete human genome, we provide the example of a segmentally duplicated region of the Chromosome 4q subtelomere that is associated with facioscapulohumeral muscular dystrophy (FSHD) (55). This region includes FSHD region gene 1 (*FRG1*), FSHD region gene 2 (*FRG2*), and an intervening D4Z4 macrosatellite repeat containing the double homeobox 4 (*DUX4*) gene that has been implicated in the etiology of FSHD (56). Numerous duplications of this region throughout the genome have complicated past genetic analyses of FSHD.

The T2T-CHM13 assembly reveals 23 paralogs of *FRG1* spread across all acrocentric chromosomes as well as chromosomes 9 and 20 (Fig. 5A). This gene appears to have undergone recent amplification in the great apes (57), and approximate locations of *FRG1* paralogs were previously identified by fluorescence in situ hybridization (58). However, only 9 *FRG1* paralogs are found in GRCh38, hampering sequence-based analysis.

One of the few *FRG1* paralogs included in GRCh38, *FRG1DP*, is located in the centromeric region of Chromosome 20 and shares high identity (97%) with several paralogs (*FRG1BP4–10*) (23) (fig. S39 and tables S15 and S16). When mapping HiFi reads, absence of the additional *FRG1* paralogs in GRCh38 causes their reads to incorrectly align to *FRG1DP* resulting in many false-positive variants (Fig. 5B). Most *FRG1* paralogs appear present in other human genomes (Fig. 5C), and all except *FRG1KP2* and *FRG1KP3* have upstream CpG islands and some degree of expression evidence in CHM13 (Fig. 5D, table S17). Any variants within these paralogs, and others like them, will be overlooked when using GRCh38 as a reference.

Future of the human reference genome

The T2T-CHM13 assembly adds five full chromosome arms and more additional sequence than any genome reference release in the past 20 years (Fig. 1D). This 8% of the genome has not been overlooked due to its lack of importance, but rather due to technological limitations. High accuracy long-read sequencing has finally removed this technological barrier, enabling comprehensive studies of genomic variation across the entire human

genome, which we expect to drive future discovery in human genomic health and disease. Such studies will necessarily require a complete and accurate human reference genome.

CHM13 lacks a Y chromosome, and homozygous Y-bearing CHMs are non-viable, so a different sample type will be required to complete this last remaining chromosome. However, given its haploid nature, it should be possible to assemble the Y chromosome from a male sample using the same methods described here, and supplement the T2T-CHM13 reference assembly with a Y chromosome as needed.

Extending beyond the human reference genome, large-scale resequencing projects have revealed genomic variation across human populations. Our reanalyses of 1KGP (25) and SGDP (42) datasets have already shown the advantages of T2T-CHM13, even for short-read analyses. However, these studies give only a glimpse of the extensive structural variation that lies within the most repetitive regions of the genome assembled here. Long-read resequencing studies are now needed to comprehensively survey polymorphic variation and reveal any phenotypic associations within these regions.

Although CHM13 represents a complete human haplotype, it does not capture the full diversity of human genetic variation. To address this bias, the Human Pangenome Reference Consortium (HPRC) (59) has joined with the T2T Consortium to build a collection of high-quality reference haplotypes from a diverse set of samples. Ideally, all genomes could be assembled at the quality achieved here, but automated T2T assembly of diploid genomes presents a difficult challenge that will require continued development. Until this goal is realized, and any human genome can be completely sequenced without error, the T2T-CHM13 assembly represents a more complete, representative, and accurate reference than GRCh38.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Sergey Nurk[1,†], Sergey Koren[1,†], Arang Rhie[1,†], Mikko Rautiainen[1,†], Andrey V. Bzikadze[2], Alla Mikheenko[3], Mitchell R. Vollger[4], Nicolas Altemose[5], Lev Uralsky[6,7], Ariel Gershman[8], Sergey Aganezov[9,‡], Savannah J. Hoyt[10], Mark Diekhans[11], Glennis A. Logsdon[4], Michael Alonge[9], Stylianos E. Antonarakis[12], Matthew Borchers[13], Gerard G. Bouffard[14], Shelise Y. Brooks[14], Gina V. Caldas[15], Nae-Chyun Chen[9], Haoyu Cheng[16,17], Chen-Shan Chin[18], William Chow[19], Leonardo G. de Lima[13], Philip C. Dishuck[4], Richard Durbin[19,20], Tatiana Dvorkina[3], Ian T. Fiddes[21], Giulio Formenti[22,23], Robert S. Fulton[24], Arkarachai Fungtammasan[18], Erik Garrison[11,25], Patrick G.S. Grady[10], Tina A. Graves-Lindsay[26], Ira M. Hall[27], Nancy F. Hansen[28], Gabrielle A. Hartley[10], Marina Haukness[11], Kerstin Howe[19], Michael W. Hunkapiller[29], Chirag Jain[1,30], Miten Jain[11], Erich D. Jarvis[22,23], Peter Kerpedjiev[31], Melanie Kirsche[9], Mikhail Kolmogorov[32], Jonas Korlach[29], Milinn Kremitzki[26], Heng Li[16,17], Valerie V. Maduro[33], Tobias Marschall[34], Ann M. McCartney[1], Jennifer McDaniel[35], Danny E. Miller[4,36], James C. Mullikin[14,28],

Eugene W. Myers[37], Nathan D. Olson[35], Benedict Paten[11], Paul Peluso[29], Pavel A. Pevzner[32], David Porubsky[4], Tamara Potapova[13], Evgeny I. Rogaev[6,7,38,39], Jeffrey A. Rosenfeld[40], Steven L. Salzberg[9,41], Valerie A. Schneider[42], Fritz J. Sedlazeck[43], Kishwar Shafin[11], Colin J. Shew[44], Alaina Shumate[41], Ying Sims[19], Arian F. A. Smit[45], Daniela C. Soto[44], Ivan Sović[29,46], Jessica M. Storer[45], Aaron Streets[5,47], Beth A. Sullivan[48], Françoise Thibaud-Nissen[42], James Torrance[19], Justin Wagner[35], Brian P. Walenz[1], Aaron Wenger[29], Jonathan M. D. Wood[19], Chunlin Xiao[42], Stephanie M. Yan[49], Alice C. Young[14], Samantha Zarate[9], Urvashi Surti[50], Rajiv C. McCoy[49], Megan Y. Dennis[44], Ivan A. Alexandrov[3,7,51], Jennifer L. Gerton[13,52], Rachel J. O'Neill[10], Winston Timp[8,41], Justin M. Zook[35], Michael C. Schatz[9,49], Evan E. Eichler[4,23,*], Karen H. Miga[11,53,*], Adam M. Phillippy[1,*]

Affiliations

[1]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health; Bethesda, MD USA

[2]Graduate Program in Bioinformatics and Systems Biology, University of California, San Diego; La Jolla, CA, USA

[3]Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, Saint Petersburg State University; Saint Petersburg, Russia

[4]Department of Genome Sciences, University of Washington School of Medicine; Seattle, WA, USA

[5]Department of Bioengineering, University of California, Berkeley; Berkeley, CA, USA

[6]Sirius University of Science and Technology; Sochi, Russia

[7]Vavilov Institute of General Genetics; Moscow, Russia

[8]Department of Molecular Biology and Genetics, Johns Hopkins University; Baltimore, MD, USA

[9]Department of Computer Science, Johns Hopkins University; Baltimore, MD, USA

[10]Institute for Systems Genomics and Department of Molecular and Cell Biology, University of Connecticut; Storrs, CT, USA

[11]UC Santa Cruz Genomics Institute, University of California, Santa Cruz; Santa Cruz, CA, USA

[12]University of Geneva Medical School; Geneva, Switzerland

[13]Stowers Institute for Medical Research; Kansas City, MO, USA

[14]NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health; Bethesda, MD, USA

[15]Department of Molecular and Cell Biology, University of California, Berkeley; Berkeley, CA, USA

[16]Department of Data Sciences, Dana-Farber Cancer Institute; Boston, MA

[17]Department of Biomedical Informatics, Harvard Medical School; Boston, MA

[18]DNAnexus; Mountain View, CA, USA

[19]Wellcome Sanger Institute; Cambridge, UK

[20]Department of Genetics, University of Cambridge; Cambridge, UK

[21]Inscripta; Boulder, CO, USA

[22]Laboratory of Neurogenetics of Language and The Vertebrate Genome Lab, The Rockefeller University; New York, NY, USA

[23]Howard Hughes Medical Institute; Chevy Chase, MD, USA

[24]Department of Genetics, Washington University School of Medicine; St. Louis, MO, USA

[25]University of Tennessee Health Science Center; Memphis, TN, USA

[26]McDonnell Genome Institute, Washington University in St. Louis; St. Louis, MO, USA

[27]Department of Genetics, Yale University School of Medicine; New Haven, CT, USA

[28]Comparative Genomics Analysis Unit, Cancer Genetics and Comparative Genomics Branch, National Human Genome Research Institute, National Institutes of Health; Bethesda, MD, USA

[29]Pacific Biosciences; Menlo Park, CA, USA

[30]Department of Computational and Data Sciences, Indian Institute of Science; Bangalore KA, India

[31]Reservoir Genomics LLC; Oakland, CA

[32]Department of Computer Science and Engineering, University of California, San Diego; San Diego, CA, USA

[33]Undiagnosed Diseases Program, National Human Genome Research Institute, National Institutes of Health; Bethesda, MD, USA

[34]Heinrich Heine University Düsseldorf, Medical Faculty, Institute for Medical Biometry and Bioinformatics; Düsseldorf, Germany

[35]Biosystems and Biomaterials Division, National Institute of Standards and Technology; Gaithersburg, MD, USA

[36]Department of Pediatrics, Division of Genetic Medicine, University of Washington and Seattle Children’s Hospital; Seattle, WA, USA

[37]Max-Planck Institute of Molecular Cell Biology and Genetics; Dresden, Germany

[38]Department of Psychiatry, University of Massachusetts Medical School; Worcester, MA, USA

[39]Faculty of Biology, Lomonosov Moscow State University; Moscow, Russia

[40]Cancer Institute of New Jersey; New Brunswick, NJ, USA

[41]Department of Biomedical Engineering, Johns Hopkins University; Baltimore, MD, USA

[42]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health; Bethesda, MD, USA

[43]Human Genome Sequencing Center, Baylor College of Medicine; Houston TX, USA

[44]Genome Center, MIND Institute, Department of Biochemistry and Molecular Medicine, University of California, Davis; CA, USA

[45]Institute for Systems Biology; Seattle, WA, USA

[46]Digital BioLogic d.o.o.; Ivanić-Grad, Croatia

[47]Chan Zuckerberg Biohub; San Francisco, CA, USA

[48]Department of Molecular Genetics and Microbiology, Duke University School of Medicine; Durham, NC, USA

[49]Department of Biology, Johns Hopkins University; Baltimore, MD, USA

[50]Department of Pathology, University of Pittsburgh; Pittsburgh, PA, USA

[51]Research Center of Biotechnology of the Russian Academy of Sciences; Moscow, Russia

[52]Department of Biochemistry and Molecular Biology, University of Kansas Medical School; Kansas City, MO, USA

[53]Department of Biomolecular Engineering, University of California Santa Cruz, CA, USA

Acknowledgements:

We would like to thank M. Akeson, A. Carroll, P.-C. Chang, A. Delcher, M. Nattestad, and M. Pop for discussions on sequencing, assembly, and analysis; AnVIL, Amazon Web Services, DNAnexus, UW Genome Sciences IT Group, and the UConn Computational Biology Core for computational support; the NIH Intramural Sequencing Center, the UConn Center for Genome Innovation, and the Stowers Imaging Facility for experimental support. This work utilized the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov). Certain commercial equipment, instruments, or materials are identified to specify adequately experimental conditions or reported results. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the equipment, instruments, or materials identified are necessarily the best available for the purpose.

Funding:

Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health (AMP, ACY, AMM, MR, AR, BPW, GGB, JCM, NFH, SK, SN, SYB); NIH U01HG010971 (EEE, HL, KHM, MK, RSF, TAGL); NIH R01HG002385, R01HG010169 (EEE); NIH R01HG009190 (AG, WT); NIH R01HG010485, U01HG010961 (BP, EG, KS); NIH U41HG010972 (IMH, BP, EG, KS); NSF 1627442, 1732253, 1758800, NIH U24HG006620, U01CA253481, R24DK106766 (MCS); NIH U24HG010263 (SZ, MCS); Mark Foundation for Cancer Research 19-033-ASP (SA, MCS); NIH R01HG006677 (ASh, SLS); NIH U24HG009081 (RSF, TAGL); Intramural funding at the National Institute of Standards and Technology (JM, JMZ, JW, NDO); St. Petersburg State University grant 73023573 (AM, IAA, TD); NIH R01HG002939 (AFAS, ITF, JMS); NIH

R01GM124041, R01GM129263, R21CA238758 (BAS); Intramural Research Program of the National Library of Medicine, National Institutes of Health (CX, FTN, VAS); NIH F31HG011205 (CJS); Damon Runyon Postdoctoral Fellowship and PEW Latin American Fellowship (GVC); Fulbright Fellowship (DCS); HHMI (EDJ, GF); NIH R01AG054712 (EIR); NIH UM1HG008898 (FJS); NIH R01GM123312, NSF 1613806 (GAH, PGSG, SJH, RJO); NIH R21CA240199, NSF 643825, Connecticut Innovations 20190200 (RJO); NIH F32GM134558 (GAL); NIH R01HG010040 (HL); St. Petersburg State University grant 73023573 (IAA); Wellcome WT206194 (JT, JMDW, KH, WC, YS); Wellcome WT207492 (RD); Stowers Institute for Medical Research (JLG); NIH R01HG011274 (KHM); Ministry of Science and Higher Education of the RF 075-10-2020-116 / 13.1902.21.0023 (LU); LU is supported by Sirius University; RSF 19-75-30039 Analysis of genomic repeats (IAA); NIH U41HG007234 (MD); NIH DP2MH119424 (MYD); HHMI Hanna H. Gray Fellowship (NA); NIH R35GM133747 (RCM); Childcare Foundation, Swiss National Science Foundation, ERC 249968 (SEA); German Federal Ministry for Research and Education 031L0184A (TM); Chan Zuckerberg Biohub Investigator Award (ASt); Common Fund, Office of the Director, National Institutes of Health (VVM); Max Planck Society (EWM); EEE and EDJ are investigators of the Howard Hughes Medical Institute.

Competing interests:

AF and CSC are employees of DNAnexus; IS, JK, MWH, PP, and AW are employees of Pacific Biosciences; SA is an employee and stock holder of Oxford Nanopore Technologies; EEE is an SAB member of Variant Bio; KHM is an SAB member of Centaura; PK owns and receives income from Reservoir Genomics LLC; FJS has received travel funds to speak at events hosted by Pacific Biosciences; SK, DEM, FJS, and KHM have received travel funds to speak at events hosted by Oxford Nanopore Technologies. WT has licensed two patents to Oxford Nanopore Technologies (US 8748091 and 8394584).

References and Notes:

1. Schneider VA et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 27, 849–864 (2017). [PubMed: 28396521]
2. International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. Nature. 409, 860–921 (2001). [PubMed: 11237011]
3. Venter JC et al. The sequence of the human genome. Science. 291, 1304–1351 (2001). [PubMed: 11181995]
4. Myers EW et al. A whole-genome assembly of Drosophila. Science. 287, 2196–2204 (2000). [PubMed: 10731133]
5. Eichler EE, Clark RA, She X, An assessment of the sequence gaps: unfinished business in a finished human genome. Nat. Rev. Genet 5, 345–354 (2004). [PubMed: 15143317]
6. Miga KH et al. Centromere reference models for human chromosomes X and Y satellite arrays. Genome Res 24, 697–707 (2014). [PubMed: 24501022]
7. Gupta M, Dhanasekaran AR, Gardiner KJ, Mouse models of Down syndrome: gene content and consequences. Mamm. Genome. 27, 538–555 (2016). [PubMed: 27538963]
8. Chaisson MJP et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature. 517, 608–611 (2015). [PubMed: 25383537]
9. International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. Nature. 431, 931–945 (2004). [PubMed: 15496913]
10. Eid J et al. Real-time DNA sequencing from single polymerase molecules. Science. 323, 133–138 (2009). [PubMed: 19023044]
11. Berlin K et al. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat. Biotechnol 33, 623–630 (2015). [PubMed: 26006009]
12. Jain M et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol 36, 338–345 (2018). [PubMed: 29431738]
13. Jain M et al. Linear assembly of a human centromere on the Y chromosome. Nat. Biotechnol 36, 321–323 (2018). [PubMed: 29553574]
14. Miga KH et al. Telomere-to-telomere assembly of a complete human X chromosome. Nature. 585, 79–84 (2020). [PubMed: 32663838]
15. Wenger AM et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol 37, 1155–1162 (2019). [PubMed: 31406327]

16. Nurk S et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res. 30, 1291–1305 (2020). [PubMed: 32801147]
17. Huddleston J et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Research. 27 (2017), pp. 677–685. [PubMed: 27895111]
18. Eichler EE, Surti U, Ophoff R, Proposal for Construction a Human Haploid BAC library from Hydatidiform Mole Source Material (2002), (available at https://www.genome.gov/Pages/Research/Sequencing/BACLibrary/HydatidiformMoleBAC021203.pdf).
19. Steinberg KM et al. Single haplotype assembly of the human genome from a hydatidiform mole. Genome Res. 24, 2066–2076 (2014). [PubMed: 25373144]
20. Vollger MR et al. Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. Ann. Hum. Genet (2019), doi:10.1111/ahg.12364.
21. Logsdon GA et al. The structure, function and evolution of a complete human chromosome 8. Nature. 593, 101–107 (2021). [PubMed: 33828295]
22. Fan J-B et al. Paternal origins of complete hydatidiform moles proven by whole genome single-nucleotide polymorphism haplotyping. Genomics. 79, 58–62 (2002). [PubMed: 11827458]
23. See supplementary materials.
24. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. Nature. 526, 68–74 (2015). [PubMed: 26432245]
25. Aganezov S et al. A complete reference genome improves analysis of human genetic variation. Science 375, eabl3533 (2022). doi:10.1126/science.abl3533
26. Myers EW, The fragment assembly string graph. Bioinformatics. 21 Suppl 2, ii79–85 (2005). [PubMed: 16204131]
27. Li H, Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics. 32, 2103–2110 (2016). [PubMed: 27153593]
28. Rautiainen M, Marschall T, GraphAligner: rapid and versatile sequence-to-graph alignment. Genome Biol. 21, 253 (2020). [PubMed: 32972461]
29. Gershman A et al. Epigenetic patterns in a complete human genome. Science375, eabj5089 (2022). doi:10.1126/science.abj5089
30. Altemose N et al. Complete genomic and epigenetic maps of human centromeres. Science 376, eabl4178 (2022). doi:10.1126/science.abl4178. [PubMed: 35357911]
31. Mc Cartney AM et al. Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. Nat. Methods (2022), doi:10.1038/s41592-022-01440-3
32. Flynn JM, Long M, Wing RA, Clark AG, Evolutionary Dynamics of Abundant 7-bp Satellites in the Genome of Drosophila virilis. Mol. Biol. Evol 37, 1362–1375 (2020). [PubMed: 31960929]
33. Guiblet WM et al. Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate. Genome Res. 28, 1767–1778 (2018). [PubMed: 30401733]
34. Mikheenko A, Bzikadze AV, Gurevich A, Miga KH, Pevzner PA, TandemTools: mapping long reads and assessing/improving assembly quality in extra-long tandem repeats. Bioinformatics. 36, i75–i83 (2020). [PubMed: 32657355]
35. Vollger MR et al. Long-read sequence and assembly of segmental duplications. Nat. Methods. 16, 88–94 (2019). [PubMed: 30559433]
36. Parks MM et al. Variant ribosomal RNA alleles are conserved and exhibit tissue-specific expression. Sci Adv. 4, eaao0665 (2018). [PubMed: 29503865]
37. Nelson JO, Watase GJ, Warsinger-Pepe N, Yamashita YM, Mechanisms of rDNA Copy Number Maintenance. Trends Genet. 35, 734–742 (2019). [PubMed: 31395390]
38. Rautiainen M, Marschall T, MBG: Minimizer-based Sparse de Bruijn Graph Construction. Bioinformatics (2021), doi:10.1093/bioinformatics/btab004.
39. Fiddes IT et al. Comparative Annotation Toolkit (CAT)-simultaneous clade and personal genome annotation. Genome Res. 28, 1029–1038 (2018). [PubMed: 29884752]
40. Shumate A, Salzberg SL, Liftoff: accurate mapping of gene annotations. Bioinformatics (2020), doi:10.1093/bioinformatics/btaa1016.
41. Frankish A et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 47, D766–D773 (2019). [PubMed: 30357393]

42. Vollger MR et al. Segmental duplications and their variation in a complete human genome. Science 375, eabj6965 (2022), doi:10.1126/science.abj6965
43. Mallick S et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 538, 201–206 (2016). [PubMed: 27654912]
44. Lindström MS et al. Nucleolus as an emerging hub in maintenance of genome stability and cancer pathogenesis. Oncogene. 37, 2351–2366 (2018). [PubMed: 29429989]
45. Lyle R et al. Islands of euchromatin-like sequence and expressed polymorphic sequences within the short arm of human chromosome 21. Genome Res. 17, 1690–1696 (2007). [PubMed: 17895424]
46. Floutsakou I et al. The shared genomic architecture of human nucleolar organizer regions. Genome Res. 23, 2003–2012 (2013). [PubMed: 23990606]
47. Kim J-H et al. Variation in human chromosome 21 ribosomal RNA genes characterized by TAR cloning and long-read sequencing. Nucleic Acids Res. 46, 6712–6725 (2018). [PubMed: 29788454]
48. Caburet S et al. Human ribosomal RNA gene arrays display a broad range of palindromic structures. Genome Res. 15, 1079–1085 (2005). [PubMed: 16024823]
49. Hoyt SJ et al. From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. Science 376, eabk3112 (2022). doi:10.1126/science.abk3112 [PubMed: 35357925]
50. van Sluis M et al. Human NORs, comprising rDNA arrays and functionally conserved distal elements, are located within dynamic chromosomal regions. Genes Dev. 33, 1688–1701 (2019). [PubMed: 31727772]
51. Sullivan BA, Jenkins LS, Karson EM, Leana-Cox J, Schwartz S, Evidence for structural heterogeneity from molecular cytogenetic analysis of dicentric Robertsonian translocations. Am. J. Hum. Genet 59, 167–175 (1996). [PubMed: 8659523]
52. Greig GM, Warburton PE, Willard HF, Organization and evolution of an alpha satellite DNA subset shared by human chromosomes 13 and 21. J. Mol. Evol 37, 464–475 (1993). [PubMed: 8283478]
53. Jørgensen AL, Kølvraa S, Jones C, Bak AL, A subfamily of alphoid repetitive DNA shared by the NOR-bearing human chromosomes 14 and 22. Genomics. 3, 100–109 (1988). [PubMed: 3224978]
54. Kent WJ et al. The human genome browser at UCSC. Genome Res. 12, 996–1006 (2002). [PubMed: 12045153]
55. Wijmenga C et al. Chromosome 4q DNA rearrangements associated with facioscapulohumeral muscular dystrophy. Nat. Genet 2, 26–30 (1992). [PubMed: 1363881]
56. Lemmers RJLF et al. A unifying genetic model for facioscapulohumeral muscular dystrophy. Science. 329, 1650–1653 (2010). [PubMed: 20724583]
57. Grewal PK, van Geel M, Frants RR, de Jong P, Hewitt JE, Recent amplification of the human FRG1 gene during primate evolution. Gene. 227, 79–88 (1999). [PubMed: 9931447]
58. van Deutekom JC et al. Identification of the first gene (FRG1) from the FSHD region on human chromosome 4q35. Hum. Mol. Genet 5, 581–590 (1996). [PubMed: 8733123]
59. Miga KH, Wang T, The Need for a Human Pangenome Reference Sequence. Annu. Rev. Genomics Hum. Genet (2021), doi:10.1146/annurev-genom-120120-081921.
60. Wick RR, Schultz MB, Zobel J, Holt KE, Bandage: interactive visualization of de novo genome assemblies. Bioinformatics. 31, 3350–3352 (2015). [PubMed: 26099265]
61. Nurk S, Koren S, Rhie A, Rautiainen M, T2T-CHM13 supplemental code and data (2021; 10.5281/zenodo.5598253).
62. Maples BK, Gravel S, Kenny EE, Bustamante CD, RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. Am. J. Hum. Genet 93, 278–288 (2013). [PubMed: 23910464]
63. Lowy-Gallego E, Fairley S, Zheng-Bradley X, Ruffier M, Clarke L, Flicek P, Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project [version 2; peer review: 2 approved]. Wellcome Open Res. 4 (2019), doi:10.12688/wellcomeopenres.15126.2.
64. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao

H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MMY, Tsui SKW, Xue H, Wong JT-F, Galver LM, Fan J-B, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier J-F, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok P-Y, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui L-C, Mak W, Qiang Song Y, Tam PKH, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PIW, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Vernon Smith A, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Steve Qin Z, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Johnson T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CDM, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Wright Clayton E, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevêque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Ota Wang V, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, The International HapMap Consortium, Genotyping centres: Perlegen Sciences, Baylor College of Medicine and ParAllele BioScience, Beijing Genomics Institute, Broad Institute of Harvard and Massachusetts Institute of Technology, Chinese National Human Genome Center at Beijing, Chinese National Human Genome Center at Shanghai, Chinese University of Hong Kong, Hong Kong University of Science and Technology, Illumina, McGill University and Génome Québec Innovation Centre, University of California at San Francisco and Washington University, University of Hong Kong, University of Tokyo and RIKEN, Wellcome Trust Sanger Institute, Analysis groups: Broad Institute, Cold Spring Harbor Laboratory, Johns Hopkins University School of Medicine, University of Michigan, University of Oxford, W. T. C. for H. G. University of Oxford, RIKEN, US National Institutes of Health, US National Institutes of Health National Center for Biotechnology Information, Community engagement/public consultation and sample collection groups: Beijing Normal University and Beijing Genomics Institute, E. E. I. Health Sciences University of Hokkaido and Shinshu University, Howard University and University of Ibadan, University of Utah, legal and social issues: C. A. of S. S. Ethical, Genetic Interest Group, Kyoto University, Nagasaki University, University of Ibadan School of Medicine, University of Montréal, University of Oklahoma, Vanderbilt University, Wellcome Trust, SNP discovery: Baylor College of Medicine, Washington University, Scientific management: Chinese Academy of Sciences, Genome Canada, Génome Québec, C. Japanese Ministry of Education Sports, Science and Technology, Ministry of Science and Technology of the People's Republic of China, The Human Genetic Resource Administration of China, The SNP Consortium, A second generation human haplotype map of over 3.1 million SNPs. Nature. 449, 851–861 (2007). [PubMed: 17943122]

65. Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier L, Neale B, MacArthur D, A synthetic-diploid benchmark for accurate variant-calling evaluation. Nat. Methods. 15, 595–597 (2018). [PubMed: 30013044]
66. Chen L, Wolf AB, Fu W, Li L, Akey JM, Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals. Cell. 180, 677–687.e16 (2020). [PubMed: 32004458]
67. Prüfer K, de Filippo C, Grote S, Mafessoni F, Korlević P, Hajdinjak M, Vernot B, Skov L, Hsieh P, Peyrégne S, Reher D, Hopfe C, Nagel S, Maricic T, Fu Q, Theunert C, Rogers R, Skoglund P, Chintalapati M, Dannemann M, Nelson BJ, Key FM, Rudan P, Kućan Ž, Gušić I, Golovanova LV, Doronichev VB, Patterson N, Reich D, Eichler EE, Slatkin M, Schierup MH, Andrés AM, Kelso

J, Meyer M, Pääbo S, A high-coverage Neandertal genome from Vindija Cave in Croatia. Science. 358, 655 (2017). [PubMed: 28982794]

68. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, Fairley S, Runnels A, Winterkorn L, Lowy-Gallego E, Flicek P, Germer S, Brand H, Hall IM, Talkowski ME, Narzisi G, Zody MC, High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. bioRxiv (2021), doi:10.1101/2021.02.06.430068.
69. Logsdon GA, Vollger MR, Eichler EE, Long-read human genome sequencing and its applications. Nat. Rev. Genet 21, 597–614 (2020). [PubMed: 32504078]
70. Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, Sutton G, Aggressive assembly of pyrosequencing reads with mates. Bioinformatics. 24 (2008), pp. 2818–2824. [PubMed: 18952627]
71. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM, Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27, 722–736 (2017). [PubMed: 28298431]
72. Dawson ET, Durbin R, GFAKluge: A C++ library and command line utilities for the Graphical Fragment Assembly formats. J. Open Source Softw 4, 1083 (2019). [PubMed: 31535074]
73. Gonnella G, Kurtz S, GfaPy: a flexible and extensible software library for handling sequence graphs in Python. Bioinformatics. 33, 3094–3095 (2017). [PubMed: 28645150]
74. Li H, Feng X, Chu C, The design and construction of reference pangenome graphs with minigraph. Genome Biol. 21, 265 (2020). [PubMed: 33066802]
75. Peng Y, Leung HCM, Yiu SM, Chin FYL, IDBA – A Practical Iterative de Bruijn Graph De Novo Assembler. Res. Comput. Mol. Biol, 426–440 (2010).
76. Jain C, Koren S, Dilthey A, Phillippy AM, Aluru S, A fast adaptive algorithm for computing whole-genome homology maps. Bioinformatics. 34, i748–i756 (2018). [PubMed: 30423094]
77. Šípek A Jr, Mihalová R, Panczak A, Hrčková L, Janashia M, Kaspříková N, Kohoutová M, Heterochromatin variants in human karyotypes: a possible association with reproductive failure. Reprod. Biomed. Online. 29, 245–250 (2014). [PubMed: 24928354]
78. Jain C, Rhie A, Zhang H, Chu C, Walenz BP, Koren S, Phillippy AM, Weighted minimizer sampling improves long read mapping. Bioinformatics. 36, i111–i118 (2020). [PubMed: 32657365]
79. Jain C, Rhie A, Hansen N, Koren S, Phillippy AM, Long-read mapping to repetitive reference sequences using Winnowmap2. Nat. Methods (2022), doi:10.1038/s41592-022-01457-8
80. Bzikadze AV, Pevzner PA, Automated assembly of centromeres from ultra-long error-prone reads. Nat. Biotechnol 38, 1309–1316 (2020). [PubMed: 32665660]
81. Rhie A, Walenz BP, Koren S, Phillippy AM, Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 21, 245 (2020). [PubMed: 32928274]
82. Li H, Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 34, 3094–3100 (2018). [PubMed: 29750242]
83. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, Lee C, Ko BJ, Chaisson M, Gedman GL, Cantin LJ, Thibaud-Nissen F, Haggerty L, Bista I, Smith M, Haase B, Mountcastle J, Winkler S, Paez S, Howard J, Vernes SC, Lama TM, Grutzner F, Warren WC, Balakrishnan CN, Burt D, George JM, Biegler MT, Iorns D, Digby A, Eason D, Robertson B, Edwards T, Wilkinson M, Turner G, Meyer A, Kautt AF, Franchini P, Detrich HW, Svardal H, Wagner M, Naylor GJP, Pippel M, Malinsky M, Mooney M, Simbirsky M, Hannigan BT, Pesout T, Houck M, Misuraca A, Kingan SB, Hall R, Kronenberg Z, Sović I, Dunn C, Ning Z, Hastie A, Lee J, Selvaraj S, Green RE, Putnam NH, Gut I, Ghurye J, Garrison E, Sims Y, Collins J, Pelan S, Torrance J, Tracey A, Wood J, Dagnew RE, Guan D, London SE, Clayton DF, Mello CV, Friedrich SR, Lovell PV, Osipova E, Al-Ajli FO, Secomandi S, Kim H, Theofanopoulou C, Hiller M, Zhou Y, Harris RS, Makova KD, Medvedev P, Hoffman J, Masterson P, Clark K, Martin F, Howe K, Flicek P, Walenz BP, Kwak W, Clawson H, Diekhans M, Nassar L, Paten B, Kraus RHS, Crawford AJ, Gilbert MTP, Zhang G, Venkatesh B, Murphy RW, Koepfli K-P, Shapiro B, Johnson WE, Di Palma F, Marques-Bonet T, Teeling EC, Warnow T, Graves JM, Ryder OA, Haussler D, O'Brien SJ, Korlach J, Lewin HA, Howe K, Myers EW,

Durbin R, Phillippy AM, Jarvis ED, Towards complete and error-free genome assemblies of all vertebrate species. Nature. 592, 737–746 (2021). [PubMed: 33911273]

84. Li H, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv Prepr. ArXiv13033997 (2013).
85. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, Gross SS, Dorfman L, McLean CY, DePristo MA, A universal SNP and small-indel variant caller using deep neural networks. Nat. Biotechnol 36, 983–987 (2018). [PubMed: 30247488]
86. Shafin K, Pesout T, Chang P-C, Nattestad M, Kolesnikov A, Goel S, Baid G, Eizenga JM, Miga KH, Carnevali P, Jain M, Carroll A, Paten B, Haplotype-aware variant calling enables high accuracy in nanopore long-reads using deep neural networks. bioRxiv (2021), doi:10.1101/2021.03.04.433952.
87. Formenti G, Rhie A, Walenz BP, Thibaud-Nissen F, Shafin K, Koren S, Myers EW, Jarvis ED, Phillippy AM, Merfin: improved variant filtering and polishing via k-mer validation. bioRxiv (2021), doi:10.1101/2021.07.16.452324.
88. Zarate S, Carroll A, Krashenina O, Sedlazeck FJ, Jun G, Salerno W, Boerwinkle E, Gibbs R, Parliament2: Fast Structural Variant Calling Using Optimized Combinations of Callers. bioRxiv, 424267 (2018).
89. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC, Accurate detection of complex structural variations using single-molecule sequencing. Nat. Methods. 15, 461–468 (2018). [PubMed: 29713083]
90. Kirsche M, Prabhu G, Sherman R, Ni B, Aganezov S, Schatz MC, Jasmine: Population-scale structural variant comparison and analysis. bioRxiv (2021), doi:10.1101/2021.05.27.445886.
91. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP, Integrative genomics viewer. Nat. Biotechnol 29, 24–26 (2011). [PubMed: 21221095]
92. Schmutz J, Wheeler J, Grimwood J, Dickson M, Yang J, Caoile C, Bajorek E, Black S, Chan YM, Denys M, Escobar J, Flowers D, Fotopulos D, Garcia C, Gomez M, Gonzales E, Haydu L, Lopez F, Ramirez L, Retterer J, Rodriguez A, Rogers S, Salazar A, Tsai M, Myers RM, Quality assessment of the human genome sequence. Nature. 429 (2004), pp. 365–8. [PubMed: 15164052]
93. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, Henaff E, McIntyre ABR, Chandramohan D, Chen F, Jaeger E, Moshrefi A, Pham K, Stedman W, Liang T, Saghbini M, Dzakula Z, Hastie A, Cao H, Deikus G, Schadt E, Sebra R, Bashir A, Truty RM, Chang CC, Gulbahce N, Zhao K, Ghosh S, Hyland F, Fu Y, Chaisson M, Xiao C, Trow J, Sherry ST, Zaranek AW, Ball M, Bobe J, Estep P, Church GM, Marks P, Kyriazopoulou-Panagiotopoulou S, Zheng GXY, Schnall-Levin M, Ordonez HS, Mudivarti PA, Giorda K, Sheng Y, Rypdal KB, Salit M, Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci. Data. 3, 1–26 (2016).
94. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, Sedlazeck FJ, Marschall T, Mayes S, Costa V, Zook JM, Liu KJ, Kilburn D, Sorensen M, Munson KM, Vollger MR, Monlong J, Garrison E, Eichler EE, Salama S, Haussler D, Green RE, Akeson M, Phillippy A, Miga KH, Carnevali P, Jain M, Paten B, Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Nat. Biotechnol (2020), doi:10.1038/s41587-020-0503-6.
95. Kolmogorov M, Yuan J, Lin Y, Pevzner PA, Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol 37, 540–546 (2019). [PubMed: 30936562]
96. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy AM, De novo assembly of haplotype-resolved genomes with trio binning. Nat. Biotechnol (2018), doi:10.1038/nbt.4277.
97. Sullivan LL, Boivin CD, Mravinac B, Song IY, Sullivan BA, 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 (2011). [PubMed: 21484447]
98. Mravinac B, Sullivan LL, Reeves JW, Yan CM, Kopf KS, Farr CJ, Schueler MG, Sullivan BA, Histone Modifications within the Human X Centromere Region. PLOS ONE. 4, e6602 (2009). [PubMed: 19672304]

99. Phillippy AM, Schatz MC, Pop M, Genome assembly forensics: finding the elusive mis-assembly. Genome Biol. 9 (2008), pp. R55–R55. [PubMed: 18341692]
100. Tischler G, Leonard S, biobambam: tools for read pair collation based algorithms on BAM files. Source Code Biol. Med 9, 13 (2014).
101. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome S Project Data Processing, The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25 (2009), pp. 2078–9. [PubMed: 19505943]
102. Falconer E, Hills M, Naumann U, Poon SSS, Chavez EA, Sanders AD, Zhao Y, Hirst M, Lansdorp PM, DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. Nat. Methods. 9, 1107–1112 (2012). [PubMed: 23042453]
103. Sanders AD, Falconer E, Hills M, Spierings DCJ, Lansdorp PM, Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. Nat. Protoc 12, 1151–1176 (2017). [PubMed: 28492527]
104. Porubsky D, Sanders AD, Taudt A, Colomé-Tatché M, Lansdorp PM, Guryev V, breakpointR: an R/Bioconductor package to localize strand state changes in Strand-seq data. Bioinformatics. 36, 1260–1261 (2020). [PubMed: 31504176]
105. Porubsky D, Ebert P, Audano PA, Vollger MR, Harvey WT, Marijon P, Ebler J, Munson KM, Sorensen M, Sulovari A, Haukness M, Ghareghani M, Lansdorp PM, Paten B, Devine SE, Sanders AD, Lee C, Chaisson MJP, Korbel JO, Eichler EE, Marschall T, 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). [PubMed: 33288906]
106. Kerpedjiev P, Abdennur N, Lekschas F, McCallum C, Dinkla K, Strobelt H, Luber JM, Ouellette SB, Azhir A, Kumar N, Hwang J, Lee S, Alver BH, Pfister H, Mirny LA, Park PJ, Gehlenborg N, HiGlass: web-based visual exploration and analysis of genome interaction maps. Genome Biol. 19, 125 (2018). [PubMed: 30143029]
107. Krishnakumar R, Sinha A, Bird SW, Jayamohan H, Edwards HS, Schoeniger JS, Patel KD, Branda SS, Bartsch MS, Systematic and stochastic influences on the performance of the MinION nanopore sequencer across a range of nucleotide bias. Sci. Rep 8, 3159 (2018). [PubMed: 29453452]
108. Uralsky LI, Shepelev VA, Alexandrov AA, Yurov YB, Rogaev EI, Alexandrov IA, 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). [PubMed: 30989093]
109. Wheeler TJ, Eddy SR, nhmmer: DNA homology search with profile HMMs. Bioinformatics. 29, 2487–2489 (2013). [PubMed: 23842809]
110. Vaser R, Sović I, Nagarajan N, Šikić M, Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27, 737–746 (2017). [PubMed: 28100585]
111. Gel B, Serra E, karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics. 33, 3088–3090 (2017). [PubMed: 28575171]
112. Harris RS, thesis, Pennsylvania State University, USA (2007).
113. Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D, Cactus: Algorithms for genome multiple sequence alignment. Genome Res. 21, 1512–1528 (2011). [PubMed: 21665927]
114. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M, Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 20, 278 (2019). [PubMed: 31842956]
115. Stanke M, Diekhans M, Baertsch R, Haussler D, Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 24, 637–644 (2008). [PubMed: 18218656]
116. Miller CA, Walker JR, Jensen TL, Hooper WF, Fulton RS, Painter JS, Sekeres MA, Ley TJ, Spencer DH, Goll JB, Walter MJ, Failure to detect mutations in U2AF1 due to changes in the GRCh38 reference sequence. bioRxiv (2021), doi:10.1101/2021.05.07.442430.
117. Pockrandt C, Alzamel M, Iliopoulos CS, Reinert K, GenMap: ultra-fast computation of genome mappability. Bioinformatics. 36, 3687–3692 (2020). [PubMed: 32246826]

118. Katoh K, Standley DM, MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol 30, 772–780 (2013). [PubMed: 23329690]

119. Kumar S, Stecher G, Li M, Knyaz C, Tamura K, MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol. Biol. Evol 35, 1547–1549 (2018). [PubMed: 29722887]

120. Saitou N, Nei M, The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol 4, 406–425 (1987). [PubMed: 3447015]

121. Kimura M, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol 16, 111–120 (1980). [PubMed: 7463489]

122. Stamatakis A, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30, 1312–1313 (2014). [PubMed: 24451623]

123. Nei M, Kumar S, Molecular evolution and phylogenetics (Oxford university press, 2000).

124. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C, Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods. 14, 417–419 (2017). [PubMed: 28263959]

125. Soneson C, Love M, Robinson M, Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 2; peer review: 2 approved]. F1000Research. 4 (2016), doi:10.12688/f1000research.7563.2.

126. Delcher AL, Phillippy A, Carlton J, Salzberg SL, MUMmer: comparative applications Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 30 (2002), pp. 2478–2483. [PubMed: 12034836]

127. Ghareghani M, Porubskỳ D, Sanders AD, Meiers S, Eichler EE, Korbel JO, Marschall T, Strand-seq enables reliable separation of long reads by chromosome via expectation maximization. Bioinformatics. 34, i115–i123 (2018). [PubMed: 29949971]

128. Cheng H, Concepcion GT, Feng X, Zhang H, Li H, Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat. Methods 18, 170–175 (2021). [PubMed: 33526886]

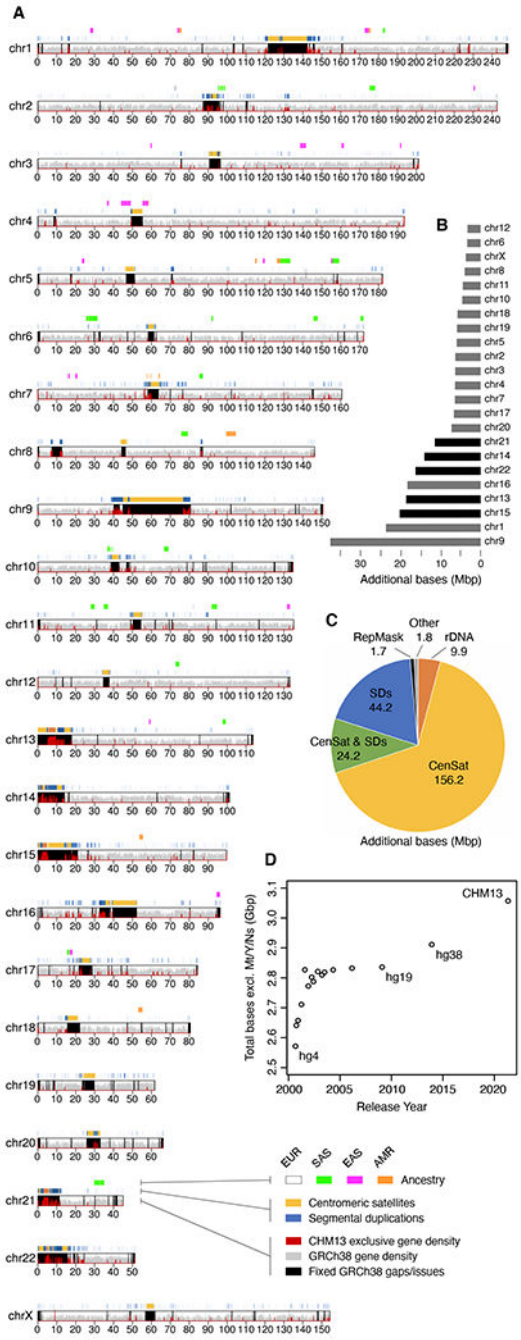

Fig. 1. Summary of the complete T2T-CHM13 human genome assembly.

(**A**) Ideogram of T2T-CHM13v1.1 assembly features. Bottom to top: gaps/issues in GRCh38 fixed by CHM13 overlaid with the density of genes exclusive to CHM13 in red; segmental duplications (SDs) (42) and centromeric satellites (CenSat) (30); and CHM13 ancestry predictions (EUR, European; SAS, South Asian; EAS, East Asian; AMR, Ad Mixed American). (**B**) Additional (non-syntenic) bases in the CHM13 assembly relative to GRCh38 per chromosome, with the acrocentrics highlighted in black, and (**C**) by sequence type (note that the CenSat and SD annotations overlap). (**D**) Total non-gap bases in UCSC reference

genome releases dating back to September 2000 (hg4) and ending with T2T-CHM13 in 2021.

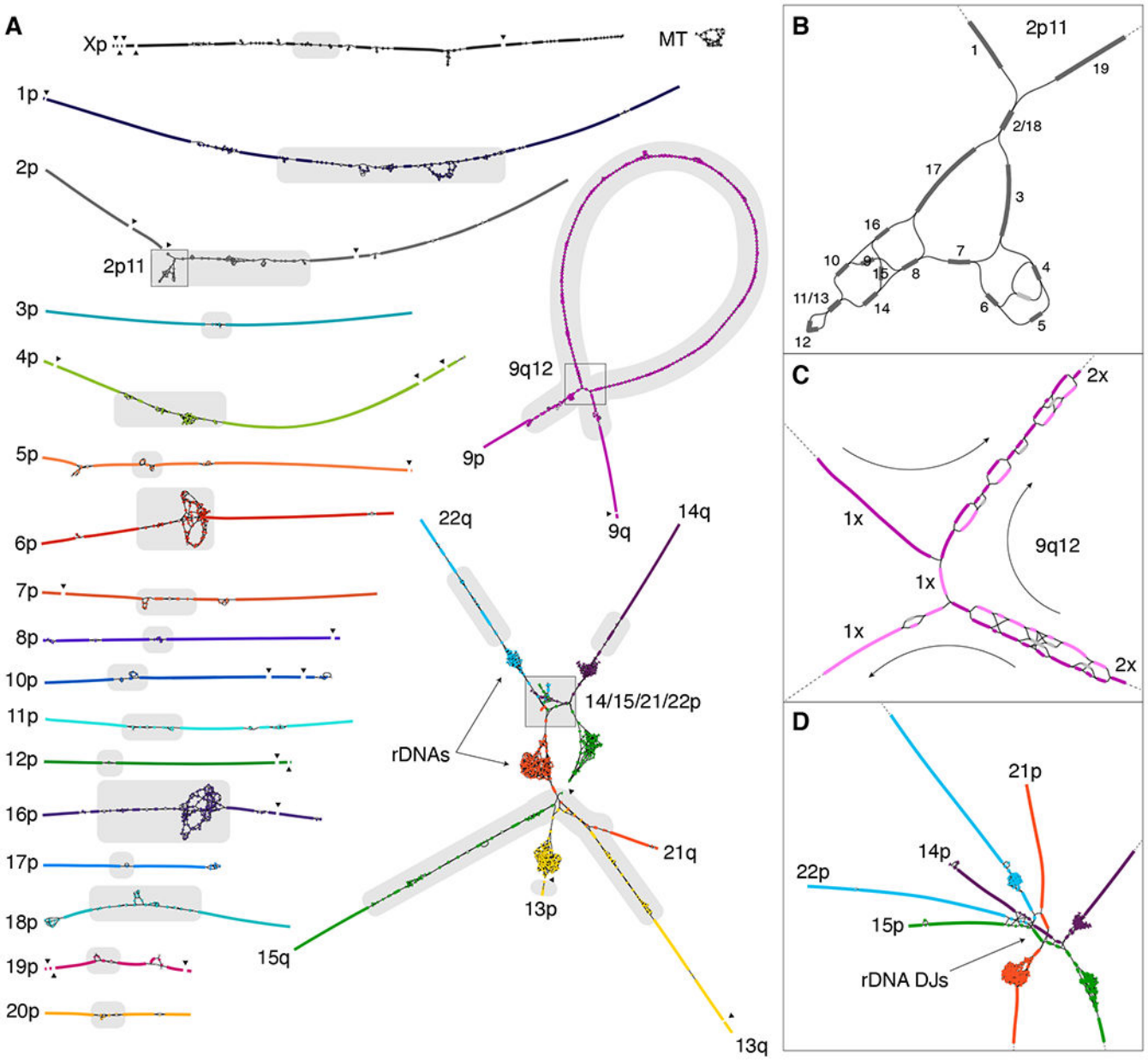

Fig. 2. High-resolution assembly string graph of the CHM13 genome.

(**A**) Bandage (60) visualization, where nodes represent unambiguously assembled sequences scaled by length, and edges correspond to the overlaps between node sequences. Each chromosome is both colored and numbered on the short (p) arm. Long (q) arms are labeled where unclear. The five acrocentric chromosomes (bottom right) are connected due to similarity between their short arms, and the rDNA arrays form five dense tangles due to their high copy number. The graph is partially fragmented due to HiFi coverage dropout surrounding GA-rich sequence (black triangles). Centromeric satellites (30) are the source of most ambiguity in the graph (gray highlights). (**B**) The ONT-assisted graph traversal for the 2p11 locus is given by numerical order. Based on low depth-of-coverage, the unlabeled light gray node represents an artifact or heterozygous variant and was not used. (**C**) The

multi-megabase tandem HSat3 duplication (9qh+) at 9q12 requires two traversals of the large loop structure (the size of the loop is exaggerated because graph edges are of constant size). Nodes used by the first traversal are in dark purple and the second traversal in light purple. Nodes used by both traversals typically have twice the sequencing coverage. (**D**) Enlargement of the distal short arms of the acrocentrics, showing the colored graph walks and edges between highly similar sequences in the distal junctions (DJs) adjacent to the rDNA arrays.

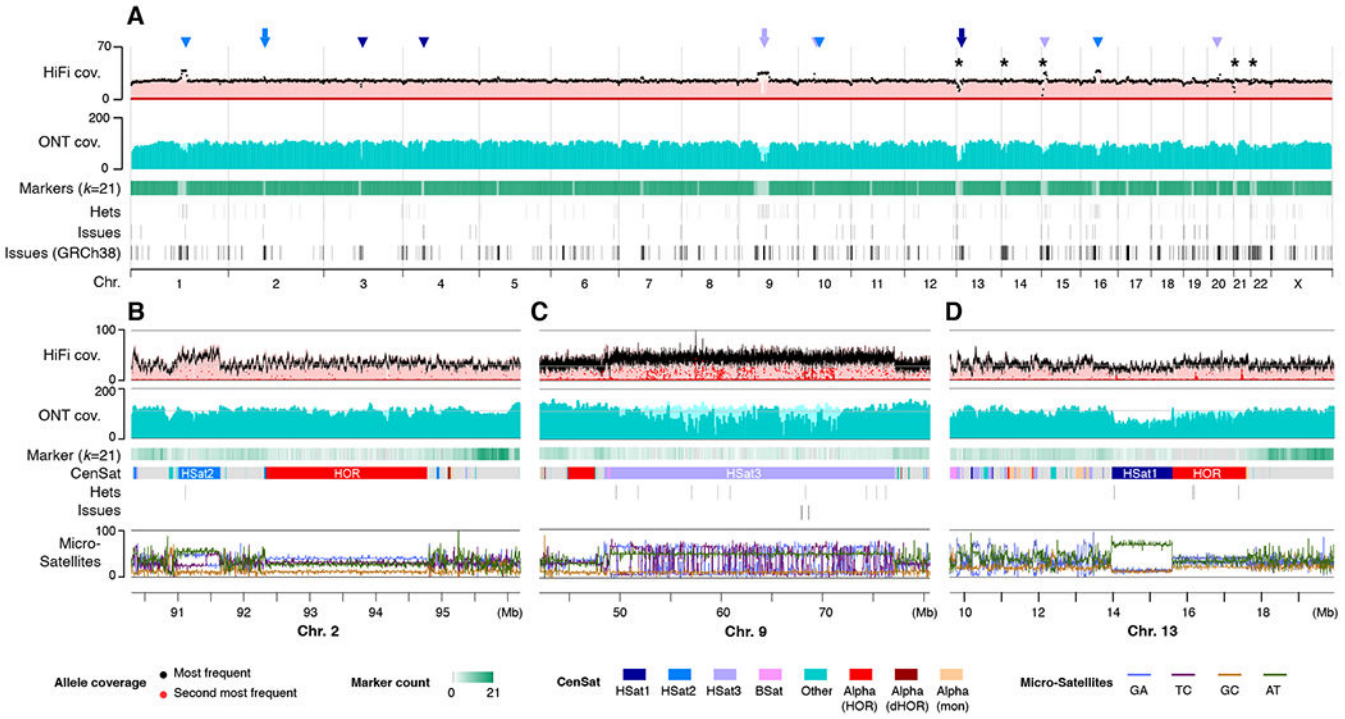

Fig. 3. Sequencing coverage and assembly validation.

(**A**) Uniform whole-genome coverage of mapped HiFi and ONT reads is shown with primary alignments in light shades and marker-assisted alignments overlaid in dark shades. Large HSat arrays (30) are noted by triangles, with inset regions are marked by arrowheads and the location of the rDNA arrays marked with asterisks. Regions with low unique marker frequency (light green) correspond to drops in unique marker density, but are recovered by the lower-confidence primary alignments. Annotated assembly issues are compared for T2T-CHM13 and GRCh38. (**B–D**) Enlargements corresponding to regions of the genome featured in Fig. 2. Uniform coverage changes within certain satellites are reproducible and likely caused by sequencing bias. Identified heterozygous variants and assembly issues are marked below and typically correspond with low coverage of the primary allele (black) and elevated coverage of the secondary allele (red). % microsatellite repeats for every 128 bp window is shown at the bottom.

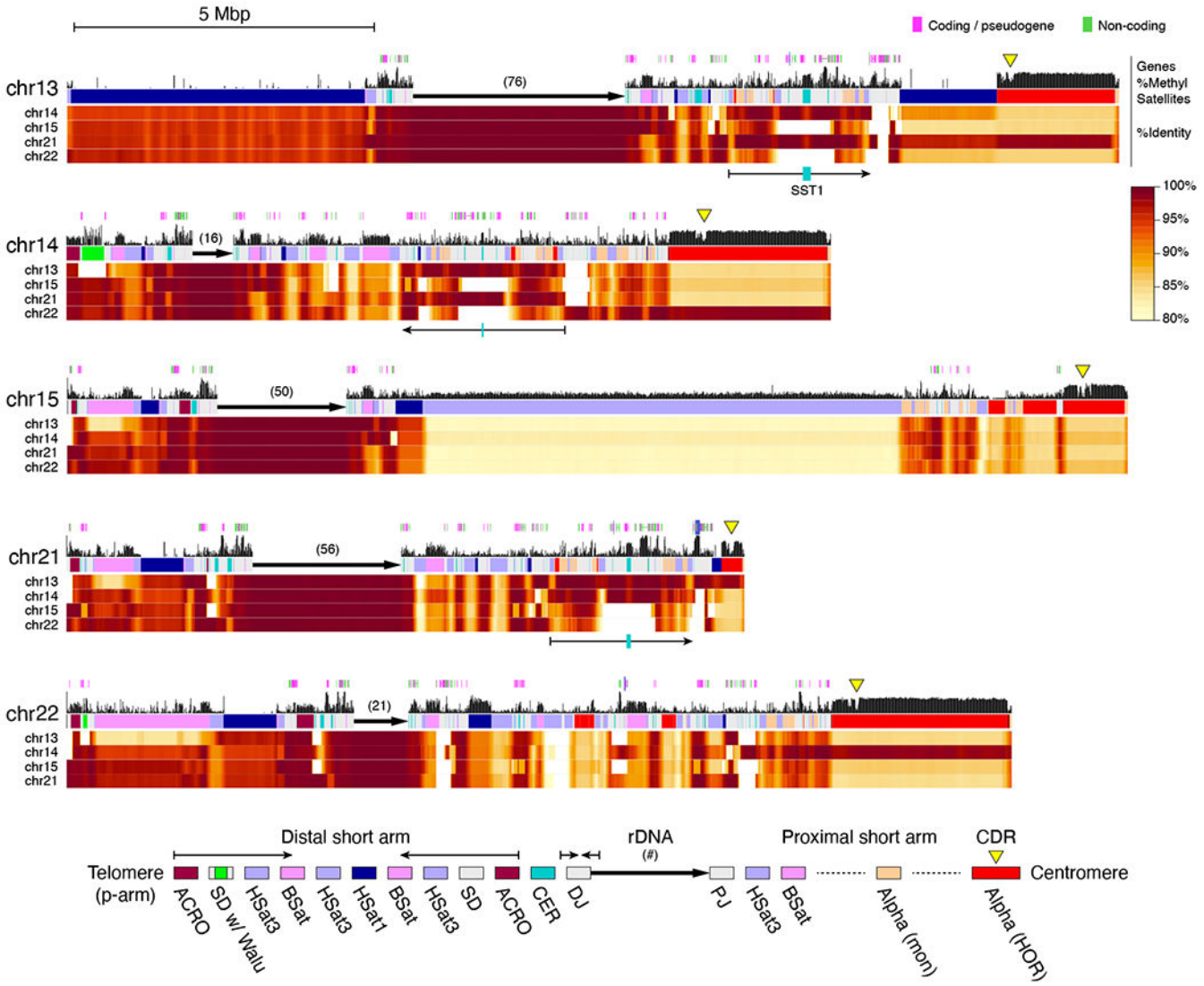

Fig. 4. Short arms of the acrocentric chromosomes.

Each short arm is shown along with annotated genes, percent of methylated CpGs (29), and a color-coded satellite repeat annotation (30). The rDNA arrays are represented by a directional arrow and copy number due to their high self-similarity, which prohibits ONT mapping. Percent identity heatmaps versus the other four arms were computed in 10 kbp windows and smoothed over 100 kbp intervals. Each position shows the maximum identity of that window to any window in the other chromosome. The distal short arms include conserved satellite structure and inverted repeats (thin arrows), while the proximal short arms show a diversity of structures. The proximal short arms of Chromosomes 13, 14, and 21 share a segmentally duplicated core, including small alpha satellite HOR arrays and a central, highly methylated, SST1 array (thin arrows with teal block). Yellow triangles indicate hypomethylated centromeric dip regions (CDRs), marking the sites of kinetochore assembly (29).

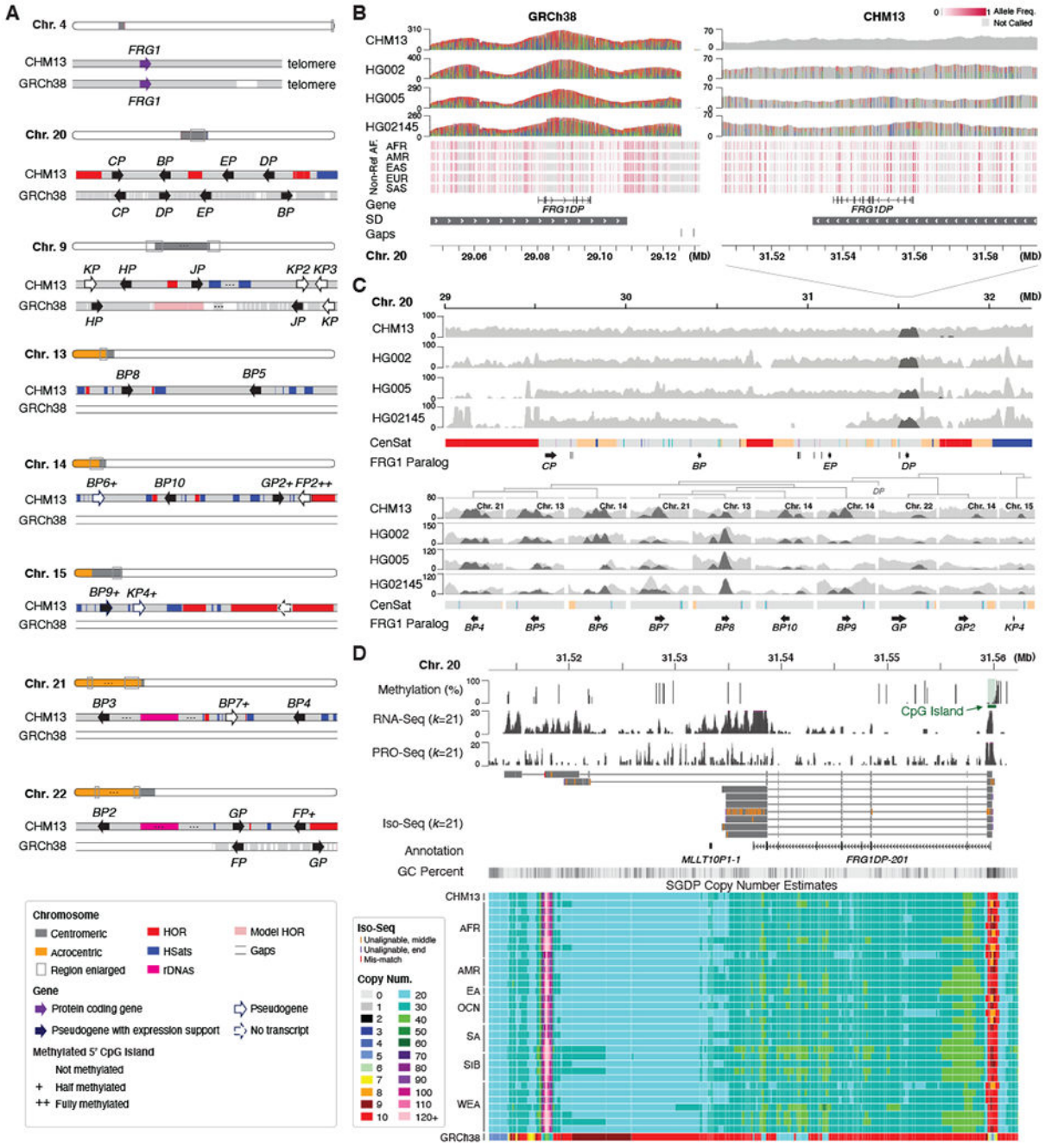

Fig. 5. Resolved FRG1 paralogs.

(**A**) Protein-coding gene *FRG1* and its 23 paralogs in CHM13. Only 9 are found in GRCh38. Genes are drawn larger than their actual size and the “FRG1” prefix is omitted for brevity. All paralogs are found near satellite arrays. Most copies exhibit evidence of expression, including CpG islands present at the 5′ start site with varying degrees of methylation. (**B**) Reference (gray) and variant (colored) allele coverage is shown for four human HiFi samples mapped to the paralog *FRG1DP*. When mapped to GRCh38, the region shows excessive HiFi coverage and variants, indicating that reads from the missing paralogs are

mis-mapped to *FRG1DP* (variants with >80% coverage shown). When mapped to CHM13, HiFi reads show the expected coverage and a typical heterozygous variation pattern for the three non-CHM13 samples (variants >20% coverage shown). These non-reference alleles are also found in other populations from 1KGP ILMN data. (**C**) Mapped HiFi read coverage for other *FRG1* paralogs, with an extended context shown for Chromosome 20. Coverage of HiFi reads that mapped to *FRG1DP* in GRCh38 are highlighted (dark gray), showing the paralogous copies they originate from (*FRG1BP4–10*, *FRG1GP*, *FRG1GP2*, and *FRG1KP4*). Background coverage is variable for some paralogs, suggesting copy number polymorphism in the population. (**D**) Methylation and expression profiles suggest transcription of *FRG1DP* in CHM13. In the copy number display (bottom), each length *k* sequence (*k*-mer) of the CHM13 assembly is painted with a color representing the copy number of that *k*-mer sequence in an SGDP sample. The CHM13 and GRCh38 tracks show the copy number of these same *k*-mers in the respective assemblies. CHM13 copy number resembles all samples from the SGDP, whereas GRCh38 underrepresents the true copy number.

Table 1.

Comparison of GRCh38 and T2T-CHM13v1.1 human genome assemblies.

Summary	GRCh38	T2T-CHM13	±%
Assembled bases (Gbp)	2.92	3.05	+4.5%
Unplaced bases (Mbp)	11.42	0	−100.0%
Gap bases (Mbp)	120.31	0	−100.0%
# Contigs	949	24	−97.5%
Ctg NG50 (Mbp)	56.41	154.26	+173.5%
# Issues	230	46	−80.0%
Issues (Mbp)	230.43	8.18	−96.5%
Gene Annotation			
# Genes	60,090	63,494	+5.7%
protein coding	19,890	19,969	+0.4%
# Exclusive genes	263	3,604	
protein coding	63	140	
# Transcripts	228,597	233,615	+2.2%
protein coding	84,277	86,245	+2.3%
# Exclusive transcripts	1,708	6,693	
protein coding	829	2,780	
Segmental duplications (SDs)			
% SDs	5.00%	6.61%	
SD bases (Mbp)	151.71	201.93	+33.1%
# SDs	24097	41528	+72.3%
RepeatMasker			
% Repeats	51.89%	53.94%	
Repeat bases (Mbp)	1,516.37	1,647.81	+8.7%
LINE	626.33	631.64	+0.8%
SINE	386.48	390.27	+1.0%
LTR	267.52	269.91	+0.9%
Satellite	76.51	150.42	+96.6%
DNA	108.53	109.35	+0.8%
Simple repeat	36.5	77.69	+112.9%
Low complexity	6.16	6.44	+4.6%
Retroposon	4.51	4.65	+3.3%
rRNA	0.21	1.71	+730.4%

GRCh38 summary statistics exclude “alts” (110 Mbp), patches (63 Mbp), and Chromosome Y (58 Mbp). Assembled bases: all non-N bases. Unplaced bases: not assigned or positioned within a chromosome. # Contigs: GRCh38 scaffolds were split at three consecutive Ns to obtain contigs. NG50: half of the 3.05 Gbp human genome size contained in contigs of this length or greater. # Exclusive genes/transcripts: for GRCh38, GENCODE genes/transcripts not found in CHM13; for CHM13, extra putative paralogs that are not in GENCODE. Segmental duplication analysis is from (42). RepeatMasker analysis is from (49).