

2021-05-27

The complete sequence of a human genome [preprint]

Sergey Nurk
National Institutes of Health

Et al.

Let us know how access to this document benefits you.

Follow this and additional works at: https://escholarship.umassmed.edu/faculty_pubs



Part of the [Genomics Commons](#)

Repository Citation

Nurk S, Rogaev EI, Eichler EE, Miga KH, Phillippy AM. (2021). The complete sequence of a human genome [preprint]. University of Massachusetts Medical School Faculty Publications. <https://doi.org/10.1101/2021.05.26.445798>. Retrieved from https://escholarship.umassmed.edu/faculty_pubs/2037

Creative Commons License



This work is licensed under a [Creative Commons Attribution 4.0 License](#).

This material is brought to you by eScholarship@UMassChan. It has been accepted for inclusion in University of Massachusetts Medical School Faculty Publications by an authorized administrator of eScholarship@UMassChan. For more information, please contact Lisa.Palmer@umassmed.edu.

The complete sequence of a human genome

Sergey Nurk^{1,*}, Sergey Koren^{1,*}, Arang Rhie^{1,*}, Mikko Rautiainen^{1,*}, Andrey V. Bzikadze², Alla Mikheenko³, Mitchell R. Vollger⁴, Nicolas Altemose⁵, Lev Uralsky^{6,7}, Ariel Gershman⁸, Sergey Aganezov⁹, Savannah J. Hoyt¹⁰, Mark Diekhans¹¹, Glennis A. Logsdon⁴, Michael Alonge⁹, Stylianos E. Antonarakis¹², Matthew Borchers¹³, Gerard G. Bouffard¹⁴, Shelise Y. Brooks¹⁴, Gina V. Caldas¹⁵, Haoyu Cheng^{16,17}, Chen-Shan Chin¹⁸, William Chow¹⁹, Leonardo G. de Lima¹³, Philip C. Dishuck⁴, Richard Durbin²¹, Tatiana Dvorkina³, Ian T. Fiddes²², Giulio Formenti^{23,24}, Robert S. Fulton²⁵, Arkarachai Fungtammasan¹⁸, Erik Garrison^{11,26}, Patrick G.S. Grady¹⁰, Tina A. Graves-Lindsay²⁷, Ira M. Hall²⁸, Nancy F. Hansen²⁹, Gabrielle A. Hartley¹⁰, Marina Haukness¹¹, Kerstin Howe¹⁹, Michael W. Hunkapiller³⁰, Chirag Jain^{1,31}, Miten Jain¹¹, Erich D. Jarvis^{23,24}, Peter Kerpedjiev³², Melanie Kirsche⁹, Mikhail Kolmogorov³³, Jonas Korlach³⁰, Milinn Kremitzki²⁷, Heng Li^{16,17}, Valerie V. Maduro³⁴, Tobias Marschall³⁵, Ann M. McCartney¹, Jennifer McDaniel³⁶, Danny E. Miller^{4,37}, James C. Mullikin^{14,29}, Eugene W. Myers³⁸, Nathan D. Olson³⁶, Benedict Paten¹¹, Paul Peluso³⁰, Pavel A. Pevzner³³, David Porubsky⁴, Tamara Potapova¹³, Evgeny I. Rogae^{6,7,39,40}, Jeffrey A. Rosenfeld⁴¹, Steven L. Salzberg^{9,42}, Valerie A. Schneider⁴³, Fritz J. Sedlazeck⁴⁴, Kishwar Shafin¹¹, Colin J. Shew²⁰, Alaina Shumate⁴², Yumi Sims¹⁹, Arian F. A. Smit⁴⁵, Daniela C. Soto²⁰, Ivan Sovic^{30,46}, Jessica M. Storer⁴⁵, Aaron Streets^{5,47}, Beth A. Sullivan⁴⁸, Françoise Thibaud-Nissen⁴³, James Torrance¹⁹, Justin Wagner³⁶, Brian P. Walenz¹, Aaron Wenger³⁰, Jonathan M. D. Wood¹⁹, Chunlin Xiao⁴³, Stephanie M. Yan⁴⁹, Alice C. Young¹⁴, Samantha Zarate⁹, Urvashi Surti⁵⁰, Rajiv C. McCoy⁴⁹, Megan Y. Dennis²⁰, Ivan A. Alexandrov^{3,7,51}, Jennifer L. Gerton¹³, Rachel J. O'Neill¹⁰, Winston Timp^{8,42}, Justin M. Zook³⁶, Michael C. Schatz^{9,49}, Evan E. Eichler^{4,24,†}, Karen H. Miga^{11,†}, Adam M. Phillippy^{1,†}

¹⁻⁵¹ Affiliations are listed at the end

* Equal contribution

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

Abstract

In 2001, Celera Genomics and the International Human Genome Sequencing Consortium published their initial drafts of the human genome, which revolutionized the field of genomics. While these drafts and the updates that followed effectively covered the euchromatic fraction of the genome, the heterochromatin and many other complex regions were left unfinished or erroneous. Addressing this remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium has finished the first truly complete 3.055 billion base pair (bp) sequence of a human genome, representing the largest improvement to the human reference genome since its initial release. The new T2T-CHM13 reference includes gapless assemblies for all 22 autosomes plus Chromosome X, corrects numerous errors, and introduces nearly 200 million bp of novel sequence containing 2,226 paralogous gene copies, 115 of which are predicted to be protein coding. The newly completed regions include all centromeric satellite arrays and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies for the first time.

Introduction

The latest major update to the human reference genome was released by the Genome Reference Consortium (GRC) in 2013 and most recently patched in 2019 (GRCh38.p13) (1). This assembly traces its origin to the publicly funded Human Genome Project (2) and has been

continually improved over the past two decades. Unlike the competing Celera assembly (3), and most modern genome projects that are also based on shotgun sequence assembly (4), the GRC human reference assembly is primarily based on Sanger sequencing data derived from bacterial artificial chromosome (BAC) clones that were ordered and oriented along the genome via radiation hybrid, genetic linkage, and fingerprint maps (5). This laborious approach resulted in what remains one of the most continuous and accurate reference genomes today. However, reliance on these technologies limited the assembly to only the euchromatic regions of the genome that could be reliably cloned into BACs, mapped, and assembled. Restriction enzyme biases led to the underrepresentation of many long, tandem repeats in the resulting BAC libraries, and the opportunistic assembly of BACs derived from multiple different individuals resulted in a mosaic assembly that does not represent a continuous haplotype. As such, the current GRC assembly contains several unsolvable gaps, where a correct genomic reconstruction is impossible due to incompatible structural polymorphisms associated with segmental duplications on either side of the gap (6). As a result of these shortcomings, many repetitive and polymorphic regions of the genome have been left unfinished or incorrectly assembled for over 20 years.

The current GRCh38.p13 reference genome 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 the entire p-arms (short arms) of all five acrocentric chromosomes (Chr13, Chr14, Chr15, Chr21, and Chr22), and large human satellite arrays (e.g., Chr1, Chr9, and Chr16), which are currently represented in the reference simply as multi-megabase stretches of unknown bases ('N's). In addition to these apparent gaps, other regions of the current reference are artificial or are otherwise incorrect. The centromeric alpha satellite arrays, for example, are represented in GRCh38 as computationally generated models of alpha satellite monomers to serve as decoys for resequencing analyses (7). In the case of the acrocentrics, some sequence is included for the p-arm of Chromosome 21 but appears incorrectly localized and poorly assembled, resulting in false gene duplications that complicate downstream analyses (8). When compared to other human genomes, the current reference also shows a genome-wide deletion bias, suggesting the systematic collapse of repeats during its initial cloning and/or assembly (9).

Despite the functional importance of these missing or erroneous regions, the Human Genome Project was officially declared complete in 2003 (10), and there was limited progress towards closing the remaining gaps in the years that followed. This was largely due to limitations of its construction discussed above, but also due to the sequencing technologies of the time, which were dominated by low-cost, high-throughput methods capable of sequencing only a few hundred bases per read. Thus, shotgun-based assembly methods were unable to surpass the quality of the existing reference. However, recent advances in long-read genome sequencing and assembly methods have enabled the complete assembly of individual human chromosomes from telomere to telomere without gaps (11, 12). In addition to using long reads, these T2T projects have targeted the genomes of clonal, complete hydatidiform mole (CHM) cell lines, which are almost completely homozygous and therefore easier to assemble than heterozygous

diploid genomes (13). This single-haplotype, *de novo* strategy overcomes the limitations of the GRC's mosaic BAC-based legacy, bypasses the challenges of structural polymorphism, and allows the use of modern genome sequencing and assembly methods.

Application of long-read sequencing for the improvement of the human reference genome followed the introduction of PacBio's single-molecule, polymerase-based technology (14). This was the first commercial sequencing technology capable of producing multi-kilobase sequence reads, which, even with a 15% error rate, proved capable of resolving complex forms of structural variation and gaps in GRCh38 (9, 15). The next major advance in sequencing read lengths came from Oxford Nanopore's single-molecule, nanopore-based technology, capable of sequencing "ultra-long" reads in excess of 1 Mbp (16), but again with an error rate of 15%. By spanning most genomic repeats, these ultra-long reads enabled highly continuous *de novo* assembly (17), including the first complete assemblies of a human centromere (ChrY) (18) and a human chromosome (ChrX) (11). However, due to their high error rate, these long-read technologies have posed considerable algorithmic challenges, especially for the reliable assembly of long, highly similar repeat arrays (19). Improved sequencing accuracy simplifies the problem, but past technologies have excelled at either accuracy or length, not both. PacBio's recent "HiFi" circular consensus sequencing offers a compromise of 20 kbp read lengths and a median accuracy of 99.9% (20, 21), which has resulted in unprecedented assembly accuracy with relatively minor adjustments to standard assembly approaches (22, 23). Whereas ultra-long nanopore sequencing excels at spanning long, identical repeats, HiFi sequencing excels at differentiating subtly diverged repeat copies or haplotypes.

In order to create a complete and gapless human genome assembly, we leveraged the complementary aspects of PacBio HiFi and Oxford Nanopore ultra-long read sequencing, combined with the essentially haploid nature of the CHM13hTERT cell line (hereafter, CHM13) (24). 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 the first truly complete human reference genome and discuss its potential impact on the field.

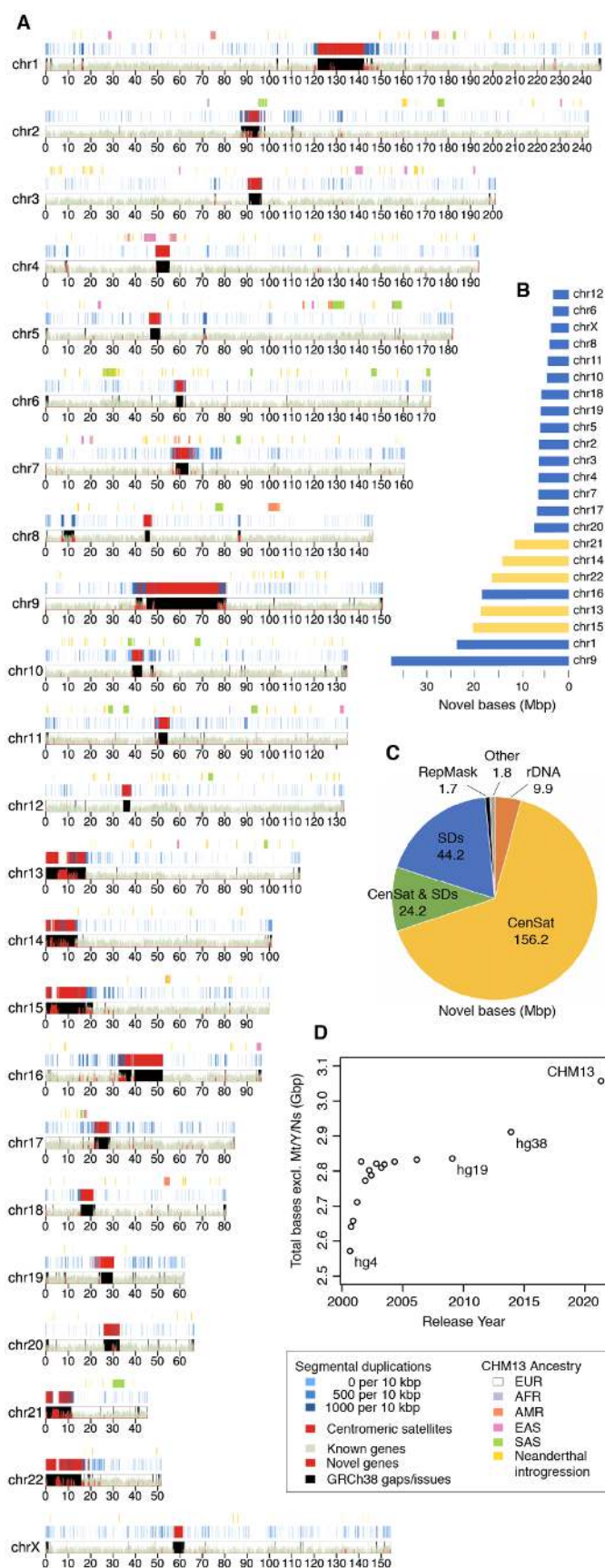


Fig. 1. Summary of the complete T2T-CHM13 human genome assembly.

(A) karyoplotR (25) ideogram of the T2T-CHM13v.1.1 assembly improvements. The bottom track shows the density of known genes in green and new paralogs in red. GRCh38 gaps and issues that are resolved by the CHM13 assembly are highlighted by black rectangles. Above, the density of segmental duplications is given in blue (26) and centromeric satellites (CenSat) in red (27). The top track is a local ancestry analysis where the majority of the genome is predicted to be of European ancestry (1000 Genomes EUR), with regions of admixture colored as specified in the legend. **(B)** New bases in the CHM13 assembly relative to GRCh38 per chromosome, with the acrocentrics highlighted in yellow. **(C)** New or structurally variable bases added by sequence type (“CenSat & SDs” is the overlap between these two annotations). **(D)** Total non-gap bases in UCSC reference genome releases dating back to September 2000 (hg4) and ending with T2T-CHM13 in 2021.

Cell line and sequencing

As with many prior reference genome improvement efforts (1, 9, 13, 24, 28, 29), including the T2T assemblies of human chromosomes X (11) and 8 (12), we utilized a complete hydatidiform mole for sequencing. CHM genomes arise from the loss of the maternal complement and duplication of the paternal complement postfertilization and are, therefore, homozygous for one set of alleles. This simplifies the genome assembly problem by removing the confounding effect of heterozygous variation. We selected CHM13 for its stable 46,XX karyotype compared to other CHMs (11), but later found that CHM13 does possess a low level of heterozygosity, notably including a megabase-scale heterozygous deletion within the rDNA array on Chromosome 15, which was revealed by both FISH and nanopore sequencing (Figs. S1-2, Note S1). This and other identified heterozygous variants appear fixed in CHM13 and may have arisen during growth of the mole or passaging of the cell line. Local ancestry analysis shows the majority of the CHM13 genome is of European origin, including regions of Neanderthal introgression, with some predicted admixture from other populations (30) (**Fig. 1A**, Note S2).

Over the past 6 years, we have extensively sequenced CHM13 with multiple technologies (Note S3), including 30× PacBio circular consensus sequencing (HiFi) (29), 120× Oxford Nanopore ultra-long read sequencing (ONT) (11, 12), 100× Illumina PCR-Free sequencing (ILMN) (1), 70× Illumina / Arima Genomics Hi-C (Hi-C) (11), BioNano optical maps (11), and Strand-seq (29). Here we developed new methods for assembly, polishing, and validation that better utilize these datasets. In contrast to the first T2T assembly of Chromosome X (11)—which relied on ONT sequencing to create a backbone that was then polished with other technologies—we shifted to a new strategy that leverages the combined accuracy and length of HiFi reads to enable assembly of highly repetitive centromeric satellite arrays and closely related segmental duplications (12, 22, 29).

Genome assembly

The basis of the T2T-CHM13 assembly is a high-resolution assembly string graph (31) 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 HiFi-based string graph was constructed using a purpose-built method that combines components from the HiCanu (22) and Miniasm (32) assemblers along with specialized graph processing. Although HiFi reads are very accurate, their primary error mode is small insertions or deletions within homopolymer runs, so, like HiCanu, the first step of the T2T string graph construction process was to “compress” homopolymer runs in the reads to a single nucleotide (e.g., $[A]_n$ becomes $[A]_1$ for $n > 1$) (33). All compressed reads were then aligned to one another to identify and correct small errors, and differences within simple sequence repeats were masked to overcome this other known source of HiFi errors (22). After compression, correction, and masking, only exact overlaps were considered during graph construction, and new methods were developed for iterative graph simplification, as described in the supplementary methods (Fig. S3, Note S4). Edges in the resulting string graph correspond to exact overlaps of at least 8 kbp in homopolymer-compressed space.

In the resulting graph, most chromosomes are represented by one or more connected components, each having a mostly linear structure (**Fig. 2A**). This suggests very few perfect repeats greater than roughly 10 kbp exist between different chromosomes or distant loci, with the exception of the five acrocentric chromosomes, which form a single connected component in the graph. Another complex region is the HSat3 array on Chromosome 9, which includes a recent multi-megabase tandem HSat3 duplication consistent with the 9qh+ (34) karyotype of CHM13 (Fig. S4). Minor fragmentation of the chromosomes into multiple connected components resulted from HiFi sequencing dropout across some GA-rich simple sequence repeats, presumably due to a bias of the HiFi sequencing or base-calling process (22). These gaps were later filled using a prior ONT-based assembly (CHM13v0.7) (11).

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) (Fig. S5). To help identify the correct walks, we estimated coverage depth and multiplicity of the string graph nodes (Note S4), which allowed most tangles to be resolved as unique walks visiting each node the appropriate number of times (**Fig. 2B-D**, Fig. S5). Most of these walks were identified via manual curation of the graph. Low coverage nodes (e.g., <15% the average coverage) were presumed to stem from sequencing errors or somatic variants and were either pruned from the graph or omitted from the walks. Half-coverage nodes arranged as simple bubbles were most often heterozygous variants and the longer of the two was chosen with ties broken arbitrarily. In the remaining cases, the correct path was ambiguous and required integration of ONT reads. Where possible, raw ONT reads were aligned to the HiFi-based string graph using GraphAligner (35) to guide the correct walk, but more elaborate strategies were required for the large and highly similar satellite array duplications on chromosomes 6 and 9 (Fig. S6, Note S4). Only the five rDNA arrays, constituting approximately 10 Mbp of sequence, were not resolved as walks through the string graph and required a specialized assembly approach (described below). With this exception, an accurate consensus sequence for each chromosome was obtained from a multi-alignment of HiFi reads corresponding to the selected graph walks (Note S4), resulting in the CHM13v0.9 draft assembly.

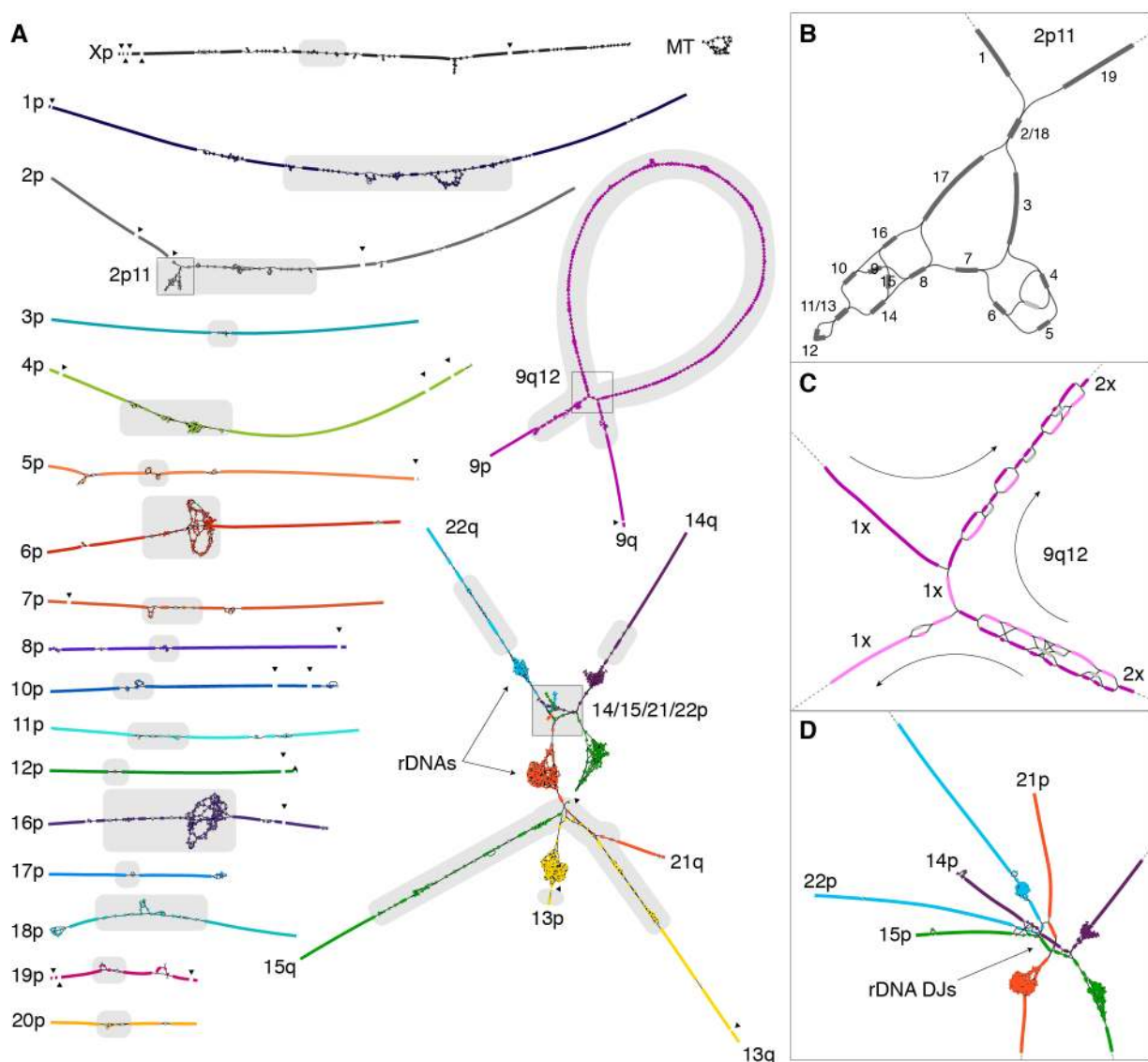


Fig. 2. HiFi-based assembly string graph of the CHM13 genome. (A) Bandage (36) string graph visualization, where nodes represent unambiguously assembled sequences colored by source chromosome and scaled by length. Edges correspond to the overlaps between node sequences, due to either repeats or true adjacencies in the underlying genome. Centromeric satellite sequences are the source of most ambiguity in the graph (gray highlights). The graph is partially fragmented due to HiFi coverage dropout surrounding GA-rich sequence (black triangles). Local graph structures are enlarged with insets. The correct graph walks through these complex structures were resolved and confirmed with ultra-long ONT reads. **(B)** The identified graph traversal for the 2p11 locus is given by numerical order. Based on a depth-of-coverage analysis, the unlabeled light gray node represents an artifact or possible heterozygous variant and was not used. **(C)** The multi-megabase tandem HSat3 duplication (9qh+) at 9q12 requires two traversals of the large loop structure (note that the size of the loop is exaggerated because graph edges are of constant size). The first traversal of the loop is given in dark purple and the second in light purple. Nodes used by both traversals are also in dark purple and typically have twice the sequencing coverage. **(D)** The telomeric ends of four acrocentric p-arms form an ambiguous graph structure due to the highly similar sequence shared between all four chromosomes, specifically within the distal junction (DJ) sequence adjacent to the rDNA arrays, which themselves form dense, but separate, clusters of small nodes.

Assembly validation and polishing

The first step of genome assembly validation is to confirm that the constructed assembly is consistent with the data used to generate it (37). 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 v0.9 draft assembly using Winnowmap2 (38) for long reads and BWA (39) for short reads (Note S5). Structural variants were identified with Sniffles (40), and small variants were called with DeepVariant (41) for ILMN and HiFi and PEPPER-DeepVariant for ONT (42). Small variants were further filtered using Merfin (43) to exclude any corrections that were not supported by the underlying ILMN and HiFi reads. After manual curation to differentiate true errors from heterozygous variants and mapping artifacts, a total of 4 large variants and 993 small variants were corrected, 52% of which were small indels within homopolymers. An additional 44 large and 3,901 small heterozygous variants were cataloged during curation (44), including the hemizygous insertion of an hTERT vector on Chromosome 21, consistent with the immortalization process used to create the CHM13 line (this insertion was excluded from the final assembly). The assembled sizes of major repeat arrays were consistent with ddPCR copy-number estimates for those tested (Tables S1-2, Fig. S7), and both Strand-seq (Figs. S8-9) and Hi-C (Fig. S10) data were concordant with the overall structure of the assembly, showing no signs of misorientations or other large-scale structural errors. In addition, the assembly correctly resolved 644 of 647 previously sequenced CHM13 BACs at >99.99% identity, with the three unresolved BACs appearing to be errors in the BACs themselves rather than the T2T assembly (Figs. S11-14).

The entire validation process was then repeated on the polished assembly, and investigation of the remaining variant calls revealed additional base-calling errors within some telomeric $[\text{TTAGGG}]_n$ repeats. These putative errors were primarily a result of decreased coverage by both HiFi and ONT technologies towards the telomeres and not flagged by the initial variant-calling pipeline due to a telomere-associated strand bias in the ONT data. Telomeric ONT reads were only found oriented in the direction of the chromosome end and never away from it, which led to low-confidence variant calls and omission from polishing. After further curation and adjustment of the variant calling strategy (Note S4), an additional 454 corrections were made to the telomeres using PEPPER (42), followed by addition of the rDNA arrays as described below, resulting in a gapless CHM13v1.1 assembly—the first telomere-to-telomere representation of a human genome.

Mapped sequencing read depth across the final assembly shows uniform coverage across all chromosomes (**Fig. 3A**), with 99.86% of the assembly within three standard deviations of the mean coverage for both HiFi and 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 (Note S5). This is consistent with uniform coverage of the genome and confirms both the overall accuracy of the assembly and the absence of aneuploidy in the sequenced CHM13 cells. Copy-number concordance with raw ILMN and HiFi data also increased with successive versions of the assembly (Figs. S15-16). Local coverage anomalies were, however, observed across multiple satellite arrays (Table S3, Note S6). Given the uniformity of coverage increases

and decreases 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 systematic biases introduced during either sample preparation (e.g., shearing bias) or sequencing (e.g., polymerase kinetics), rather than assembly error (Note S6). For example, HiFi coverage is consistently elevated across HSat2 and HSat3 arrays, while ONT coverage remains normal but with an apparent strand bias and reduced read lengths for HSat2 (**Fig. 3B-C**, Figs. S17-S21). On the other hand, both HiFi and ONT coverage is depleted across the AT-rich HSat1 arrays, with ONT reads also showing shorter read lengths (**Fig. 3D**, Figs. S17-18, Table S3). 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 (45, 46).

Due to the challenge of assembling them correctly, we performed targeted validation of all large satellite arrays and segmental duplications (Note S7). For centromeric alpha satellite arrays, we used the TandemTools package (47) to catalog additional variants that were missed by the standard approach. TandemTools was used throughout the process to guide development of the assembly method, and analysis of the final assembly shows high accuracy across all centromeric arrays (Fig. S22, Table S4). Independent ILMN-based copy number estimates of alpha satellite higher-order repeats (HOR) also correlate strongly with the assembly (Fig. S23). The beta satellite (BSat) and HSat arrays were separately validated by measuring the frequency of secondary variants identified by HiFi read mappings using a technique previously developed to identify collapsed segmental duplications (48). Because CHM13 is mostly homozygous, we expect to find very few heterozygous variants when mapping the raw reads back to the assembly and any variant clusters would indicate potential mis-assembly. This analysis shows consistent coverage across all satellite arrays, with only a handful of potential variants flagged (Fig. S24). A companion study (26) used this same approach to validate segmentally duplicated regions of the genome, along with an analysis of copy number variation compared to a collection of diverse human genomes, demonstrating that T2T-CHM13 represents these complex regions better than GRCh38.

In addition to high structural accuracy, we estimate the average consensus accuracy of the assembly to be between Phred Q67 and Q73 (Note S5), which is equivalent to 1 error per 10 Mbp and far exceeds the original Q40 definition of “finished” sequence (49). However, this represents an average across the entire genome and some regions are expected to be higher quality than others. In particular, regions of low HiFi coverage were found to be associated with an enrichment of potential consensus errors, as estimated from both HiFi and ILMN data (44). To guide future use of the assembly, we provide a curated list of all low-coverage and known heterozygous sites identified by the above validation procedures (Note S5). 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**), making T2T-CHM13 a more complete, accurate, and representative reference sequence for both short- and long-read variant calling across human samples of all ancestries (50). Compared to GRCh38, T2T-CHM13 reduces false negative variant calls by adding 182 Mbp of novel sequence and removing 1.2 Mbp of falsely duplicated sequence, while simultaneously reducing false positive variant calls by fixing collapsed segmental duplications and other errors, affecting a total of at least 388 genes (68

protein coding) in GRCh38. Lastly, the T2T-CHM13 haplotype structure and SNP density is much more consistent than the mosaic GRCh38 when calling variants. A full comparison of GRCh38 versus CHM13 as a reference for variant calling is provided by Aganezov *et al.* (50), and a discussion of validation and polishing strategies for T2T genome assemblies by McCartney *et al.* (44).

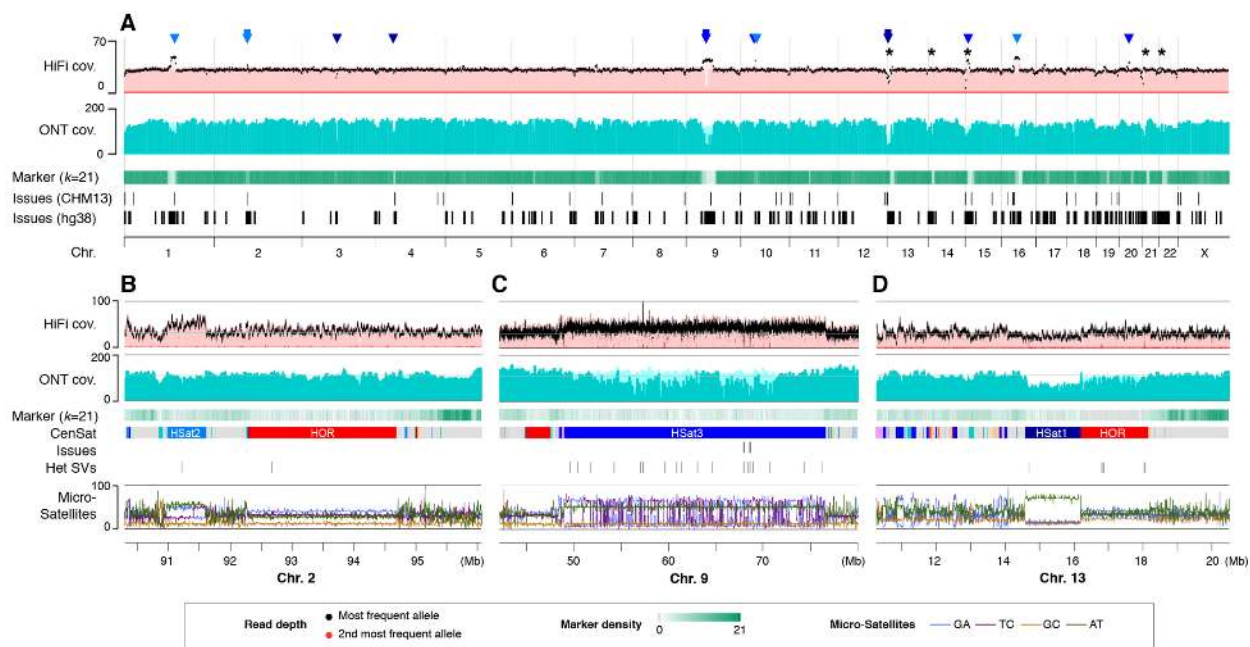


Fig. 3. Sequencing coverage and assembly validation. Both HiFi and ONT sequencing reads mapped to the assembly show uniform coverage (cov.) across the whole genome as visualized by IGV (51), with the exception of certain human satellite repeat classes. Coverage deviations in these regions were found to be caused by sequencing biases associated with specific repeats rather than misassembly. **(A)** Whole-genome coverage of HiFi and ONT reads with primary alignments is shown in light shades and marker-assisted alignments overlaid in dark shades. Large HSat2 and HSat3 arrays are noted with light and dark blue triangles, respectively, and the location of the rDNA arrays is marked with asterisks (the inset regions are marked with arrowheads). Regions with low marker-assisted alignment coverage correspond with a lack of unique 21-mer markers (density shown in green), but are recovered by the primary alignments, albeit with low mapping quality. Suspected assembly issues in T2T-CHM13 are compared to known assembly gaps and issues in GRCh38/hg38 below, as reported by the GRC. **(B–D)** Enlargements corresponding to regions of the genome featured in Figure 2, along with annotations of the major satellite repeats contained (primarily HSat1, HSat2, HSat3, and alpha satellite HOR arrays). Elevated HiFi sequencing coverage is observed for HSat2 and HSat3, while reduced ONT coverage is observed for HSat1. Identified errors (Issues) and heterozygous variants (Het SVs) are shown below, which typically correspond with low HiFi coverage of the primary allele (black) and elevated coverage of a secondary allele (red). Microsatellite repeats (%) in every 128 bp window are shown at the bottom, labeled with dimer notation in homopolymer compressed space.

rDNA assembly

The most complex region of the HiFi string graph involves the human ribosomal DNA arrays and their surrounding sequence (Fig. 2). Human rDNAs are 45 kbp near-identical repeats that encode the 45S rRNA and are arranged in large, tandem repeat arrays embedded within the

acrocentric p-arms. The length of these arrays varies widely between individuals (52) and even somatically, especially with aging and certain cancers (53). Based on ddPCR (Fig. S7) and whole-genome ILMN coverage (Figs. S15, S25), we estimate that the CHM13 genome contains approximately 200 rDNA copies per haplotype. 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 that was confirmed by FISH and a single 450 kbp spanning ONT read (Figs. S1-2, Note S1). In all cases we chose to include the variant that was most similar to the canonical rDNA unit, or in the case of Chromosome 15, the larger variant since it contained additional rDNA units. No other ONT reads spanned entire rDNA arrays, which is unsurprising given their multi-megabase size, and so additional copy number variation within the interior of the arrays could not be ruled out. However, all constructed assembly graphs exclude the presence of additional, non-rDNA, sequences within the arrays.

To assemble these highly dynamic regions of the genome, and overcome potential limitations of the string graph construction (Notes S4,8), we used a sparse de Bruijn graph approach (54). After recruiting HiFi reads from the rDNAs, we built a homopolymer-compressed sparse de Bruijn graph using a large word (k -mer) size, which formed a separate connected component for each rDNA array. We used this graph to separate all HiFi and ONT reads by chromosome, and then rebuilt the de Bruijn graph for each chromosome using a smaller word size and no homopolymer compression to improve connectivity. 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. S26). ONT reads were then aligned to this graph with GraphAligner (35) 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. 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. S26). A consensus sequence was then constructed for each morph and copy number roughly estimated from the number of supporting ONT reads. Adjacency information was inferred from ONT reads spanning two or more rDNA units, and this was used to build a graph of morphs representing the structure of each rDNA array (Fig. S26). The most common morphs are self-adjacent in this graph, meaning they are arranged in continuous, head-to-tail, repeating blocks. The arrays on chromosomes 14 and 22 were found to contain a single primary morph repeated approximately 20 times and flanked by low-copy, more diverged units at the boundaries. In these cases, ONT reads extended far enough into the arrays to reconstruct the boundary units from the original HiFi string graph, and an appropriate number of polished primary morph copies were inserted into the center of the array. Chromosomes 13, 15, and 21 however, exhibited a more mosaic structure involving multiple, interspersed morphs, and the ONT reads were not long enough to determine the correct traversal of the morph graph. In these cases, after reconstructing the boundary units as before, the remaining primary morphs were polished and arranged in consecutive blocks, based on their coverage and order in the graph. The resulting (haploid) assembly contains 219 complete rDNA copies, totalling 9.9 Mbp of sequence. These rDNA assemblies approximate the true copy number and capture the common morphs on each

chromosome, but the internal portions of arrays 13, 15, and 21 have been artificially ordered and should be treated as model sequences.

A truly complete genome

The T2T-CHM13v1.1 assembly 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 (CHM13 does not have a Y chromosome). This complete assembly adds or corrects 238 Mbp of sequence compared to GRCh38, defined as regions of the T2T-CHM13 assembly that do not linearly align to GRCh38 over a 1 Mbp interval (i.e., are non-syntenic). The bulk of this sequence comprises centromeric satellites (180 Mbp), segmental duplications (68 Mbp), and rDNAs (10 Mbp), noting that there is overlap between regions identified as centromeric and segmentally duplicated (**Fig. 1B-C**). Of these regions 182 Mbp of sequence is uncovered by any primary alignments from GRCh38 and, therefore, completely novel to the CHM13 assembly. As a result, T2T-CHM13v1.1 substantially increases the number of known genes and repeats in the human genome (**Table 1**).

Table 1. Comparison of GRCh38 and T2T-CHM13 human genome assemblies.

Summary	GRCh38p13	CHM13v1.1	±%
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	50.03%	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%

GRCh38p13 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 (26). RepeatMasker (55) analysis is from (56), with the "Unknown" category not shown.

To provide an initial annotation, we used both the Comparative Annotation Toolkit (CAT) (57) and Liftoff (58) to project the GENCODE v35 (59) reference annotation onto the new T2T-CHM13 assembly. Additionally, CHM13 Iso-Seq transcriptome data was processed using StringTie2 (60) to generate a set of *de novo* assembled transcripts, which were provided as complementary input to CAT. Genes identified by Liftoff, but missing from the CAT annotation, were then added to provide a comprehensive annotation of the newly assembled sequence (Note S9). This represents an initial draft annotation of the newly completed regions of the genome that will require further analysis and validation that is beyond the scope of this study.

The draft T2T-CHM13 annotation totals 63,494 genes and 233,615 transcripts, of which 19,969 and 86,245 are predicted to be protein coding, respectively, with 469 predicted frameshifts in 387 genes (**Table 1**, Fig. S27, Tables S5-7). Only 263 GENCODE genes (448 transcripts) are exclusive to GRCh38 and have no assigned ortholog in the CHM13 annotation (Fig. S28, Tables S8-9). At least 24 of these genes correspond to presumed false duplications and other errors in GRCh38 (50) (Fig. S29), while most others originate from repetitive regions that may be successfully annotated in the future. In comparison, a total of 3,604 genes (6,693 transcripts) are exclusive to CHM13 (Tables S10-11). Due to the segmentally duplicated nature of the newly assembled regions, these genes represent putative paralogs and localize to pericentromeric regions and the acrocentric p-arms, including 876 rRNA transcripts (**Fig. 1A**). Of these new genes, 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 S12). Due to CHM13's more representative copy number in complex regions of the genome, many of these additions represent gene family expansions, which has the dual benefit of adding new paralogs to the reference and improving variant analyses of the previously known copies (50). While some of the new paralogs in the CHM13 annotation may also be present (but unannotated) in GRCh38, 2,226 of the genes exclusive to CHM13 (115 protein coding) fall within regions of the genome with no GRCh38 primary alignments (Note S9) and are verifiably novel.

Acrocentric chromosomes

We focus here on describing the genomic structure of the newly completed p-arms of the five acrocentric chromosomes, which, despite their obvious importance for basic cellular function, have remained largely unsequenced to date. This omission has been due to the high enrichment for satellite repeats and segmental duplications throughout the acrocentric p-arms, which has prohibited *de novo* sequence assembly and limited their past characterization to cytogenetics (61), restriction mapping (62), and BAC sequencing (63–65). All five acrocentric p-arms follow a similar structure consisting of a variable-sized rDNA array embedded within distal and proximal satellite arrays, but the size of these arrays is highly polymorphic (52, 53).

Their importance to genome biology includes ribosome biogenesis, nucleolus formation (66), chromosomal instability (67), and known genetic conditions including ribosomopathies (68), Robertsonian translocations (69), and Down Syndrome (70). However, the lack of a reference sequence for these five chromosome arms has resulted in their exclusion from high-resolution genomic analysis and association studies.

The T2T-CHM13 assembly uncovers the complete sequence of all acrocentric p-arms for the first time, which vary in size from 10.1 Mbp to 16.7 Mbp each and amount to 66.1 Mbp of new sequence (**Fig. 4**). Each p-arm contains a single rDNA array varying in size from 0.7 Mbp (Chr14) to 3.6 Mbp (Chr13), surrounded by abundant satellite repeats and segmental duplications. Compared to other human chromosomes, the acrocentrics are also unusually similar to one another in terms of their p-arm sequence and composition and, as a result, were the only chromosomes not separated into individual components during string graph construction (**Fig. 2D**). Specifically, we find that 5 kbp windows on the acrocentric p-arms align with a median identity of 98.7% between chromosomes, creating many opportunities for interchromosomal exchange. This high degree of similarity is presumably due to recent non-allelic or ectopic recombination stemming from their colocalization in the nucleolus (64). Additionally, no 5 kbp window is unique considering alignments >80% identity, and 96% of the non-rDNA sequence can be found elsewhere in the genome using a similar criteria, suggesting that the acrocentric p-arms are frequent and dynamic sources of segmental duplication.

Situated between the distal and proximal short arms, CHM13's five rDNA arrays 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 the vast majority of rDNA units are full length, in contrast to some prior studies that reported embedded inversions and other non-canonical structures (65, 71). Each array appears highly homogenized, and there is more variation between rDNA units on different chromosomes than within chromosomes (Fig. S30), which suggests that intra-chromosomal exchange of rDNA units via non-allelic homologous recombination is more common than inter-chromosomal exchange. For example, most 45S gene copies on the same chromosome are 100.0% identical to one another (with some alternative morphs in the longer arrays), while the identity of the most frequent 45S morphs between chromosomes ranges from 99.4–99.7%. A minor chromosome 15 morph shows the most similarity to the KY962518.1 reference sequence, which was originally derived from a human Chromosome 21 BAC clone (65). These two sequences align across the entirety of the 45 kbp rDNA at 98.9% identity. The primary Chromosome 21 morph is also highly similar to KY962518.1, but includes a 96 bp deletion in the intergenic spacer (IGS). As expected, the 13 kbp 45S is more highly conserved than the IGS, with all major CHM13 45S morphs aligning between 99.4–99.6% identity to KY962518.1. Certain rDNA variants appear chromosome-specific in CHM13, including single-nucleotide variants within the 45S and its upstream promoter region (Fig. S31). The most evident variants are repeat expansions and contractions within the tandem repeat that immediately follows the 45S and the large, CT-rich simple sequence repeat located in the middle of the IGS. Most CHM13 rDNA morphs can be distinguished based solely on structural differences within these two polymorphic sites, and each array contains a major morph that is unique to that chromosome (Fig. S32).

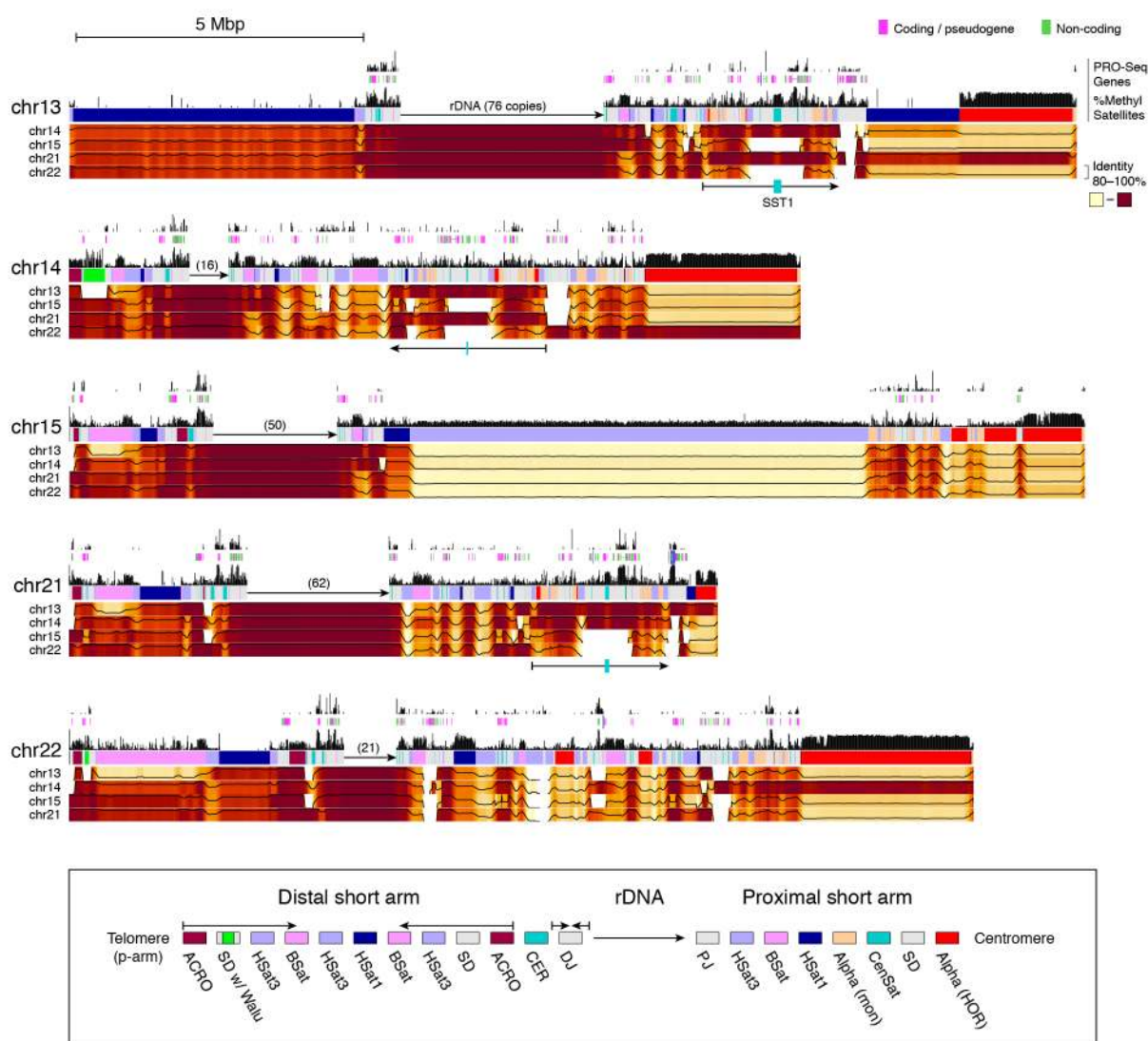


Fig. 4. Short arms of the human acrocentric chromosomes in CHM13. Each p-arm is shown along with RNA polymerase activity (PRO-Seq), annotated genes, percent of methylated CpGs, and color-coded satellite repeat annotation. ILMN PRO-Seq alignments were filtered by unique 51-mers to avoid cross mapping in these repetitive regions, and so may appear sparser than reality. Methylation data is derived from ultra-long ONT reads and is more comprehensive. The rDNA arrays are represented only by a directional arrow and copy number due to their high self-similarity, which limits mapping. Below the annotations are percent identity heatmaps versus the other four p-arms, where sequence identity was 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 legend shows the names and organization of satellite repeats, which is generally conserved on the distal short arms, including a large symmetric duplication (long arrows) and a highly conserved inverted repeat in the DJ (short arrows). The proximal short arms show a greater diversity of structures, and so the legend simply lists some common elements. The proximal short arms of Chromosomes 13, 14, and 21 share a large segmentally duplicated core, including small alpha satellite HOR arrays and a central, highly methylated, SST1 array (long arrows with a central teal block). Chromosomes 13/21 and 14/22 also share a striking degree of similarity across the centromere and pericentromere.

Beginning with the p-arm telomere and extending 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 widely among chromosomes. CHM13 Chromosome 13 is most unique in its distal structure, having lost half of the inverted satellite duplication and gained a massively expanded HSat1 array relative to the other four chromosomes. 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 newly discovered 64-bp Alu-associated satellite repeat (“Walu”) within the distal inverted duplication (56), the location of which was confirmed via FISH (Fig. S33). Several of the non-satellite sequences show evidence of transcription, including a number of gene predictions. 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 the few hundred kilobases of DJ sequence previously characterized via BACs (64, 72).

Extending from the rDNA array to the centromere, the proximal short arms are larger in size and show a higher diversity of structures compared to the distal arrays, but they remain highly similar at the sequence level with each arm being entirely composed of shuffled segmental duplications, transposable element composite arrays (e.g. ACRO_COMP), and satellite repeats including HSat3, BSat, HSat1, HSat5, and various centromeric satellites including both monomeric and higher-order repeat (HOR) alpha satellite arrays. With the exception of Chromosome 13—which includes a large, proximal HSat1 array—HSat3 and BSat are most prevalent in terms of total bases. Chromosome 15 includes a distinctive 8 Mbp HSat3 array that comprises the majority of its proximal short arm. Many of the proximal BSat arrays show a previously unreported mosaic inversion structure that we also discovered in some HSat arrays elsewhere in the genome (27) (Fig. S34). All regions of segmental duplication on the acrocentric p-arms appear highly similar between chromosomes, often exceeding 99% identity, and these regions also show evidence of active transcription (56) (**Fig. 4**, PRO-Seq). Notably, the proximal short arms of chromosomes 13, 14, and 21 appear to share the highest degree of structural and sequence similarity, including a large region of segmental duplication with a central and highly methylated SST1 array (**Fig. 4**). This similarity coincides with these three chromosomes being most frequently involved in Robertsonian translocations (73), providing possible mechanistic insight into their formation. As previously described (74, 75), alpha satellite HORs in the arrays of active centromeres of chromosomes 13/21 and chromosomes 14/22 are almost identical within each pair, but not between them. However, small alpha satellite HOR arrays within segmental duplications of the proximal short arms show a different pattern (**Fig. 4**). Here, chromosomes 13, 14 and 21 share similar HOR subsets, while Chromosome 22 is distinguished by two copies of a segmental duplication in which a Chromosome 22-specific alpha satellite HOR has amplified (27). Using this new reference sequence as a basis, further study of additional genomes is needed to understand which features of the acrocentric p-arms observed in CHM13 are conserved across the human population.

Analyses and resources

A number of companion studies were carried out to fully characterize the new sequences and structural variants revealed by the T2T-CHM13 assembly. These include comprehensive analyses of centromeres, segmental duplications, transcriptional and epigenetic profiles, mobile elements, and variant calls. Altemose *et al.* (27) carried out the first high-resolution study of satellite sequence organization at sites of centromere formation and discovered new sequences coincident with centromere-specific chromatin. Vollger *et al.* (26) identified nearly double the number of previously known near-identical segmental duplication alignments in the human genome, thereby identifying many new genes and regions susceptible to rearrangement via unequal crossing over. Additionally, comparison of T2T-CHM13, GRCh38, and high-quality primate assemblies revealed unprecedented patterns of structural heterozygosity and massive evolutionary differences in duplicated gene content between humans and the other apes (26). Gershman *et al.* (76) developed methods for epigenetic profiling of complete human genomes and uncovered novel methylation patterns across all families of human repeats, including a hypomethylated and highly dense chromatin region within the centromere which may guide kinetochore formation. Hoyt *et al.* (56) carried out a comprehensive update of human repeat models and annotations, revealing previously unknown repeat arrays, transposable element (TE) derived composite elements, and transcriptionally active TEs embedded within large satellite arrays. Coupled with PRO-Seq (77) and the ONT-derived methylation data, sites of engaged RNA polymerase were annotated genome-wide, revealing unique transcriptional and epigenetic profiles across highly repetitive regions of the human genome. Lastly, Aganezov *et al.* (50) compared variant calling results for GRCh38 and T2T-CHM13, and observed improvements in both short- and long-read variant calling with the use of CHM13, including within medically relevant genes that were found to be mis-assembled or falsely duplicated within GRCh38.

These studies have generated a rich variety of omics datasets for CHM13 that are all browsable via a UCSC Assembly Hub (**Data and materials**). In addition to the sequencing data, assembly, and gene annotation described here, the browser provides a central resource for all results of the above noted companion studies, including detailed annotations of centromeres and satellite DNAs (27); genome-wide repeat and transposable element annotations (56); segmental duplications and population copy number estimates (26); variant calls and mappability scores (50); and whole-genome alignments to GRCh38. Importantly, the newly finished regions of the genome are readily accessible to long-read mapping and variant calling, and many others can be sparsely mapped with short reads using unique marker sequences (Fig. S35, Table S13, Note S11). We have surveyed these new regions using a variety of modern omics approaches, all of which are available in the browser, including RNA-Seq (27), Iso-Seq (26), PRO-Seq (56), CUT&RUN (27), and ONT methylation (76) experiments. An additional interactive dotplot visualization (78) is also provided to explore the newly uncovered genomic repeats. Finally, the CHM13hTERT cell line itself is in the process of being banked at Coriell (Camden, NJ, USA) and will be made available for research use, providing a physical reference material alongside the digital reference sequence, something that is not possible for GRCh38.

To highlight the utility of these genetic and epigenetic resources mapped to a complete human genome, we provide the specific example of Facioscapulohumeral muscular dystrophy (FSHD) region gene 1 (*FRG1*), which is a poorly understood candidate gene for FSHD located in the subtelomeric region of human Chromosome 4q35. The precise etiology of FSHD remains unclear due to complex epigenetic factors surrounding *FRG1* (79), including the neighboring double homeobox 4 (*DUX4*) transcription factor (80) and polymorphic D4Z4 macrosatellite array (81). A near-identical duplication of the D4Z4 array on Chromosome 10q26 was previously characterized, but the T2T-CHM13 assembly reveals 23 paralogs of *FRG1* duplicated throughout previously unfinished regions of chromosomes 9, 13, 14, 15, 20, 21, and 22 (**Fig. 5**). These duplicated genes were previously identified by fluorescence *in situ* hybridization (82) and underwent recent amplification in the great apes (83), but only 9 paralogs are found in GRCh38, hampering sequence-based analysis. The new paralogs' association with all five acrocentrics is particularly notable, given that *FRG1* also localizes to nucleoli and is thought to play a role in RNA biogenesis (84). One of the few paralogs included in GRCh38, *FRG1DP*, is located in the centromeric region of Chromosome 20 and shares high identity (97%) with some of the newly assembled paralogs (*FRG1BP4–10*) situated in or near the acrocentrics (Fig. S36, Table S14-15, Note S12). When aligning HiFi reads to GRCh38, absence of these paralogs from the reference caused their reads to incorrectly align to *FRG1DP*, demonstrating the difficulty of investigating these genes with GRCh38 (**Fig. 5B**). Most of the new paralogs are found in other human genomes (**Fig. 5C**), and all paralogs except *FRG1KP2* and *FRG1KP3* have CpG islands overlapping the transcription start site with varying degrees of methylation and expression evidence in CHM13 (**Fig. 5D**, Table S16). This is just one example of the thousands of putative new paralogs uncovered by the T2T-CHM13 assembly, which we expect to drive future discovery in human genomic health and disease.

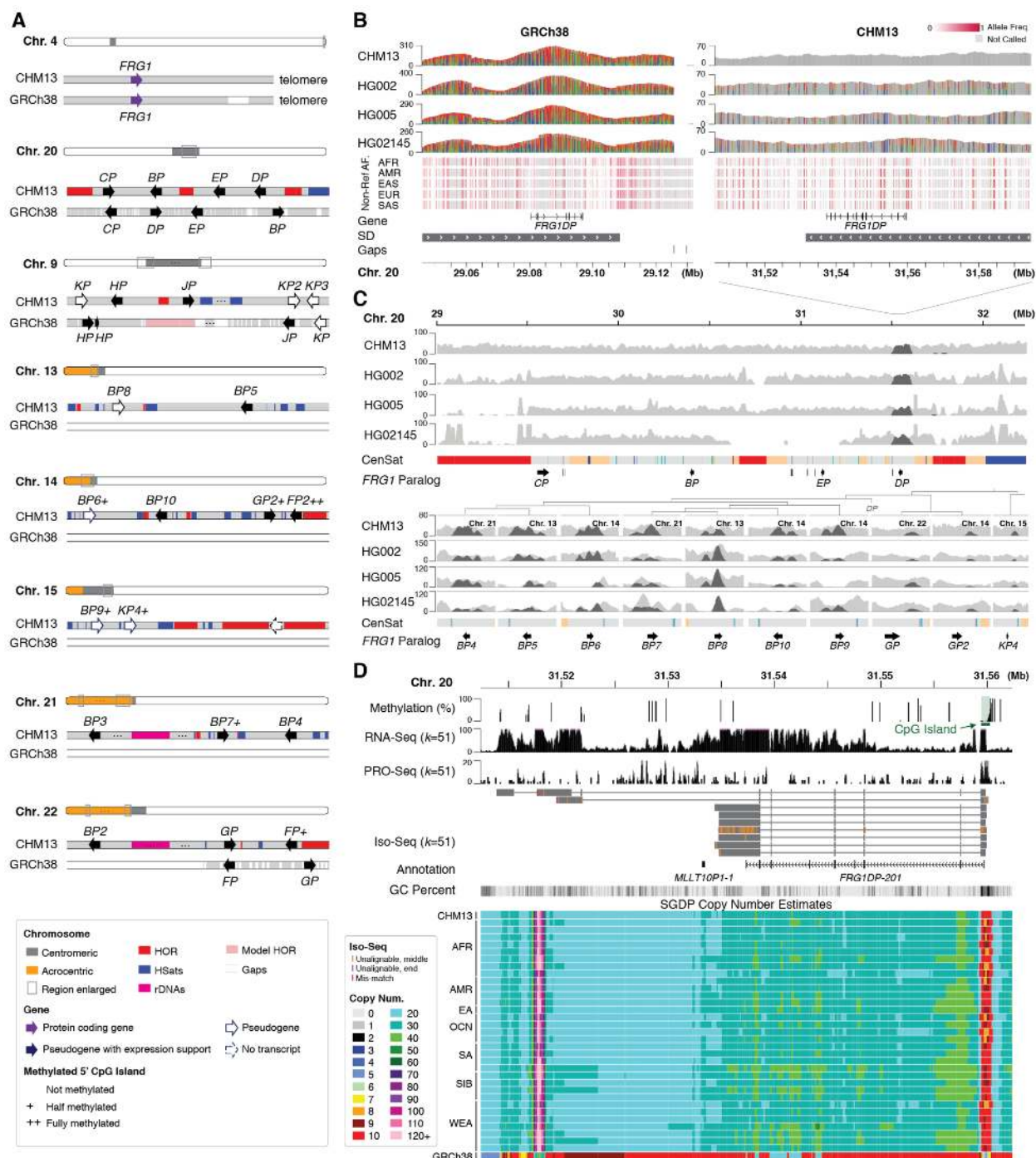


Fig. 5. Newly resolved *FRG1* paralogs in the T2T-CHM13 assembly. (A) Diagram of the segmentally duplicated protein-coding gene *FRG1* and its 23 paralogs located on multiple chromosomes of CHM13. Paralog names are presented without the “*FRG1*” prefix for brevity and genes are drawn larger than their actual size. All paralogs were found near centromeric satellite arrays, either adjacent to centromeres or on the acrocentric p-arms, with most copies having expression support and CpG islands present at the 5' start site with varying degrees of methylation (see legend). **(B)** HiFi read coverage and variants are shown for CHM13 and three other human samples mapped to the paralog *FRG1DP* in both GRCh38 and CHM13 references. When using GRCh38 as the reference, most of the region has excessive HiFi read coverage and non-reference variants, indicating that reads from the missing paralogs are mis-mapped to

this *FRG1DP* (only variants with >80% coverage shown). When CHM13 is used as the reference, HiFi read coverage follows the expected 34–40x coverage, with no variants for CHM13 (as expected) and a typical heterozygous variation pattern for the other three samples (all variants >20% coverage shown). These non-reference alleles were also found in other populations from 1000 Genomes Project ILMN data, confirming them as true polymorphic variants. **(C)** Mapped HiFi read coverage (light gray) from the same four samples is also shown for other *FRG1* paralogs in CHM13, with an extended context shown for Chromosome 20. Coverage of all HiFi reads that mapped to *FRG1DP* in GRCh38 are highlighted (dark gray), showing the various paralogous copies they actually originate from (*FRG1BP4–10*, *FRG1GP*, *FRG1GP2*, and *FRG1KP4*). Background read coverage shows copy number variation within the surrounding regions, but with relatively even coverage across *FRG1* paralogs, with the exception of *FRG1EP*, which appears to be absent from HG02145. Most other paralogs have diploid or haploid coverage levels, except *FRG1BP7* and *FRG1BP9*, which show higher coverage in some genomes, suggesting copy number polymorphism. **(D)** Methylation and expression results suggest transcription of *FRG1DP* in CHM13 (annotated as *FRG1DP-201*), with methylation depressed in the 5' CpG island, a number of well-aligned ISO-Seq reads, and both RNA-Seq and PRO-Seq evidence (again filtered with 51-mer unique markers to prevent mis-mapping). In the copy number display, each *k*-mer of the CHM13 assembly is painted with a color representing the copy number of that *k*-mer sequence in a sample from the Simons Genome Diversity Panel (SGDP) (85). The CHM13 and GRCh38 tracks show the copy number of these same *k*-mers in the respective assemblies. The region surrounding *FRG1DP* shows an average of 20–40 (diploid) *k*-mer copies in CHM13, with a diploid copy number of 10 in the first exon (red) and a highly repetitive sequence towards the tail (purple/pink). This matches all samples from the SGDP, whereas GRCh38 looks entirely unlike any sample and systematically underrepresents the true copy number. All displayed browser tracks are available from the T2T-CHM13 UCSC Assembly Hub.

Future of the human reference genome

The complete, telomere-to-telomere assembly of a human genome marks a new era of genomics where no region of the genome is beyond reach. Prior updates to the human reference genome have been incremental and the high cost of switching to a new assembly has outweighed the marginal gains for many researchers. In contrast, the T2T-CHM13 assembly presented here includes five entirely new chromosome arms and is the single largest addition of new content to the human genome 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. Such studies will necessarily require a complete and accurate human reference genome, ultimately driving adoption of the T2T-CHM13 assembly presented here.

One limitation of CHM13 is its lack of a Y chromosome. In order to finish a T2T reference sequence for all human chromosomes, we are in the process of sequencing and assembling the Y chromosome from the benchmark HG002 cell line, which has a 46,XY karyotype. The effectively haploid X and Y chromosomes can be assembled using the same methods developed here for the homozygous CHM13 genome. We have already sequenced, assembled, and validated the X chromosome from this line for use in the companion studies (27, 76) (Note S13, Figs. S37-39), and work on the HG002 Y chromosome is underway. Although much of the human Y chromosome is known to be heterochromatic and highly repetitive (86), we are confident that the strategy of combining HiFi and ONT reads used here will enable its completion (12).

Extending beyond the human reference genome, projects such as HapMap (87) and the 1000 Genomes Project (30) have been key to revealing genomic variation across human populations. However, technology limitations forced these early projects to focus mostly on a single reference genome and single-nucleotide variants. In contrast, long-read sequencing can recover the full spectrum of genomic variation including single-nucleotide variants, variable number short tandem repeats, and larger structural variants (88). Although short-read data cannot be confidently mapped to all regions of the human genome, reanalysis of 1000 Genomes short-read data (50) and copy number variation across a diverse set of human samples (26) has already shown that T2T-CHM13 constitutes a better reference even for short-read analyses. Improved computational methods and long-read omics assays are now needed to comprehensively survey polymorphic variation within the new regions of the genome assembled here, which we expect to reveal new phenotypic associations.

Highly accurate, long-read sequencing, combined with tailored algorithms, promises the *de novo* assembly of individual haplotypes and sequence-level resolution of complex structural variation. This will require the routine and complete *de novo* assembly of diploid human genomes, as planned by the Human Pangenome Reference Consortium (89). A collection of high-quality, complete reference haplotypes will transition the field away from a single linear reference and towards a reference pangenome that captures the full diversity of human genetic variation. Complete assembly of the CHM13 genome and our companion analyses have given only a small glimpse of the extensive structural variation that lies within the complete genome. Ideally, every genome could be assembled at the quality achieved here, since the small variants recovered by short-read resequencing approaches represent only a fraction of total genomic variation. However, taking the next step towards the telomere-to-telomere assembly of heterozygous diploid genomes, and further automating this process, presents a difficult challenge that will require the continued improvement of sequencing and assembly technologies. Until this ultimate goal is realized, and every genome can be completely sequenced without error, the T2T-CHM13 assembly represents a more complete, representative, and accurate reference than GRCh38, and we suggest it succeed GRCh38 in all studies requiring a linear reference sequence.

Data and materials availability

Sequencing data and assemblies (NCBI BioProject PRJNA559484):

<https://www.ncbi.nlm.nih.gov/bioproject/559484>

Sequencing data, assemblies, and other supporting data on AWS:

<https://github.com/marbl/CHM13>

Assembly issues and known heterozygous sites:

<https://github.com/marbl/CHM13-issues>

UCSC assembly hub browser:

<http://genome.ucsc.edu/cgi-bin/hgTracks?genome=t2t-chm13-v1.0&hubUrl=http://t2t.gi.ucsc.edu/chm13/hub/hub.txt>

Dotplot visualization and browser:

<https://resgen.io/paper-data/T2T-Nurk-et-al-2021/views/t2t-identity>

T2T Consortium homepage:

<https://sites.google.com/ucsc.edu/t2tworkinggroup>

Supplementary materials

SupplementaryMaterials.pdf

- Notes S1–S13
- Figs. S1–S39
- Tables S3-4, S13–S16

SupplementaryTables.xlsx

- Tables S1, S2, S5–S12

Acknowledgements

We would like to thank Mark Akeson, Andrew Carroll, Pi-Chuan Chang, Arthur Delcher, Maria Nattestad, and Mihai 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.

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*, WC, LGL, TD, GF, AF, KH, CJ, EDJ, DP, VAS, KS, 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*, 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.

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); 1U01HG010971 (EEE, HL, KHM, MK, RSF, TAGL); NIH/NHGRI R01HG002385, R01HG010169 (EEE); NIH/NHGRI R01HG009190 (AG, WT); NIH/NHGRI R01HG010485, U01HG010961 (BP, EG, KS); NIH/NHGRI U41HG010972 (IMH, BP, EG, KS); NSF DBI-1627442, NSF IOS-1732253, NSF IOS-1758800, NHGRI U24HG006620, NCI U01CA253481, NIDDK R24 DK106766-01A1 (MCS); NIH/NHGRI U24HG010263 (SZ, MCS); Mark Foundation for Cancer Research 19-033-ASP (SA, MCS); NIH/NHGRI R01HG006677 (ASh, SLS); NIH/NHGRI 5U24HG009081-03 (MK, 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/NHGRI 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/NHGRI F31HG011205 (CJS); Fulbright Fellowship (DCS); HHMI (EDJ, GF); Ministry of Science and Higher Education of the RF 075-10-2020-116 / 13.1902.21.0023 (EIR); NIH UM1 HG008898 (FJS); NIH R01GM123312-02, NSF1613806 (GAH, PGSG, SJH, RJO); NIH R21CA240199, NSF 643825, Connecticut Innovations 20190200 (RJO); NIGMS F32GM134558 (GAL); NHGRI 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/NHGRI R011R01HG011274-01 (KHM); Sirius University (LU); RSF 19-75-30039 Analysis of genomic repeats (LU, IAA); NIH/NHGRI U41HG007234 (MD); NIH/OD/NIMH DP2OD025824 (MYD); HHMI Hanna H. Gray Fellowship (NA); NIH/NIGMS 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; FJS has received travel funds to speak at events hosted by Pacific Biosciences; SK and FJS 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).

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 Genome Center, MIND Institute, Department of Biochemistry and Molecular Medicine, University of California, Davis, CA, USA
- 21 Department of Genetics, University of Cambridge, Cambridge, UK
- 22 Inscripta, Boulder, CO, USA
- 23 Laboratory of Neurogenetics of Language and The Vertebrate Genome Lab, The Rockefeller University, New York, NY, USA
- 24 Howard Hughes Medical Institute, Chevy Chase, MD, USA
- 25 Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA
- 26 University of Tennessee Health Science Center, Memphis, TN, USA
- 27 McDonnell Genome Institute, Washington University in St. Louis, MO, USA
- 28 Department of Genetics, Yale University School of Medicine, New Haven, CT, USA
- 29 Comparative Genomics Analysis Unit, Cancer Genetics and Comparative Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
- 30 Pacific Biosciences, Menlo Park, CA, USA
- 31 Department of Computational and Data Sciences, Indian Institute of Science, Bangalore KA, India
- 32 Reservoir Genomics LLC, Oakland, CA
- 33 Department of Computer Science and Engineering, UC San Diego, CA, USA
- 34 Undiagnosed Diseases Program, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
- 35 Heinrich Heine University Düsseldorf, Medical Faculty, Institute for Medical Biometry and Bioinformatics, Düsseldorf, Germany
- 36 Biosystems and Biomaterials Division, National Institute of Standards and Technology, Gaithersburg, MD, USA
- 37 Department of Pediatrics, Division of Genetic Medicine, University of Washington and Seattle Children's Hospital, Seattle, WA, USA
- 38 Max-Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
- 39 Department of Psychiatry, University of Massachusetts Medical School, Worcester, MA, USA
- 40 Faculty of Biology, Lomonosov Moscow State University, Moscow, Russia
- 41 Cancer Institute of New Jersey, New Brunswick, NJ, USA
- 42 Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA
- 43 National Center for Biotechnology Information, National Library of Medicine, NIH, Bethesda, MD, USA
- 44 Human Genome Sequencing Center, Baylor College of Medicine, Houston TX, 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

References

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).
2. International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature.* **409**, 860–921 (2001).
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. Hancherhalli, 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. Ferreira, 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. Yoosheph, 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).
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. 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).
5. J. D. McPherson, M. Marra, L. Hillier, R. H. Waterston, A. Chinwalla, J. Wallis, M. Sekhon, K. Wylie, E. R. Mardis, R. K. Wilson, R. Fulton, T. A. Kucaba, C. Wagner-McPherson, W. B. Barbazuk, S. G. Gregory, S. J. Humphray, L. French, R. S. Evans, G. Bethel, A. Whittaker, J. L. Holden, O. T.

- McCann, A. Dunham, C. Soderlund, C. E. Scott, D. R. Bentley, G. Schuler, H. C. Chen, W. Jang, E. D. Green, J. R. Idol, V. V. Maduro, K. T. Montgomery, E. Lee, A. Miller, S. Emerling, Kucherlapati, R. Gibbs, S. Scherer, J. H. Gorrell, E. Sodergren, K. Clerc-Blankenburg, P. Tabor, S. Naylor, D. Garcia, P. J. de Jong, J. J. Catanese, N. Nowak, K. Osoegawa, S. Qin, L. Rowen, A. Madan, M. Dors, L. Hood, B. Trask, C. Friedman, H. Massa, V. G. Cheung, I. R. Kirsch, T. Reid, R. Yonescu, J. Weissenbach, T. Bruls, R. Heilig, E. Branscomb, A. Olsen, N. Doggett, J. F. Cheng, T. Hawkins, R. M. Myers, J. Shang, L. Ramirez, J. Schmutz, O. Velasquez, K. Dixon, N. E. Stone, D. R. Cox, D. Haussler, W. J. Kent, T. Furey, S. Rogic, S. Kennedy, S. Jones, A. Rosenthal, G. Wen, M. Schilhabel, G. Gloeckner, G. Nyakatura, R. Siebert, B. Schlegelberger, J. Korenberg, X. N. Chen, A. Fujiyama, M. Hattori, A. Toyoda, T. Yada, H. S. Park, Y. Sakaki, N. Shimizu, S. Asakawa, K. Kawasaki, T. Sasaki, A. Shintani, A. Shimizu, K. Shibuya, J. Kudoh, S. Minoshima, J. Ramser, P. Seranski, C. Hoff, A. Poustka, R. Reinhardt, H. Lehrach, International Human Genome Mapping Consortium, A physical map of the human genome. *Nature*. **409**, 934–941 (2001).
6. 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).
 7. 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).
 8. M. Gupta, A. R. Dhanasekaran, K. J. Gardiner, Mouse models of Down syndrome: gene content and consequences. *Mamm. Genome.* **27**, 538–555 (2016).
 9. 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).
 10. International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. *Nature*. **431**, 931–945 (2004).
 11. 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).
 12. G. A. Logsdon, M. R. Vollger, P. Hsieh, Y. Mao, M. A. Liskovych, 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).
 13. E. E. Eichler, U. Surti, R. Ophoff, 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>).
 14. 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. Veceli, 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).
15. 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).
 16. 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).
 17. 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).
 18. 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).
 19. A. V. Bzikadze, P. A. Pevzner, Automated assembly of centromeres from ultra-long error-prone reads. *Nat. Biotechnol.* **38**, 1309–1316 (2020).
 20. A. M. Wenger, P. Peluso, W. J. Rowell, P.-C. Chang, R. J. Hall, G. T. Concepcion, J. Ebler, A. Functamasan, 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).
 21. G. A. Logsdon, M. R. Vollger, E. E. Eichler, Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
 22. 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).
 23. 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).
 24. 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 Research.* **27** (2017), pp. 677–685.
 25. B. Gel, E. Serra, karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics.* **33**, 3088–3090 (2017).
 26. M. R. Vollger, X. Guitart, P. C. Dishuck, L. Mercuri, W. T. Harvey, A. Gershman, M. Diekhans, A. Sulovari, K. M. Munson, A. M. 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. *bioRxiv* (2021).
 27. N. Altomose, *et al.*, Genetic and epigenetic maps of endogenous human centromeres. *bioRxiv (to appear)* (2021).
 28. 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).
29. 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.* (2019), doi:10.1111/ahg.12364.
30. 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbek, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, A global reference for human genetic variation. *Nature.* **526**, 68–74 (2015).
31. E. W. Myers, The fragment assembly string graph. *Bioinformatics.* **21 Suppl 2**, ii79–85 (2005).
32. H. Li, Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics.* **32**, 2103–2110 (2016).
33. 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).
34. A. Šípek Jr, R. Mihalová, A. Panczak, L. Hřčková, M. Janashia, N. Kaspřiková, M. Kohoutová, Heterochromatin variants in human karyotypes: a possible association with reproductive failure. *Reprod. Biomed. Online.* **29**, 245–250 (2014).
35. M. Rautiainen, T. Marschall, GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol.* **21**, 253 (2020).
36. R. R. Wick, M. B. Schultz, J. Zobel, K. E. Holt, Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics.* **31**, 3350–3352 (2015).
37. A. M. Phillippy, M. C. Schatz, M. Pop, Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol.* **9**, R55 (2008).
38. C. Jain, A. Rhie, N. Hansen, S. Koren, A. M. Phillippy, A long read mapping method for highly repetitive reference sequences. *bioRxiv* (2020), p. 2020.11.01.363887.
39. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013), (available at <http://arxiv.org/abs/1303.3997>).
40. 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).
41. 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).
42. K. Shafin, T. Pesout, P. C. Chang, M. Nattestad, Haplotype-aware variant calling enables high accuracy in nanopore long-reads using deep neural networks. *bioRxiv* (2021) (available at <https://www.biorxiv.org/content/10.1101/2021.03.04.433952v1.abstract>).
43. G. Formenti, A. Rhie, B. P. Walenz, F. Thibaud-Nissen, S. Koren, E. Myers, E. D. Jarvis, A. M. Phillippy, Merfin: improved variant filtering and polishing via k-mer validation. *bioRxiv (to appear)* (2021).

44. A. M. McCartney, *et al.*, Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *bioRxiv (to appear)* (2021).
45. 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).
46. 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).
47. 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).
48. 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).
49. 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).
50. S. Aganezov, *et al.*, A complete human reference genome improves variant calling for population and clinical genomics. *bioRxiv (to appear)* (2021).
51. 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).
52. D. M. Stults, M. W. Killen, H. H. Pierce, A. J. Pierce, Genomic architecture and inheritance of human ribosomal RNA gene clusters. *Genome Res.* **18**, 13–18 (2008).
53. J. O. Nelson, G. J. Watase, N. Warsinger-Pepe, Y. M. Yamashita, Mechanisms of rDNA Copy Number Maintenance. *Trends Genet.* **35**, 734–742 (2019).
54. M. Rautiainen, T. Marschall, MBG: Minimizer-based Sparse de Bruijn Graph Construction. *Bioinformatics* (2021), doi:10.1093/bioinformatics/btab004.
55. Smit AFA, Hubley R, Green, P, *RepeatMasker Open-4.0* (2015; <http://www.repeatmasker.org>).
56. S. J. Hoyt, *et al.*, From telomere to telomere: characterizing the transcriptional and epigenetic state of repeat elements. *bioRxiv (to appear)* (2021).
57. 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).
58. A. Shumate, S. L. Salzberg, Liftoff: accurate mapping of gene annotations. *Bioinformatics* (2020), doi:10.1093/bioinformatics/btaa1016.
59. 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. Zerbin, 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).
60. 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).
 61. A. S. Henderson, D. Warburton, K. C. Atwood, Location of ribosomal DNA in the human chromosome complement. *Proc. Natl. Acad. Sci. U. S. A.* **69**, 3394–3398 (1972).
 62. H. E. Trowell, A. Nagy, B. Vissel, K. H. Choo, Long-range analyses of the centromeric regions of human chromosomes 13, 14 and 21: identification of a narrow domain containing two key centromeric DNA elements. *Hum. Mol. Genet.* **2**, 1639–1649 (1993).
 63. R. Lyle, P. Prandini, K. Osoegawa, B. ten Hallers, S. Humphray, B. Zhu, E. Eyraas, 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).
 64. 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).
 65. 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).
 66. O. V. Iarovaia, E. P. Minina, E. V. Sheval, D. Onichtchouk, S. Dokudovskaya, S. V. Razin, Y. S. Vassetzky, Nucleolus: A Central Hub for Nuclear Functions. *Trends Cell Biol.* **29**, 647–659 (2019).
 67. M. S. Lindström, D. Jurada, S. Bursac, I. Orsolich, J. Bartek, S. Volarevic, Nucleolus as an emerging hub in maintenance of genome stability and cancer pathogenesis. *Oncogene.* **37**, 2351–2366 (2018).
 68. K. R. Kampen, S. O. Sulima, S. Vereecke, K. De Keersmaecker, Hallmarks of ribosomopathies. *Nucleic Acids Res.* **48**, 1013–1028 (2020).
 69. M. Jarmuz-Szymczak, J. Janiszewska, K. Szyfter, L. G. Shaffer, Narrowing the localization of the region breakpoint in most frequent Robertsonian translocations. *Chromosome Res.* **22**, 517–532 (2014).
 70. S. E. Antonarakis, B. G. Skotko, M. S. Rafii, A. Strydom, S. E. Pape, D. W. Bianchi, S. L. Sherman, R. H. Reeves, Down syndrome. *Nat Rev Dis Primers.* **6**, 9 (2020).
 71. 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).
 72. 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).
 73. 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).
 74. 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).
 75. A. L. Jørgensen, S. Kølvråa, 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).
76. A. Gershman, M. Sauria, P. W. Hook, S. Hoyt, R. Razaghi, S. Koren, N. Altomose, G. V. Caldas, M. R. Vollger, G. A. Logsdon, A. Rhie, E. E. Eichler, M. C. Schatz, R. O'Neill, A. M. Phillippy, K. H. Miga, W. Timp, Epigenetic patterns in a complete human genome. *bioRxiv* (2021).
 77. H. Kwak, N. J. Fuda, L. J. Core, J. T. Lis, Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science*. **339**, 950–953 (2013).
 78. P. Kerpedjiev, N. Abdennur, F. Lekschas, C. McCallum, K. Dinkla, H. Strobelt, J. M. Lubner, 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).
 79. M. Richards, F. Coppée, N. Thomas, A. Belayew, M. Upadhyaya, Facioscapulohumeral muscular dystrophy (FSHD): an enigma unravelled? *Hum. Genet.* **131**, 325–340 (2012).
 80. G. Ferri, C. H. Huichalaf, R. Caccia, D. Gabellini, Direct interplay between two candidate genes in FSHD muscular dystrophy. *Hum. Mol. Genet.* **24**, 1256–1266 (2015).
 81. 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, Chromosome 4q DNA rearrangements associated with facioscapulohumeral muscular dystrophy. *Nat. Genet.* **2**, 26–30 (1992).
 82. 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).
 83. 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).
 84. S. van Koningsbruggen, K. R. Straasheijm, E. Sterrenburg, N. de Graaf, H. G. Dauwerse, R. R. Frants, S. M. van der Maarel, FRG1P-mediated aggregation of proteins involved in pre-mRNA processing. *Chromosoma*. **116**, 53–64 (2007).
 85. 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).
 86. N. Altomose, K. H. Miga, M. Maggioni, H. F. Willard, Genomic characterization of large heterochromatic gaps in the human genome assembly. *PLoS Comput. Biol.* **10**, e1003628 (2014).
 87. International HapMap Consortium, A haplotype map of the human genome. *Nature*. **437**, 1299–1320 (2005).
 88. P. Ebert, P. A. Audano, Q. Zhu, B. Rodriguez-Martin, D. Porubsky, M. J. Bonder, A. Sulovari, J. Ebler, W. Zhou, R. Serra Mari, F. Yilmaz, X. Zhao, P. Hsieh, J. Lee, S. Kumar, J. Lin, T. Rausch, Y. Chen, J. Ren, M. Santamarina, W. Höps, H. Ashraf, N. T. Chuang, X. Yang, K. M. Munson, A. P. Lewis, S. Fairley, L. J. Tallon, W. E. Clarke, A. O. Basile, M. Byrska-Bishop, A. Corvelo, U. S. Evani, T.-Y. Lu,

M. J. P. Chaisson, J. Chen, C. Li, H. Brand, A. M. Wenger, M. Ghareghani, W. T. Harvey, B. Raeder, P. Hasenfeld, A. A. Regier, H. J. Abel, I. M. Hall, P. Flicek, O. Stegle, M. B. Gerstein, J. M. C. Tubio, Z. Mu, Y. I. Li, X. Shi, A. R. Hastie, K. Ye, Z. Chong, A. D. Sanders, M. C. Zody, M. E. Talkowski, R. E. Mills, S. E. Devine, C. Lee, J. O. Korb, T. Marschall, E. E. Eichler, Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*. **372** (2021), doi:10.1126/science.abf7117.

89. K. H. Miga, T. Wang, The Need for a Human Pangenome Reference Sequence. *Annu. Rev. Genomics Hum. Genet.* (2021), doi:10.1146/annurev-genom-120120-081921.