Article

A global view of human centromere variation and evolution

https://doi.org/10.1038/s41586-026-10841-9

Received: 10 December 2025

Accepted: 24 June 2026

Published online: 29 July 2026

Open access

Check for updates

Shenghan Gao[1,114], Keisuke K. Oshima[1,114], Shu-Cheng Chuang[1], Mark Loftus[2,3,4], Tamara A. Potapova[5], Annalaura Montanari[6], David S. Gordon[7], Zikun Yang[8,9], Human Genome Structural Variation Consortium*, Human Pangenome Reference Consortium*, Yafei Mao[8,9,10,11], PingHsun Hsieh[7,12], Jennifer L. Gerton[5,13], Miriam K. Konkel[2,3], Mario Ventura[6] & Glennis A. Logsdon[1]✉

Centromeres are essential chromosomal regions that ensure accurate chromosome segregation during cell division, yet their highly repetitive sequence has historically hindered their complete assembly and characterization[1]. Consequently, the full spectrum of centromere diversity across individuals, populations and evolutionary contexts remains largely unexplored. Here we address this gap in knowledge by assembling and characterizing 2,110 centromeres from diverse individuals representing 5 continental and 28 population groups. Using bioinformatic tools tailored for centromeres, we identify variation, including 226 centromere haplotypes and 1,870 α-satellite higher-order repeat variants. While most centromeres have a single kinetochore site, we find that around 6% have di-kinetochores, and less than 1% have tri-kinetochores, which we confirm using long-read chromatin profiling and multigenerational inheritance. We also show that kinetochore position is closely associated with the underlying sequence and structure of the centromere. To understand the nature of evolutionary change, we compared these centromeres to 5,747 centromeres assembled by the Human Pangenome Reference Consortium. We show that centromeres have a 20-fold variation in mutation rate, and a subset of centromeres has evidence of archaic hominin introgression. We validate these mutation rates in a 4-generation, 28-member family and show that the kinetochore site is the most rapidly mutating region in the centromere. We propose a model that reveals an 'arms race' between centromeric sequence and proteins, with frequent mutations within the kinetochore site that lead to changes in genetic and epigenetic landscapes and, ultimately, rapid evolution of these critically important regions.

Centromeres are essential chromosomal loci that mediate chromosome segregation during mitosis and meiosis. Despite their fundamental importance, their sequence, structure and variation among the human population are not yet fully known. Centromeres have been largely excluded from variation analyses because they are composed of near-identical tandem repeats known as α-satellites[2–5], which challenge efforts to assess their variation using standard methods and approaches. However, recent advances in long-read sequencing and telomere-to-telomere genome assembly algorithms are now making it possible to resolve entire centromeric regions[6–12]. Building on these technological developments, population-scale genome projects have begun to reveal that centromeres have far more genetic, epigenetic and structural diversity than previously appreciated[13–15]. Variation in α-satellite higher-order repeat (HOR) array length, composition, structural organization and kinetochore position has been identified in limited cohorts[16], but a comprehensive understanding of the full scope of human centromere variation among global populations remains unresolved. Moreover, fundamental questions about how centromeres evolve, how quickly they mutate and how sequence and chromatin features interact to specify kinetochore sites remain largely unclear.

Here we address this knowledge gap by assembling and characterizing 2,110 complete centromeres from 65 individuals representing

[1]Department of Genetics, Epigenetics Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [2]Department of Genetics & Biochemistry, Clemson University, Clemson, SC, USA. [3]Institute for Human Genetics, Clemson University, Greenwood, SC, USA. [4]The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA. [5]Stowers Institute for Medical Research, Kansas City, MO, USA. [6]Department of Biosciences, Biotechnology and Environment, University of Bari Aldo Moro, Bari, Italy. [7]Department of Genetics, Cell Biology, and Development, University of Minnesota, Twin Cities, MN, USA. [8]Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Ministry of Education, Shanghai Jiao Tong University, Shanghai, China. [9]Zhiyuan College, Shanghai Jiao Tong University, Shanghai, China. [10]Shanghai Key Laboratory of Embryo Original Diseases, International Peace Maternity and Child Health Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China. [11]Center for Comparative Biomedicine, Ministry of Education Key Laboratory of Systems Biomedicine, State Key Laboratory of Medical Genomics, Institute of Systems Biomedicine, Shanghai Jiao Tong University, Shanghai, China. [12]Institute for Health Informatics, University of Minnesota, Twin Cities, MN, USA. [13]Department of Biochemistry and Molecular Biology, University of Kansas Medical Center, Kansas City, KS, USA. [114]These authors contributed equally: Shenghan Gao, Keisuke K. Oshima. *Lists of authors and their affiliations appear at the end of the paper. ✉e-mail: glogsdon@pennmedicine.upenn.edu

5 continental and 28 population groups and subsequently validating our findings with a broader set of 6,230 centromeres from 227 diverse individuals and a 28-member pedigree (8,340 centromeres in total). By developing computational tools and pipelines optimized for centromeric regions, we define the landscape of centromere sequence and structural variation, identify haplotypes and α-satellite HOR variants, and map kinetochore positions across all chromosomes. Furthermore, we integrate long-read methylation data, CENP-A chromatin profiling, multigenerational inheritance and archaic sequencing data to characterize the functional and evolutionary events shaping centromeres. Together, these analyses provide a comprehensive foundation for understanding how human centromeres vary and evolve and how their underlying architecture supports faithful chromosome segregation during cell division.

Complete sequence of 2,110 centromeres

The recent release of 65 near-complete human genome assemblies spanning 5 continental and 28 population groups by the Human Genome Structural Variation Consortium[13] (HGSVC; Supplementary Table 1) provides a unique opportunity to investigate the diversity in human centromere sequence and structure, variation in epigenetic landscape and differences in evolutionary trajectories on a global scale. Therefore, we first sought to resolve the sequence of each centromere in these 65 genomes. To do this, we identified centromeres that were completely assembled from p to q arm and free of large sequence assembly errors (Methods), identifying 1,246 such centromeres[13].

Recognizing that centromeres generated by different genome assemblers (such as Verkko[17] and hifiasm[18]) have complementary error profiles that could be leveraged to repair one assembly with the other[13] (Supplementary Fig. 1), we then developed a tool, AssemblyRepairer, to repair the remaining errors in the Verkko centromeres with the corresponding centromeres assembled with hifiasm. AssemblyRepairer uses large, unique *k*-mers shared between the Verkko and hifiasm assemblies to identify corresponding contigs and then patches erroneous regions in the Verkko contigs with the correctly assembled, equivalent sequences in the hifiasm contigs (Methods and Supplementary Fig. 2a). Using this tool, we repaired an additional 864 centromeres, for a total combined set of 2,110 completely and accurately assembled centromeres (Supplementary Fig. 2b–d)—a 67% increase over the previously published dataset[13]. We confirmed the accurate assembly of each centromere with a suite of computational tools and approaches, including native long-read mapping as well as validation with HMM-Flagger[19], NucFlag[15] and GAVISUNK[20] (Supplementary Note 1 and Supplementary Table 2). Moreover, we estimated the quality value (QV) of the centromeres using short-read Illumina sequencing data generated from the same source genome with Merqury[21] and found them to have an average QV of 60.99, indicating that the centromere assemblies were over 99.9999% accurate (Supplementary Table 3).

The complete sequence of 2,110 diverse human centromeres allowed us to investigate their genetic and epigenetic variation and determine their haplotype frequency across global populations. To do this, we developed a centromere mapping and annotation pipeline (CenMAP), which automatically annotates centromeric regions at the sequence level to define the active and inactive centromeric α-satellite HOR arrays as well as divergent–monomeric α-satellites, determines the location of hypomethylated regions in the centromere known as centromere dip regions[22] (CDRs), which are associated with the site of the kinetochore[7,9], and generates local sequence identity heat maps to show the evolutionary relationship of the underlying sequences (Fig. 1 and Supplementary Figs. 3–6). Applying this pipeline to all 2,110 centromeres revealed 226 new major centromere haplotypes (Fig. 1 and Supplementary Figs. 3–6) and 1,870 new α-satellite HOR variants compared to CHM13 and CHM1 (Supplementary Table 4), all of which occur at least twice in the genome they were identified in, and 97% of which occur in at least one other genome in our dataset or in the genome assemblies released by the Human Pangenome Reference Consortium (HPRC; Methods and Supplementary Table 5). Moreover, comparison of these newly assembled centromeres to those in the CHM1 and CHM13 genomes revealed that, for 7 out of 24 chromosomes (chromosomes 1, 3, 4, 6, 10, 11 and 13), the most common centromere haplotype was not represented by CHM13 or CHM1 and was instead represented by another haplotype (Fig. 1 and Supplementary Figs. 3–6), underscoring the benefit of additional genomes for a more comprehensive understanding of centromere variation.

Global diversity among human centromeres

Using these complete centromere assemblies, we next sought to determine the extent of variation among human centromeres. We first investigated the variation in the length of α-satellite HOR arrays. Human centromeres are typically composed of one or more arrays of tandemly repeated α-satellite HORs, with one array serving as the site of the kinetochore. We therefore distinguished between kinetochore-forming and non-kinetochore-forming arrays, marked by the presence or absence of the CDR, respectively, and determined the length of each of these in our dataset (Supplementary Figs. 7 and 8, Supplementary Table 6 and Supplementary Note 2). We found that kinetochore-forming arrays are, on average, 2.5 Mb long, with the smallest array only 239 kb in length (on chromosome 21) and the largest 7.5 Mb in length (on chromosome 18; Fig. 2a, Supplementary Fig. 7a and Supplementary Table 6). Two chromosomes (chromosomes 1 and 18) have significantly larger kinetochore-forming arrays than the other chromosomes, with a mean length of 4.3 and 4.2 Mb, respectively, while seven chromosomes (chromosomes 3, 4, 13, 15, 16, 21 and Y) have significantly smaller kinetochore-forming arrays, with a mean length of 1.7, 1.5, 1.6, 1.5, 1.8, 1.4 and 0.7 Mb, respectively ($P < 0.05$, two-tailed Wilcoxson rank-sum test). The greatest variation in length occurs on chromosome 6, with a 21.5-fold difference between the smallest and largest, while the greatest absolute difference in length occurs on chromosome 18, with a 6.3 Mb difference (Fig. 2a, Supplementary Fig. 7a and Supplementary Table 6). Comparison of the lengths of kinetochore-forming arrays between haplotypes of the same individual reveals that they are, on average, only 1.5-fold different in length, with <1% of arrays more than fourfold larger than their counterpart (Supplementary Fig. 9 and Supplementary Table 6).

Across populations, we find that nine chromosomes (chromosomes 3, 7, 9, 16, 17, 19, 21, 22 and X) have significantly different kinetochore-forming α-satellite HOR array lengths (Supplementary Figs. 7b and 8 and Supplementary Table 6). For chromosome 21, for example, we find that individuals with African ancestry have α-satellite HOR arrays that are 353 kb larger, on average, than those with non-African ancestry. However, the opposite is true for chromosomes 3, 7, 9, 16, 17, 19, 22 and X, in which individuals with African ancestry have arrays that are 281, 522, 348, 180, 476, 654, 491 and 594 kb smaller, on average, than those with non-African ancestry, respectively (Supplementary Figs. 7b and 8 and Supplementary Table 6). For the remaining chromosomes, the kinetochore-forming α-satellite HOR array lengths are similar across populations, showing no significant correlation with genetic ancestry (Supplementary Figs. 7b and 8 and Supplementary Table 6).

In addition to variation in α-satellite HOR array length, we observed both large- and small-scale variation impacting the sequence, structure and organization of α-satellite HOR arrays (Fig. 2b–h and Supplementary Fig. 10). We identified ten chromosomes with large duplications, deletions or inversions (hundreds of kb long) that alter the number of α-satellite HOR arrays and/or the organization within them (chromosomes 1, 4, 5, 7, 10, 12, 13, 17, 19 and 20; Fig. 2b–d and Supplementary Fig. 10). This is perhaps best exemplified by chromosome 1, where we found that a quarter of the individuals in our



Fig. 1 | Complete maps and haplotype frequencies of diverse human centromeres. The sequence, structure, methylation pattern, local sequence identity and haplotype frequency of 2,156 completely assembled centromeres from CHM13 (ref. 9), CHM1 (ref. 10) and 65 diverse human genomes[13] (the brackets on the left apply to all columns). The α-satellite HORs are coloured by the number of α-satellite monomers within them, and the orientation of the active α-satellite HOR arrays is indicated by an arrow. The site of the putative kinetochore, marked by the CDR, is indicated by a black bar. The local sequence identity across each centromeric region is indicated by a one-dimensional heat map, and MEIs within α-satellite HOR arrays are indicated by coloured triangles. Haplotype frequencies are shown as pie charts and are coloured by population group (Supplementary Table 1). The most common haplotype for each chromosome is marked by an asterisk. All 226 new major haplotypes are shown in Supplementary Fig. 3. Magnified views of each haplotype and their pie chart are also shown in Supplementary Figs. 4–6. All centromeres can be viewed at https://www.cencyclopedia.com—an interactive and dynamic centromere browser. div., divergent; freq., frequency; methyl., methylation; mon., monomeric.

dataset have a large-scale inversion in the core of the *D1Z7* α-satellite HOR array (Supplementary Note 3) spanning 637 kb to 2.47 Mb (like CHM13 (ref. 9) and CHM1 (ref. 10)); however, the majority (68%) has a single, directly-oriented *D1Z7* α-satellite HOR array, and another 7% has a duplication or triplication of a portion of the *D1Z7* α-satellite HOR array, resulting in two or three distinct arrays spanning 631 kb to 2.81 Mb each (Fig. 2b and Supplementary Fig. 10). We also find that chromosomes 3 and 4 have multiple α-satellite HOR arrays that are interrupted by stretches of human satellite 1A (HSat1A), as previously reported[8,9]. Whereas 97% of chromosome 3 centromeres in our dataset have only two *D3Z1* α-satellite HOR arrays, we find that a much smaller proportion of individuals (56%) have two *D4Z1* α-satellite HOR arrays on chromosome 4, while the rest have three, four or five arrays (23%, 18% and 3%, respectively; Fig. 2c,d). For chromosomes 5, 7, 12, 13, 17, 19 and 20, almost all individuals in our dataset have a single active α-satellite HOR array; however, at least one individual has a duplication of the active α-satellite HOR array and flanking divergent α-satellite HORs, resulting in two or more distinct α-satellite HOR arrays (Supplementary Fig. 10). Finally, chromosomes 10 and 17 often have small active α-satellite HOR arrays that flank the larger α-satellite HOR array, occurring in 100% and 64% of individuals, respectively (Supplementary Fig. 10).

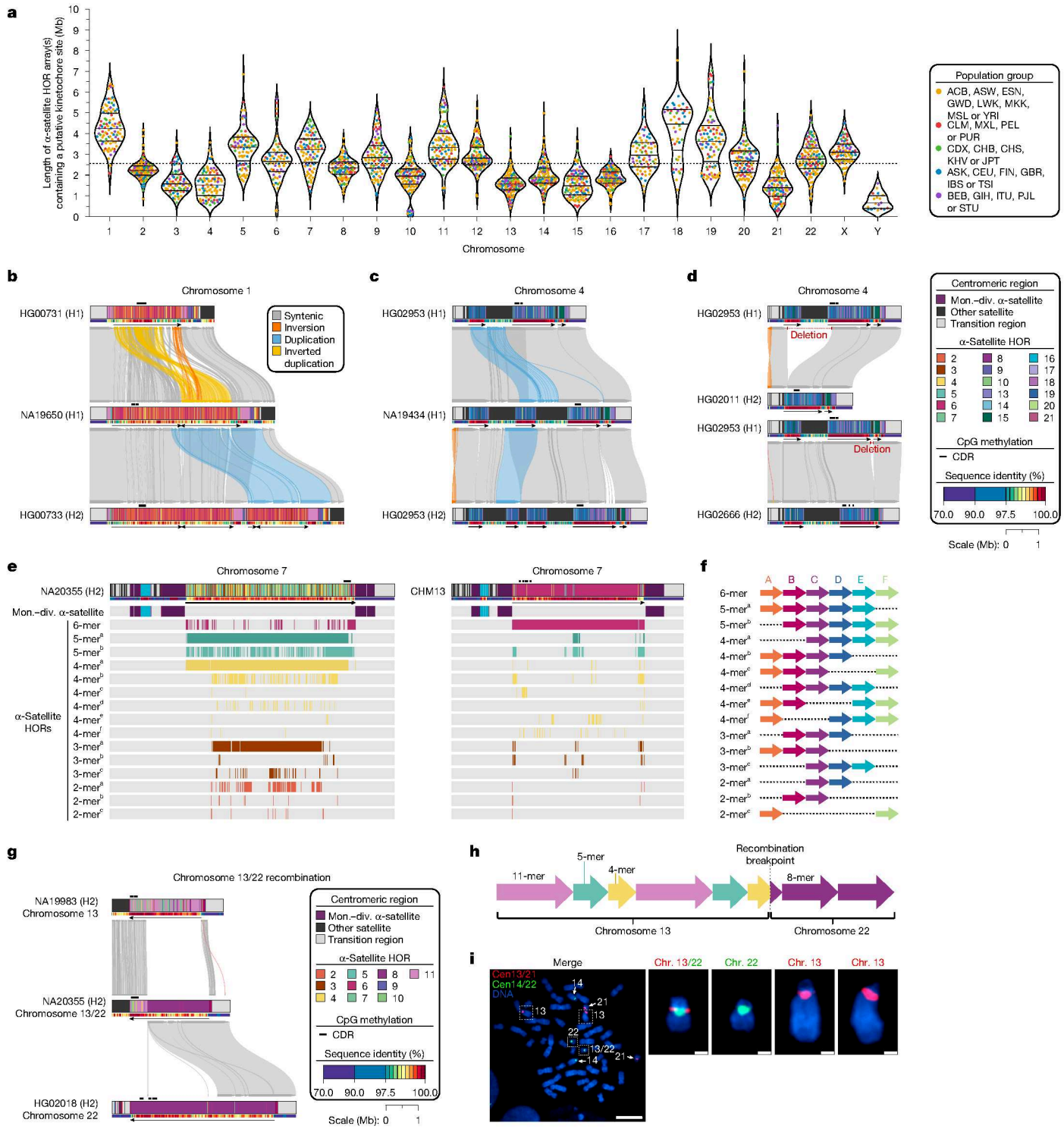


Fig. 2 | Complex sequence and structural variation among human centromeres. a, Variation in the length of α-satellite HOR arrays containing a putative kinetochore site, defined by the presence of a CDR, among 2,110 diverse human centromeres. For each chromosome, the mean (solid line) and first and third quartiles (dotted lines) are shown. The mean across all chromosomes (dashed line) is also shown. **b**–**d**, Examples of large-scale (hundreds of kb) structural variation among centromeres, including inversions, duplications, inverted duplications and deletions within the chromosome 1 *D1Z7* α-satellite HOR array (**b**) and chromosome 4 *D4Z1* α-satellite HOR arrays (**c**,**d**). **e**, Differences in the chromosome 7 *D7Z1* α-satellite HOR array organization between the NA19238 (left) and CHM13 (right) genomes. Different α-satellite HOR variants are labelled uniquely. **f**, The structure of α-satellite HOR variants present in the NA19238 (H2) and CHM13 chromosome 7 centromeres. All of them are derived from an ancestral 6-mer α-satellite HOR and are labelled uniquely. **g**, A recombination event between the chromosome 13 *D13Z1* α-satellite HOR array and chromosome 22 *D22Z1* α-satellite HOR arrays gives rise to a chromosome 13/22 hybrid α-satellite HOR array. The top and bottom centromeres are not genetically related to the hybrid centromere. **h**, Schematic of the breakpoint between the *D13Z1* and *D22Z1* α-satellite HOR arrays. The recombination event creates a new α-satellite HOR variant five monomers in length. **i**, FISH analysis of a GM20355 metaphase chromosome (Chr.) spread showing the recombined chromosome 13/22. Fluorescent probes specific to chromosome 13/21 α-satellite (Cen13/21; red) and 14/22 α-satellite (Cen14/22; green) localize together on the recombined chromosome 13/22 in 100% of cells ($n = 19$). Scale bars, 10 µm (chromosome spread) and 1 µm (individual chromosome).

Finally, we identified variation in the sequence composition of the α-satellite HORs themselves. Comparison of the α-satellite HORs in these 2,110 centromeres to those in the CHM13 and CHM1 genomes revealed 1,870 new α-satellite HOR variants (Supplementary Table 4). We confirmed that these new α-satellite HOR variants are present at least twice in the genome they were identified in, and 97% of them exist in at least one other genome in our dataset or in the broader dataset of 5,747 complete centromeres assembled by the HPRC (Supplementary Table 5), underscoring their biological relevance. Many of these new α-satellite HOR variants result in completely new α-satellite HOR array organizations relative to CHM13 and CHM1. For 12 chromosomes (chromosomes 4, 6–8, 10, 13, 14, 16–18, 20 and 21), we identified 26 distinct haplotypes that had markedly different α-satellite HOR compositions compared with CHM13 and CHM1 (Supplementary Figs. 11–22). The most pronounced differences occurred on chromosomes 7, 16 and 21, where we found that a subset of individuals harbour a centromere with α-satellite HOR variants that are up to 1,378-fold more abundant than in the CHM13 or CHM1 centromeres, altering the composition and organization of the arrays (Fig. 2e,f, Supplementary Figs. 13, 18 and 22 and Supplementary Tables 7 and 8).

Although rare, we note the discovery of a recombination event between two centromeres from acrocentric chromosomes (chromosomes 13 and 22), creating a hybrid centromere in an individual (NA20355). Previous studies have shown that acrocentric chromosomes 13, 14, 15, 21 and 22 are prone to homologous recombination, resulting in an exchange of p arms or a fusion of the q arms, as with Robertsonian translocations[23,24]. Although these recombination events have been reported to occur at the pseudohomologous regions on the p arms[23–26], they have never, to our knowledge, been shown to occur within the centromeric α-satellite itself. Here we identified an individual who carries a recombined chromosome 13/22 centromere, with the breakpoint within the α-satellite HOR array, the p arm from chromosome 13 and the q arm from chromosome 22 (Fig. 2g,h). We validated the breakpoint using both PacBio HiFi and ONT long-read sequencing data (Extended Data Fig. 1a–d) and by fluorescence in situ hybridization (FISH), which showed colocalization of chromosomes 13 and 22 alphoid DNA on the same chromosome in NA20355 (Fig. 2i and Extended Data Fig. 1e). CENP-C immunolabelling further confirmed the presence of a functional kinetochore on the chromosome 13 alphoid DNA (Extended Data Fig. 1f). Close examination of the breakpoint reveals 6 bp of homology (AAAAGA) between the α-satellite monomers on chromosome 13 and 22 (Extended Data Fig. 1d), indicating that the recombination event may have been mediated by microhomology, as previously suggested[27]. To estimate the frequency of this event in the greater population, we assessed 633 additional chromosome 13 and 22 centromeres assembled from 351 individuals (702 haplotypes) by the HPRC and found two such examples of recombined chromosome 13/22 centromeres (from individuals HG00235 and HG00639; Supplementary Fig. 23), bringing the estimated rate of occurrence to 1 in 283 in our dataset (n = 415). The two other individuals (HG00235 and HG00639) have a recombination breakpoint at the same 6-bp sequence as in NA20355, indicating a conserved mechanism for the formation of a hybrid chromosome 13/22 centromere. Although all three individuals represent different populations, they share the same centromere haplotype, suggesting that this structure evolved relatively recently and may predispose individuals to recombination between chromosomes 13 and 22. Biochemical and cell-based experiments that test the molecular basis for the formation of a hybrid chromosome 13/22 centromere and its functional consequences will be needed to confirm this hypothesis.

Epigenetic differences among centromeres

Functional centromeres are epigenetically defined by two key features: nucleosomes containing the histone H3 variant CENP-A (refs. 28–30) and underlying domains of hypomethylated DNA, known as CDRs[9,22]. Together, these marks establish a unique chromatin foundation for the assembly of the kinetochore, a multiprotein complex that binds to spindle microtubules to ensure faithful chromosome segregation during cell division. While all functional centromeres recruit a kinetochore, we lack a clear understanding of the variation in kinetochore position, distribution and organization among human centromeres, particularly across diverse haplotype structures and genetic ancestries. Using the complete centromere assemblies that we generated, we sought to determine how the underlying sequence and structure within centromeres affects the position and distribution of the kinetochore and if this stratifies by population. To do this, we first determined the CpG methylation profile of each centromere by aligning native Oxford Nanopore Technologies (ONT) long-read sequencing data from the same source genome to each assembly and identifying the location of the CDR, which spans from the first hypomethylated base to the last hypomethylated base and can contain one or multiple methylation dips[31]. We next performed CENP-A DiMeLo-seq followed by ONT long-read sequencing on 10% of the cell lines, which revealed that every CDR was enriched with CENP-A (Supplementary Fig. 24). Finally, we adapted and optimized CENP-A CUT&RUN for ONT long-read sequencing, confirming the location of CENP-A within the CDR (Methods and Supplementary Fig. 24).

Using these validated kinetochore sites, we characterized their distribution, length and position along each centromere (Fig. 3). We found that 94% of centromeres contain one kinetochore site, while 6% (119 out of 2,110) contain two sites and less than 1% (1 out of 2,110) contains three sites, separated by more than 150 kb of sequence (Fig. 3a, Supplementary Fig. 25 and Supplementary Table 9). Centromeres with multiple kinetochore sites tend to have them clustered within 390 kb of each other (Fig. 3b and Supplementary Table 9); however, rarely, they are separated by multiple megabases of sequence, as with the chromosome 11 centromere in HG00513, where the two kinetochores differ in position by 2.9 Mb (Extended Data Fig. 2a,b). Centromeres with multiple kinetochore sites are also meiotically stable, as the chromosome 11 centromere with two distant kinetochores is also observed in the child of HG00513, HG00514 (Extended Data Fig. 2c,d), indicating that this epigenetic architecture is transmissible to the next generation.

Focusing on individual kinetochore sites, we find that they are 240 kb on average, with the smallest single site being 90 kb and the largest 630 kb (Fig. 3c and Supplementary Table 9). Most kinetochore sites contain multiple dips in methylation that alternate with methylated DNA, with an average of 4.3 dips and a maximum of 10 dips (Fig. 3d,e and Supplementary Fig. 26). Although the underlying DNA sequence does not often correlate with DNA methylation patterns (Fig. 3f,g), we find that, in a subset of centromeres, a change in methylation frequency is owing to the insertion of a mobile element within the kinetochore site (Fig. 3h). In total, we identify 636 centromeres (30%) in which a mobile element (L1 or *Alu*) is present in the α-satellite HOR array (Fig. 1, Supplementary Fig. 27a and Supplementary Table 10). In 33 of these centromeres (around 5%), the mobile element is present in the putative kinetochore site, altering the DNA methylation profile of the CDR and breaking it into multiple dips in methylation (Fig. 3h and Supplementary Fig. 27b–d). While mobile element insertions (MEIs) within kinetochore sites are rare, they occur frequently within the chromosome 16 centromere, with 14% (16 out of 112) of centromeres containing an MEI within the putative kinetochore site—17 times more frequent than in the other chromosomes (Supplementary Fig. 27a–d and Supplementary Table 10). This is notable, as chromosome 16 is known to have the highest rate of nondisjunction during meiosis and is the most frequently gained or lost chromosome in aneuploid embryos[32,33]. Previous studies have suggested that errors in kinetochore–microtubule attachments can lead to increased rates of chromosome missegregation[34], and MEIs within the kinetochore site may disrupt the three-dimensional chromatin organization required for proper spindle microtubule attachment. Additional studies that test the functional consequences of MEIs within

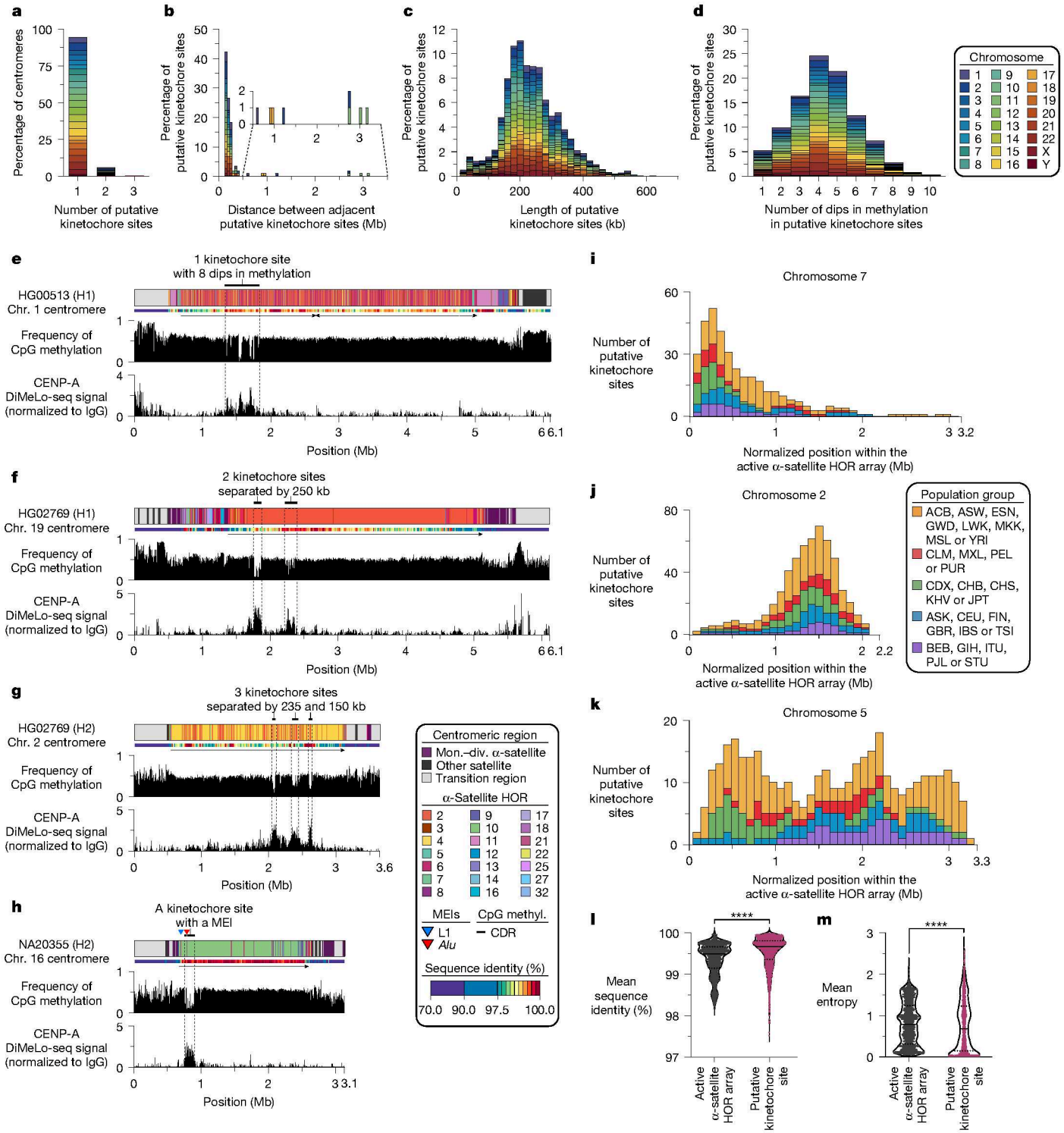


Fig. 3 | Variation in the length, distribution and position of putative kinetochore sites among human centromeres. a, The number of putative kinetochore sites within each chromosome, marked by hypomethylated DNA and separated by >150 kb of sequence. **b**, The distance between putative kinetochore sites within a single centromere. **c**, The length of putative kinetochore sites. **d**, The number of dips in methylation among putative kinetochore sites. **e**–**g**, Examples of centromeres with a single kinetochore site (**e**), two kinetochore sites separated by 250 kb of sequence (**f**) and three kinetochore sites separated by 235 and 150 kb of sequence and residing within different sequence contexts (**g**). **h**, Example of an *Alu* insertion within the kinetochore site of the NA20355 chromosome 16 *D16Z2* α-satellite HOR array. **i**–**k**, The position of the CDR across active α-satellite HOR arrays within the centromeres from chromosomes 7 (**i**), 2 (**j**) and 5 (**k**). On chromosome 7, most CDRs are located on the p arm-proximal side of the array; on chromosome 2, most CDRs are located on the q arm-proximal side of the array; and on chromosome 5, CDRs are located on different regions of the α-satellite HOR array depending on their haplotype structure and population group. The distribution of CDRs for each chromosome is shown in Extended Data Fig. 3. **l**,**m**, Comparison of the mean sequence identity (**l**) and entropy (**m**) between sequences comprising the active α-satellite HOR array versus those within the CDR ($n = 2{,}110$ centromeres; mean sequence identity = 99.36% versus 99.55%; mean entropy = 0.80 versus 0.74, for the active α-satellite HOR array versus CDR, respectively). The mean (solid line) and the first and third quartiles (dotted lines) are shown. Statistical analysis was performed using two-tailed Mann–Whitney *U*-tests; $P = 2.5 \times 10^{-73}$ (sequence identity) and $P = 2.8 \times 10^{-8}$ (entropy).

kinetochore sites, such as those found on chromosome 16, will need to be performed to confirm their contribution to increased chromosome missegregation rates.

Perhaps the most striking finding, however, is that the kinetochore position is not random and is instead closely associated with the underlying sequence and structure of the centromere. Indeed, we find that the position of the kinetochore tends to fall into one of three categories: p arm-proximal, q arm-proximal or distributed in different locations throughout the α-satellite HOR array depending on the haplotype. For 11 out of 24 chromosomes (chromosomes 1, 3, 7, 9, 13–16, 21, 22 and Y), nearly all CDRs are located on the p arm-proximal side of the array (Fig. 3i and Extended Data Fig. 3). For two chromosomes (chromosomes 2 and 20), almost all CDRs are located on the q arm-proximal side of the array (Fig. 3j and Extended Data Fig. 3). For the remaining chromosomes (chromosomes 4–6, 8, 10–12, 17–19 and X), CDRs are located on different regions of the α-satellite HOR array depending on their haplotype structure (Fig. 3k and Extended Data Fig. 3). Importantly, centromeres from individuals with different genetic ancestries have different CDR distributions for these chromosomes. This is perhaps best demonstrated with chromosome 5 (Fig. 3k), where individuals with African, Admixed American and East Asian ancestry have CDRs distributed along the length of the α-satellite HOR array, but those with European and South Asian ancestry rarely have a CDR located towards the p arm.

This bias in CDR position suggests that there may be particular sequences or structures that are more likely to serve as the site of the kinetochore than others within the centromeric α-satellite HOR array. To test this hypothesis, we calculated the mean sequence identity within the CDR and across the α-satellite HOR array and found that the CDR resides within sequences that are more highly identical to each other than the rest of the array (mean sequence identity = 99.55% versus 99.36%, respectively; $P < 0.0001$, two-tailed Mann–Whitney *U*-test; Fig. 3l, Extended Data Fig. 4a and Supplementary Table 11). Moreover, the CDR resides within regions that are more homogenous in sequence and structure and therefore have lower entropy (mean entropy = 0.74 versus 0.80, respectively; $P < 0.0001$, two-tailed Mann–Whitney *U*-test; Fig. 3m, Extended Data Fig. 4b and Supplementary Table 11). This is the case for all centromeres except for those from chromosomes 8 and 13, which have higher entropy among sequences underlying the kinetochore[7] (Fig. 3m, Extended Data Fig. 4b and Supplementary Table 11). The finding that the kinetochore typically resides on highly identical, homogenous stretches of α-satellite HORs is consistent with a model previously put forth that suggests that centromeric proteins and DNA act in concert, recombining the sequences underneath the kinetochore through a kinetochore-associated recombination machine and ultimately leading to increased homogeneity among these sequences[35,36].

Evolutionary events shape centromeres

Our findings reveal that centromeres vary considerably in size, sequence, structure and epigenetic landscape among diverse humans, with at least 226 major centromere haplotypes, 1,870 α-satellite HOR variants and distinct epigenetic landscapes that correlate with centromere structure (Figs. 1–3). However, when these haplotypes first emerged and how rapidly they are evolving remains unclear. To investigate the nature of evolutionary change among centromeres, we leveraged the suppressed meiotic recombination around these regions and identified sequences in the pericentromeres that are in linkage disequilibrium and associated with particular centromere haplogroups[10] (Supplementary Fig. 28 and Supplementary Table 12). We then built maximum-likelihood phylogenetic trees for each centromere, using a chimpanzee sequence as an outgroup and assuming a chimpanzee–human divergence time of 6.2 million years[37] (Supplementary Fig. 28). The phylogenetic trees that we built for both of the p and the q arms are highly similar, supporting the notion of limited-to-no recombination across centromeres during meiosis (Supplementary Figs. 29–56). Moreover, they reveal that centromeres with similar structures and epigenetic landscapes cluster within monophyletic clades, indicating that they have a shared evolutionary history.

Phylogenetic analysis reveals evidence that both ancient and modern evolutionary events have shaped human centromere architecture. First, we identify three chromosomes that have centromere haplotypes separated by more than 1 million years of evolution (chromosomes 10, 12 and 21; Fig. 4a and Supplementary Figs. 38, 40 and 49), as confirmed with the HPRC centromere assemblies (Supplementary Figs. 54–56). For chromosome 12, the minor haplotype is present only within a subset of individuals with African ancestry (Supplementary Figs. 40 and 55), suggesting that this haplotype arose in Africa and was probably lost during population bottlenecks after the out-of-Africa migration. By contrast, for chromosomes 10 and 21, we find minor haplotypes that are present within individuals of diverse genetic ancestry (Fig. 4a and Supplementary Figs. 38, 48, 54 and 56), suggesting that these haplotypes possibly emerged as a result of introgression with archaic hominins. To test this hypothesis, we applied an alignment-free *k*-mer-based approach[38] that leverages high-coverage Neanderthal and Denisovan whole-genome sequencing datasets[39–42] to identify archaic hominin-specific *k*-mers that are not present in modern-day African individuals (Methods). We found that the minor haplotypes from chromosomes 10 and 21 have a significant enrichment of archaic hominin-specific *k*-mers, suggesting the presence of introgressed DNA from Neanderthals and Denisovans in these sets of centromeres (Fig. 4b, Extended Data Fig. 5 and Supplementary Fig. 57). By contrast, more evolutionarily young centromeres from chromosome 21 (Fig. 4c) and the minor haplotypes from chromosome 12 (Supplementary Fig. 58) do not show this enrichment, indicating that introgression is specific to only a subset of centromere haplotypes. However, we note that alternative explanations, such as incomplete lineage sorting, ancestral polymorphism or independent convergent mutation, cannot be excluded.

In addition to ancient evolutionary events, we find that modern evolutionary changes have shaped the structure of human centromeres. In fact, we find that nearly all chromosomes have centromere haplotypes that have emerged in the past 10,000 years, such as on chromosome 1, where a 2–4 Mb duplication of the *D1Z7* active α-satellite HOR array is estimated to have occurred in a subset of African individuals over the past approximately 500 generations (Fig. 4d and Supplementary Figs. 10a, 28 and 53). Even more recently, an approximately 1 Mb duplication of the *D4Z1* active α-satellite HOR array on chromosome 4 is estimated to have occurred within the past 1,000 years (or 50 generations) among a subset of African individuals (Fig. 2c and Supplementary Fig. 32), indicating that centromeres are continually mutating and evolving.

Elevated mutation rates in centromeres

The variable evolutionary history among centromeres suggests that centromeres from different chromosomes may have different mutation rates that alter their sequence and structure. To test this hypothesis, we refined our previous mutation rate estimation method[10] to use a pan-centromere alignment framework with clade-specific reference sequences to account for lineage-level differences (Methods and Supplementary Fig. 28). Using this framework, we estimate the mean mutation rate among centromeric α-satellite HOR arrays to be 4.07×10^{-7} mutations per base pair per generation (m.b.p.g.) (95% confidence interval = $4.05–4.08 \times 10^{-7}$ m.b.p.g.), which is 16.4-fold higher than the unique portions flanking the centromeres[10,16] (Supplementary Table 13 and Supplementary Figs. 59–69). The lowest mutation rate occurs in the chromosome Y centromere (4.7×10^{-8} m.b.p.g. on average), while the highest occurs in the chromosome 1 centromere (9.5×10^{-7} m.b.p.g. on average)—a 20.1-fold difference in mutation

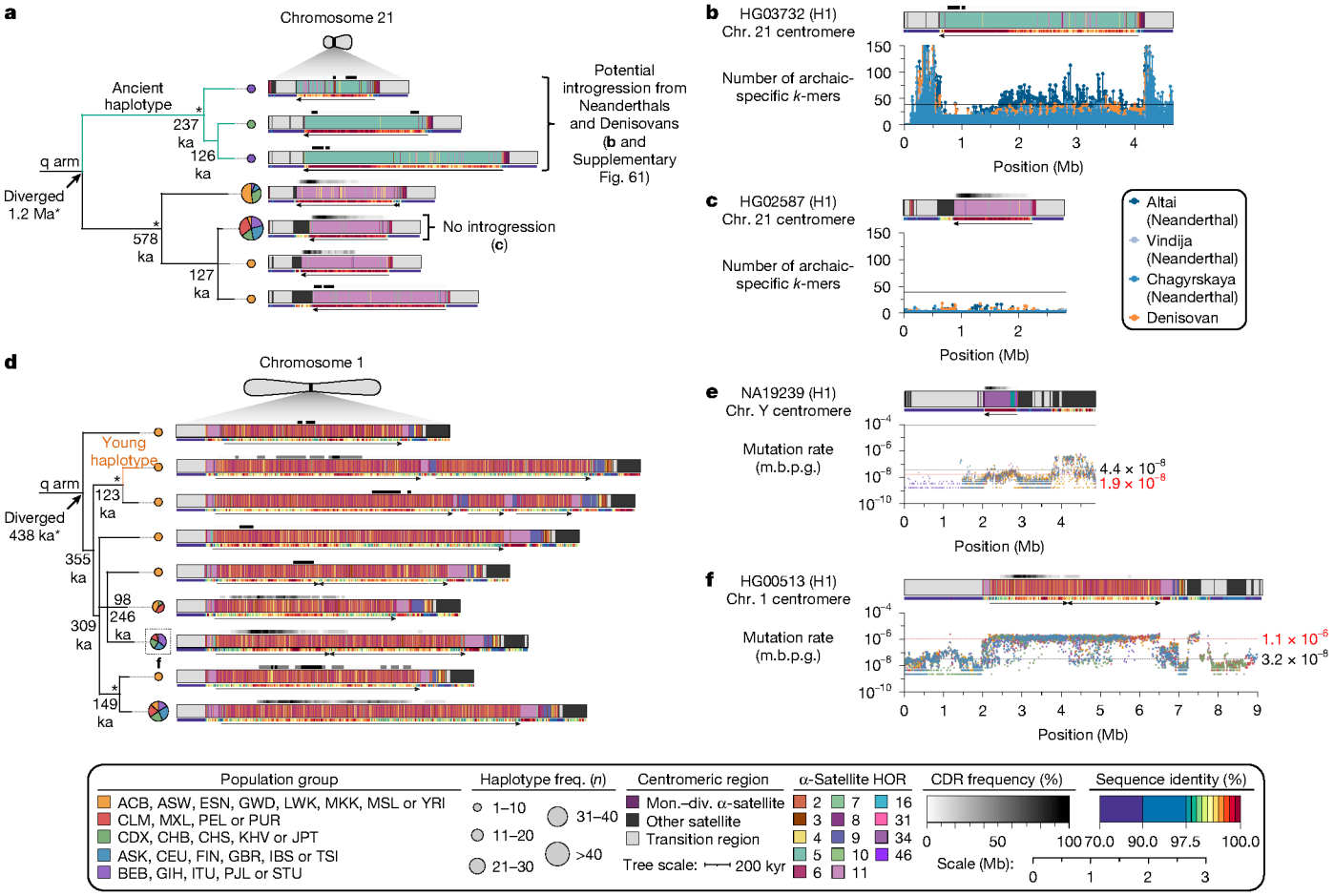


Fig. 4 | Evidence of potential introgression from archaic hominins within centromeres. a, Maximum-likelihood phylogenetic tree depicting the q arm topology for chromosome 21 centromere haplogroups along with the estimated divergence times and haplotype frequencies. The asterisks indicate nodes with 100% bootstrap support, and nodes with 90–99% bootstrap support are indicated numerically. An ancient haplotype with a different *D21Z1* α-satellite HOR array organization (composed mainly of 5-mer α-satellite HORs) is observed, with an estimated divergence time of 1.2 million years ago (Ma). This haplotype is found only within non-African individuals in our dataset, suggesting that it may have arisen owing to introgression with archaic hominins. **b**,**c**, The distribution of archaic-specific DNA *k*-mers across 20 kb windows of the HG03732 (H1) (**b**) or HG02587 (H1) (**c**) chromosome 21 centromeric regions using high-quality archaic short-read data from four genomes: Altai (Neanderthal), Vindija (Neanderthal), Chagryskaya (Neanderthal) and Denisovan. The black dashed lines indicate the genome-wide *k*-mer count cut-off, 42, which corresponds to the genome-wide archaic introgression fraction[35]. The ancient haplotype enriched with 5-mer α-satellite HORs has an abundance of archaic DNA *k*-mers, while the most frequent haplotype enriched with 11-mer α-satellite HORs does not. **d**, Maximum-likelihood phylogenetic tree depicting the q arm topology for chromosome 1 centromere haplogroups along with estimated divergence times and haplotype frequencies. A young haplotype with a large, multimegabase duplication of the α-satellite HOR array is observed and is estimated to have emerged in the past 10,000 years (or 500 generations). **e**,**f**, The estimated mutation rate across the chromosome Y (**e**) and chromosome 1 (**f**) centromeric regions. Individual datapoints from 10 kb pairwise sequence alignments are shown. The mean mutation rates across the active α-satellite HOR array and flanking non-satellite sequences are shown as dotted red and black lines, respectively. There is a 20-fold difference in mutation rate of the α-satellite HOR arrays between the chromosome Y and 1 centromeres. ka, thousand years ago; kyr, thousand years.

rate (Fig. 4e,f and Supplementary Figs. 59c, 69c and 70). Even within the same chromosome, we find that centromere haplogroups exhibit diverse mutation rates, ranging from 1.03-fold to 5.02-fold on chromosomes 18 and 12, respectively (Supplementary Figs. 64e–j, 66i,j and 71 and Supplementary Table 13). Despite these differences, we find that the α-satellite HOR mutation rates are relatively consistent across populations, ranging from an average of 4.0×10^{-7} to 4.1×10^{-7} m.b.p.g. (Supplementary Fig. 70b).

To validate the centromere mutation rates across multiple generations, we leveraged 28 near-complete genome assemblies from a four-generation pedigree (CEPH 1463) that was recently generated[16] (Supplementary Fig. 72a). These genome assemblies already had 288 centromeres completely resolved[16]; however, we reasoned that we could repair many of the remaining centromeres, taking advantage of the complementary error profiles produced by Verkko[17] and hifiasm[18]. Applying AssemblyRepairer to the Verkko genome assemblies for generations 1–4 (G1–4) enabled us to resolve an additional 195 centromeres, which we subsequently validated (Supplementary Note 1 and Supplementary Tables 14 and 15), bringing the total number of centromeres to 483—a 67.7% increase (G1, 134; G2, 70; G3, 215; G4, 64; Supplementary Fig. 72b). Among these 483 centromeres, 244 of them were transmitted from one generation to the next (Supplementary Fig. 72c and Supplementary Table 16), enabling us to identify de novo single-nucleotide variants (SNVs) and structural variants (SVs) that occur within the centromeric regions across generations.

Using these assemblies, we identified 206 de novo SNVs within centromeric regions across four generations (Supplementary Table 17), which we validated with raw PacBio HiFi and ONT reads as well as with transmission to the next generation when possible. We found that the mutation rate varies depending on the underlying sequence (Fig. 5a and Supplementary Table 18). The lowest mutation rate occurs in the unique sequences in the pericentromere (7.1×10^{-8} m.b.p.g.), while the

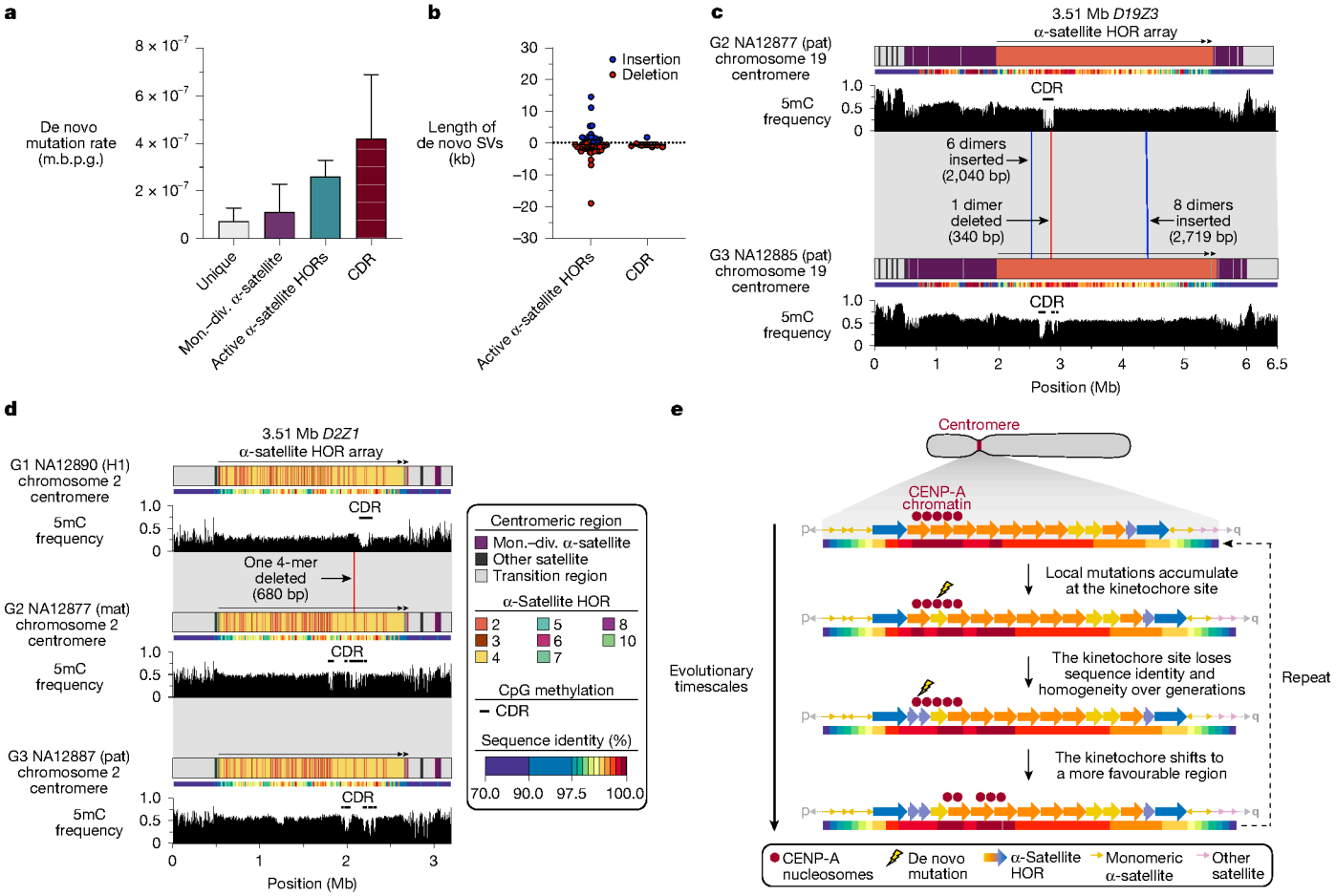


Fig. 5 | Accelerated mutation rates within the putative kinetochore site results in changes in its position, distribution and sequence composition across generations. a, Estimated de novo mutation rates for different sequences within centromeric regions across four generations of a pedigree (CEPH1463). The active α-satellite HORs and CDR are mutating 3.4- and 5.9-fold faster than the unique sequences in the pericentromere, respectively. Mutation rates were calculated as the cumulative number of de novo SNVs observed across 243 centromeres divided by the total number of analysed base pairs in each sequence class. The error bars represent the 95% confidence intervals around the rate estimate. **b**, The length of de novo SVs within the active α-satellite HORs and CDR. $n = 43$ and $n = 8$ de novo SVs for the α-satellite HOR array and CDR, respectively. The error bars represent the 95% confidence intervals around the mean. 5mC, 5-methylcytosine. **c**,**d**, Examples of SVs occurring within the active α-satellite HOR array and CDR in the chromosome 19 (**c**) and chromosome 2 (**d**) centromeres across generations. Mutations occurring within or near the CDR often result in a change in the position, distribution and/or length of the CDR from one generation to the next. **e**, Model of centromere mutation and evolution, showing that the CDR is often located within the most identical and homogenous sequences of the centromere and is prone to both SNVs and SVs that alter the genetic and epigenetic landscape of the centromere over time. mat, maternal; pat, paternal.

monomeric–divergent α-satellite HORs have a slightly higher mutation rate (1.1×10^{-7} m.b.p.g.), and the active α-satellite HORs have a mutation rate of more than twice that of the monomeric–divergent α-satellite HORs (2.6×10^{-7} m.b.p.g.). Within the active α-satellite HOR array itself, we find that the CDR has the highest mutation rate (4.2×10^{-7} m.b.p.g.), which is approximately 1.8-fold greater than the surrounding α-satellite HORs in the rest of the array (2.4×10^{-7} m.b.p.g.; Fig. 5a and Supplementary Table 18), as reported in cancer cells[43]. Importantly, these mutation rates are similar to our earlier pan-centromere-based estimates (Fig. 4c–f), which averaged 4.07×10^{-7} m.b.p.g.

In addition to de novo SNVs, we identified 43 de novo SVs (defined as insertions or deletions more than 50 bp in length) within the α-satellite HOR arrays that range in length from 340 bp to 19.4 kb (Fig. 5b and Supplementary Table 19). Among these 43 SVs, 8 of them occur within the CDR. Given that the CDR comprises only 11% of the active α-satellite HOR array on average (Supplementary Table 18), this is a nearly twofold enrichment of SVs in the CDR compared to what would be expected by chance. This elevated SV mutation rate in the CDR (Supplementary Table 20) may reflect the higher sequence identity of these regions, which are more prone to homologous recombination, leading to expansions and deletions, relative to less identical sequences. Among the eight SVs that we observed in the CDR across generations, all of them reside within highly identical sequences, such as the deletion of a dimer α-satellite HOR in the chromosome 19 centromere from G2 to G3 (Fig. 5c) as well as a deletion of a 4-mer α-satellite HOR in the chromosome 2 centromere from G1 to G2, which is maintained in G3 (Fig. 5d). When this occurs, this often alters the epigenetic landscape of the centromere, splitting the CDR into two or more sites and/or shifting its position in the next generation (Supplementary Table 21 and Supplementary Fig. 73).

Together, our findings suggest a model for centromere mutation and evolution that integrates the genetic and epigenetic variation that we observe across global populations as well as the mutational landscape among recent generations (Fig. 5e). In this model, we find that the putative kinetochore site, marked by the CDR and an enrichment of CENP-A chromatin, typically resides within the most highly identical and homogenous sequences of the centromeric α-satellite HOR array. Owing to its high sequence identity and homogeneity, this region is prone to mutations at both the single-nucleotide and structural level that alter its position, distribution and length and lead to changes in

both genetic and epigenetic landscapes over time. Cell-based assays that identify the mutational processes that shape centromere diversity across generations will need to be performed in the future to confirm this hypothesis.

Discussion

Here we present a comprehensive, population-scale analysis of human centromere sequence, structure and epigenetic variation by assembling and characterizing 2,110 complete centromeres from 65 diverse individuals and complementing them with 5,747 centromeres from 227 individuals and 483 centromeres from a four-generation pedigree (8,340 centromeres in total; Supplementary Tables 6, 22 and 23). Our work expands the number of complete human centromere sequences by more than fivefold compared to previous studies[10,13], revealing 226 major centromere haplotypes and 1,870 α-satellite HOR variants. Additionally, in contrast to previous work that either compared limited numbers of complete centromeres or analysed larger cohorts with incomplete assemblies, we combined deep sequence-level resolution across thousands of centromeres with comprehensive epigenetic profiling and multigenerational mutation tracking. Moreover, our integrated centromere maps make it possible to systematically link sequence variation, structural rearrangements, MEIs and epigenetic organization to kinetochore positioning and recent evolutionary dynamics across human populations, establishing the foundation for understanding both the mechanistic basis of centromere evolution and its implications for chromosome segregation fidelity and human disease.

Our population-scale analysis identifies multiple mechanisms that contribute to centromere diversity. We identify large duplications, deletions and inversions on ten chromosomes (chromosomes 1, 4, 5, 7, 10, 12, 13, 17, 19 and 20), as well as a rare recombination event between acrocentric centromeres (chromosomes 13/22) that alters the composition and organization of the centromeric α-satellite HOR arrays. We find that MEIs are present in around 30% of centromeres and occur within the kinetochore site in 5% of centromeres, altering local methylation patterns. Chromosome 16 shows a marked 17-fold enrichment of MEIs within the kinetochore site. These structural and insertional changes provide mechanistic routes by which centromeric sequence and epigenetic landscapes can be remodelled on timescales ranging from ancient (more than 100,000 years ago) to recent (centuries).

A central conclusion from our work is that the position of the kinetochore, marked by the CDR and an enrichment of CENP-A-containing nucleosomes, is not stochastic but is tightly coupled to local centromere sequence and structure. Across chromosomes, we observe strong positional biases of the kinetochore site and show that the sequence underlying the kinetochore is, on average, more homogeneous and of higher sequence identity than the surrounding α-satellite HORs. This association indicates that certain sequence contexts (for example, highly identical, low-entropy stretches of α-satellite DNA) are preferential substrates for kinetochore establishment and/or maintenance. This finding helps to reconcile models in which centromeric proteins and DNA co-evolve (for example, kinetochore-associated recombination machine)[35,36] and suggests that local sequence homogeneity may facilitate the assembly or maintenance of centromeric chromatin (Supplementary Note 4).

One result of particular biological importance is the elevated mutation dynamics we observe at centromeres and specifically within kinetochore sites. Using both pan-centromere phylogenetic approaches and direct observation of de novo changes in a four-generation pedigree, we show that α-satellite HOR arrays mutate substantially faster than the flanking unique sequence (mean of around 4.07×10^{-7} m.b.p.g.), that centromeres exhibit a 20-fold range of mutation rates (for example, chromosome Y is much slower than chromosome 1), and that the CDR mutates more rapidly than the rest of the active α-satellite HOR array (1.8-fold higher SNV rate and a marked enrichment of structural variants). Regardless, our data support a model in which homogeneous, kinetochore-associated regions are intrinsically prone to both nucleotide and structural changes, probably because sequence homogeneity facilitates unequal exchange and recombination events (Supplementary Note 5). The accelerated turnover at the functional site of the kinetochore offers a potential mechanistic basis for rapid centromere evolution and for a genetic–epigenetic arms race between centromeric DNA and centromere-binding proteins.

Collectively, our findings provide the foundation for a unified model in which (1) kinetochores preferentially form on homogenous stretches of α-satellite HORs; (2) those stretches are prone to elevated rates of homologous recombination and structural change; and (3) such genetic changes feed back on chromatin state and kinetochore localization, producing continual shifts in centromere architecture across generations (an evolutionary 'treadmill' of DNA–protein interactions). This model integrates our observations of haplotype structure, epigenetic patterns, MEIs and mutational spectra and is consistent with longstanding hypotheses about centromere drive[44,45] and centromere-associated recombination machinery[35,36].

There are multiple implications for human biology and disease. First, the finding that MEIs (and other structural changes) can occur within kinetochore sites raises the possibility that sequence perturbations directly influence chromosome segregation fidelity. The elevated incidence of MEIs in the chromosome 16 kinetochore site, a chromosome already known for nondisjunction propensity, is particularly intriguing, and future experiments that test the functional consequence of MEIs within kinetochore sites will need to be performed to confirm their role in increased chromosome missegregation rates. Second, population-stratified haplotypes and ancient introgressed centromeres may influence subtle differences in meiotic behaviour or fertility across lineages and could be relevant for interpreting aneuploidy risk and patterns of chromosomal instability in disease. Finally, the rapid evolution of kinetochore-proximal sequence may have consequences for cross-species compatibility of centromere proteins and could inform models of reproductive isolation.

Looking forward, a number of experiments can build on this resource: cell-based assays that measure segregation fidelity after targeted MEI or SV introduction, population genetic screens for associations between centromere haplotypes and reproductive outcomes, and single-molecule imaging of kinetochores on engineered centromeres. On the computational side, integrating our centromere maps with large-scale clinical and cytogenetic datasets may reveal links between centromere genotype and aneuploidy spectra in embryos or cancer. Finally, extending surveillance to additional pedigrees and to gamete sequencing will clarify the mutation processes that most strongly shape centromeric evolution.

In summary, the complete centromere sequence assemblies and comprehensive analyses presented here reveal centromeres as dynamic genomic regions where sequence homogeneity, structural changes, mobile elements and epigenetic state interact to shape kinetochore position and evolution. Our results provide a foundation for future mechanistic experiments and for integrating centromere biology into models of human genetic disease, fertility and genome evolution.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-026-10841-9.

1. Eichler, E. E., Clark, R. A. & She, X. An assessment of the sequence gaps: unfinished business in a finished human genome. *Nat. Rev. Genet.* **5**, 345–354 (2004).

2. Lee, C., Wevrick, R., Fisher, R. B., Ferguson-Smith, M. A. & Lin, C. C. Human centromeric DNAs. *Hum. Genet.* **100**, 291–304 (1997).
3. Rudd, M. K. & Willard, H. F. Analysis of the centromeric regions of the human genome assembly. *Trends Genet.* **20**, 529–533 (2004).
4. Rudd, M. K., Schueler, M. G. & Willard, H. F. Sequence organization and functional annotation of human centromeres. *Cold Spring Harb. Symp. Quant. Biol.* **68**, 141–149 (2003).
5. McNulty, S. M. & Sullivan, B. A. Alpha satellite DNA biology: finding function in the recesses of the genome. *Chromosome Res.* **26**, 115–138 (2018).
6. Miga, K. H. et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
7. Logsdon, G. A. et al. The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).
8. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
9. Altemose, N. et al. Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).
10. Logsdon, G. A. et al. The variation and evolution of complete human centromeres. *Nature* **629**, 136–145 (2024).
11. Volpe, E. et al. The reference genome of the human diploid cell line RPE-1. *Nat. Commun.* **16**, 7751 (2025).
12. Hansen, N. F. et al. A complete diploid human genome benchmark for personalized genomics. Preprint at *bioRxiv* https://doi.org/10.1101/2025.09.21.677443 (2025).
13. Logsdon, G. A. et al. Complex genetic variation in nearly complete human genomes. *Nature* **644**, 430–441 (2025).
14. Gao, Y. et al. A pangenome reference of 36 Chinese populations. *Nature* **619**, 112–121 (2023).
15. Lucas, J. K. et al. HPRC2: a human pangenome reference with near-complete coverage of common genetic variation. Preprint at *bioRxiv* https://doi.org/10.64898/2026.07.21.739710 (2026).
16. Porubsky, D. et al. Human de novo mutation rates from a four-generation pedigree reference. *Nature* **643**, 427–436 (2025).
17. Rautiainen, M. et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* **41**, 1474–1482 (2023).
18. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
19. Asri, M. et al. Evaluating genome assemblies with HMM-Flagger. Preprint at *bioRxiv* https://doi.org/10.64898/2026.02.27.708355 (2026).
20. Dishuck, P. C., Rozanski, A. N., Logsdon, G. A., Porubsky, D. & Eichler, E. E. GAVISUNK: genome assembly validation via inter-SUNK distances in Oxford Nanopore reads. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btac714 (2022).
21. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
22. Gershman, A. et al. Epigenetic patterns in a complete human genome. *Science* **376**, eabj5089 (2022).
23. Guarracino, A. et al. Recombination between heterologous human acrocentric chromosomes. *Nature* **617**, 335–343 (2023).
24. de Lima, L. G. et al. The formation and propagation of human Robertsonian chromosomes. *Nature* **647**, 952–961 (2025).
25. Page, S. L., Shin, J.-C., Han, J.-Y., Choo, K. H. A. & Shaffer, L. G. Breakpoint diversity illustrates distinct mechanisms for Robertsonian translocation formation. *Hum. Mol. Genet.* **5**, 1279–1288 (1996).
26. Page, S. L. & Shatter, L. G. Nonhomologous Robertsonian translocations form predominantly during female meiosis. *Nat. Genet.* **15**, 231–232 (1997).
27. Gerton, J. L. A working model for the formation of Robertsonian chromosomes. *J. Cell Sci.* **137**, jcs261912 (2024).
28. Palmer, D. K., O'Day, K., Wener, M. H., Andrews, B. S. & Margolis, R. L. A 17-kD centromere protein (CENP-A) copurifies with nucleosome core particles and with histones. *J. Cell Biol.* **104**, 805–815 (1987).
29. Sullivan, K. F., Hechenberger, M. & Masri, K. Human CENP-A contains a histone H3 related histone fold domain that is required for targeting to the centromere. *J. Cell Biol.* **127**, 581–592 (1994).
30. Warburton, P. E. et al. Immunolocalization of CENP-A suggests a distinct nucleosome structure at the inner kinetochore plate of active centromeres. *Curr. Biol.* **7**, 901–904 (1997).
31. Mastrorosa, F. K. et al. Identification and annotation of centromeric hypomethylated regions with CDR-Finder. *Bioinformatics* **40**, btae733 (2024).
32. McCoy, R. C. et al. Meiotic and mitotic aneuploidies drive arrest of in vitro fertilized human preimplantation embryos. *Genome Med.* **15**, 77 (2023).
33. Carioscia, S. A. et al. Common variation in meiosis genes shapes human recombination and aneuploidy. *Nature* **651**, 146–153 (2026).
34. Thompson, S. L. & Compton, D. A. Chromosome missegregation in human cells arises through specific types of kinetochore–microtubule attachment errors. *Proc. Natl Acad. Sci. USA* **108**, 17974–17978 (2011).
35. Alexandrov, I. A. et al. Definition of a new alpha satellite suprachromosomal family characterized by monomeric organization. *Nucleic Acids Res.* **21**, 2209–2215 (1993).
36. Shepelev, V. A., Alexandrov, A. A., Yurov, Y. B. & Alexandrov, I. A. The evolutionary origin of man can be traced in the layers of defunct ancestral alpha satellites flanking the active centromeres of human chromosomes. *PLoS Genet.* **5**, e1000641 (2009).
37. Yoo, D. et al. Complete sequencing of ape genomes. *Nature* **641**, 401–418 (2025).
38. Hsieh, P. et al. A global map for introgressed structural variation and selection in humans. *Science* **392**, eadz7518 (2026).
39. Meyer, M. et al. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).
40. Prüfer, K. et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
41. Prüfer, K. et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655–658 (2017).
42. Mafessoni, F. et al. A high-coverage Neandertal genome from Chagyrskaya Cave. *Proc. Natl Acad. Sci. USA* **117**, 15132–15136 (2020).
43. Sohn, M.-H. et al. A telomere-to-telomere map of somatic mutation burden and functional impact in cancer. Preprint at *bioRxiv* https://doi.org/10.1101/2025.10.10.681725 (2025).
44. Henikoff, S., Ahmad, K. & Malik, H. S. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**, 1098–1102 (2001).
45. Henikoff, S. & Malik, H. S. Centromeres: selfish drivers. *Nature* **417**, 227–227 (2002).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026

Human Genome Structural Variation Consortium

Hufsah Ashraf[14,15], Peter A. Audano[4], Marcelo Ayllon[16], Andrey Azov[17], Parithi Balachandran[4], Anna O. Basile[18], Christine R. Beck[4,19], Marc Jan Bonder[20,21,22], Lucy Brooks[17], Marta Byrska-Bishop[18], Mark J. P. Chaisson[23], Zechen Chong[24], André Corvelo[18], Jonathan Crabtree[25], Scott E. Devine[25], Peter Ebert[15,26], Jana Ebler[14,15], Evan E. Eichler[16,27], Aine Fairbrother-Browne[17], Chia-Hsuan Fan[24], Mallory Freeberg[17], Shenghan Gao[1,114], Mark B. Gerstein[28,29], Bida Gu[23], Pille Hallast[4], Patrick Hasenfeld[30], Mir Henglin[14,15], Kendra Hoekzema[16], Kaili Hu[24], Sarah Hunt[17], Matthew Jensen[28,29], Yunzhe Jiang[28,29], Kwondo Kim[4], Miriam K. Konkel[2,3], Jan O. Korbel[30], Youngjun Kwon[16], Peter M. Lansdorp[31], Charles Lee[4], Tiffany Leung[31], Jiaqi Li[28,29], Chong Li[32,33], Jiadong Lin[16], Mark Loftus[4], Glennis A. Logsdon[1], Tobias Marschall[14,15], Gianni V. Martino[2], Ryan E. Mills[34], Yulia Mostovoy[35], Katherine M. Munson[16], Giuseppe Narzisi[18], Lingbin Ni[16], Keisuke K. Oshima[1,114], Carolyn Paisie[4], Samarendra Pani[14,15], Zishan Peng[24], David Porubsky[16], Timofey Prodanov[14,15], Keon Rabbani[23], Tobias Rausch[30], Xinghua Shi[32,33], Yuwei Song[24], Kaitlyn Sun[16], Likhitha Surapaneni[17], Michael E. Talkowski[36,37,38], Vasiliki Tsapalou[30], Andres Veidenberg[17], Feyza Yilmaz[4], DongAhn Yoo[16], Xuefang Zhao[36,37,38], Weichen Zhou[34], Qihui Zhu[4] & Michael C. Zody[18]

[14]Institute for Medical Biometry and Bioinformatics, Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University, Düsseldorf, Germany. [15]Center for Digital Medicine, Heinrich Heine University, Düsseldorf, Germany. [16]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [17]European Molecular Biology Laboratory, Wellcome Genome Campus, European Bioinformatics Institute, Cambridge, UK. [18]New York Genome Center, New York, NY, USA. [19]The University of Connecticut Health Center, Farmington, CT, USA. [20]Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands. [21]Oncode Institute, Utrecht, The Netherlands. [22]Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [23]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA. [24]Department of Biomedical Informatics and Data Science, Heersink School of Medicine, University of Alabama, Birmingham, AL, USA. [25]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA. [26]Core Unit Bioinformatics, Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University, Düsseldorf, Germany. [27]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [28]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. [29]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA. [30]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [31]Terry Fox Laboratory, BC Cancer Agency, Vancouver, British Columbia, Canada. [32]Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA. [33]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA. [34]Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [35]Cardiovascular Research Institute and Institute for Human Genetics, UCSF School of Medicine, San Francisco, CA, USA. [36]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [37]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [38]Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA.

Human Pangenome Reference Consortium

Derek Albracht[39], Ivan A. Alexandrov[40], Jamie Allen[41], Alawi A. Alsheikh-Ali[42], Nicolas Altemose[43], Casey Andrews[44], Dmitry Antipov[45], Lucinda Antonacci-Fulton[39], Mobin Asri[46], Marcelo Ayllon[16], Jennifer R. Balacco[47], Floris P. Barthel[48], Edward A. Belter Jr[39], Halle D. Bender[46], Andrew P. Blair[46], Davide Bolognini[49], Katherine E. Bonini[50], Christina Boucher[51], Guillaume Bourque[52,53,54], Silvia Buonaiuto[55], Shuo Cao[55], Andrew Carroll[56], Ann M. Mc Cartney[46], Monika Cechova[46], Mark J. P. Chaisson[23], Pi-Chuan Chang[56], Xian Chang[46], Jitender Cheema[41], Haoyu Cheng[57], Claudio Ciofi[58], Hiram Clawson[46], Sarah Cody[39], Vincenza Colonna[55], Holland C. Conwell[59],

Robert Cook-Deegan[60], Mark Diekhans[46], Maria Angela Diroma[58], Daniel Doerr[15,61,62], Zheng Dong[44], Danilo Dubocanin[43], Richard Durbin[63,64], Jana Ebler[14,15], Evan E. Eichler[16,27], Jordan M. Eizenga[46], Parsa Eskandar[46], Eddie Ferro[51], Anna-Sophie Fiston-Lavier[65,66], Sarah M. Ford[59], Willard W. Ford[67], Giulio Formenti[47], Adam Frankish[41], Mallory A. Freeberg[41], Qichen Fu[44], Stephanie M. Fullerton[68], Robert S. Fulton[39], Shenghan Gao[1,114], Yan Gao[69], Gage H. Garcia[16], Obed A. Garcia[70], Joshua M. V. Gardner[46], Shilpa Garg[71], Erik Garrison[55], Nanibaa' A. Garrison[72,73,74], John E. Garza[39], Margarita Geleta[75], Mohammadmersad Ghorbani[76], Tina A. Graves-Lindsay[39], Richard E. Green[59], Cristian Groza[77], Bida Gu[23], Andrea Guarracino[48,55], Melissa Gymrek[78], Maximilian Haeussler[46], Leanne Haggerty[41], Ira M. Hall[79,80], Nancy F. Hansen[45], Yue Hao[48], Mohammad Amiruddin Hashmi[42], David Haussler[46], Prajna Hebbar[46], Peter Heringer[15,61,62], Glenn Hickey[46], Todd L. Hillaker[46], S. Nakib Hossain[41], Neng Huang[69,81], Sarah E. Hunt[41], Toby Hunt[41], Alexander G. Ioannidis[43,46], Nafiseh Jafarzadeh[46], Nivesh Jain[47], Erich D. Jarvis[27,47], Maryam Jehangir[48], Juan Jiang[44], Eimear E. Kenny[50], Juhyun Kim[45], Bonhwang Koo[47], Sergey Koren[45], Milinn Kremitzki[39,44], Charles H. Langley[82], Ben Langmead[83], Heather A. Lawson[44], Daofeng Li[44], Heng Li[69,81], Wen-Wei Liao[79,80], Jiadong Lin[16], Tianjie Liu[44], Glennis A. Logsdon[1], Ryan Lorig-Roach[46], Jonathan LoTempio Jr[84], Hailey Loucks[46], Jane E. Loveland[41], Jianguo Lu[85], Shuangjia Lu[79,80], Julian K. Lucas[46], Walfred Ma[23], Juan F. Macias-Velasco[39,44,86], Kateryna D. Makova[87], Maximillian G. Marin[69,81], Christopher Markovic[39], Tobias Marschall[14,15], Franco L. Marsico[55], Fergal J. Martin[41], Mira Mastoras[46], Capucine Mayoud[65], Brandy McNulty[46], Jack A. Medico[47], Julian M. Menendez[46], Karen H. Miga[46], Anna Minkina[88], Matthew W. Mitchell[4], Saswat K. Mohanty[89], Younes Mokrab[76,90,91], Jean Monlong[92], Shabir Moosa[76], Avelina Moreno-Ochando[93,94], Shinichi Morishita[95], Jonathan M. Mudge[41], Katherine M. Munson[16], Njagi Mwaniki[96], Nasna Nassir[42], Chiara Natali[58], Shloka Negi[46], Lingbin Ni[16], Adam M. Novak[46], Faith Okamoto[46], Keisuke K. Oshima[1,114], Pilar N. Ossorio[97], Chie Owa[95], Sadye Paez[47], Benedict Paten[46], Clelia Peano[49,98], Adam M. Phillippy[45], Brandon D. Pickett[45], Laura Pignata[55], Nadia Pisanti[96], David Porubsky[16,99], Pjotr Prins[55], Anandi Radhakrishnan[46], T. Rhyker Ranallo-Benavidez[48], Brian J. Raney[46], Mikko Rautiainen[100], Alessandro Raveane[49], Luyao Ren[16,27], Arang Rhie[45], Fedor Ryabov[101,102], Samuel Sacco[59], Farnaz Salehi[55], Michael C. Schatz[83,103], Laura B. Scheinfeldt[104], Aarushi Sehgal[67], William E. Seligmann[59], Mahsa Shabani[105], Kishwar Shafin[56], Shadi Shahatit[65], Ruhollah Shemirani[50], Vikram S. Shivakumar[83], Swati Sinha[41], Jouni Sirén[46], Linnéa Smeds[89], Steven J. Solar[45], Marco Sollitto[47,58], Nicole Soranzo[49,63,106], Andrew B. Stergachis[16,88], Marie-Marthe Suner[41], Yoshihiko Suzuki[95], Arda Söylev[14,15], Ahmad Abou Tayoun[107,108], Jack A. S. Tierney[41], Chad Tomlinson[39], Francesca Floriana Tricomi[41], Mohammed Uddin[42,109], Matteo Tommaso Ungaro[59,110], Rahul Varki[51], Flavia Villani[55], Ivo Violich[46], Mitchell R. Vollger[88], Brian P. Walenz[45], Charles Wang[111], Lisa E. Wang[50], Ting Wang[39,44,86], Aaron M. Wenger[112], Conor V. Whelan[47], Zilan Xin[44], Zheng Xu[44], Kai Ye[113], DongAhn Yoo[16], Wenjin Zhang[44], Ying Zhou[69], Xiaoyu Zhuo[44] & Giulia Zunino[49]

[39]McDonnell Genome Institute, Washington University School of Medicine, St Louis, MO, USA. [40]Department of Human Molecular Genetics and Biochemistry, Faculty of Medical and Health Sciences, Tel Aviv University, Tel Aviv, Israel. [41]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK. [42]Center for Applied and Translational Genomics (CATG), Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai Health, Dubai, UAE. [43]Department of Genetics, Stanford University, Palo Alto, CA, USA. [44]Department of Genetics, Washington University School of Medicine, St Louis, MO, USA. [45]Genome Informatics Section, Center for Genomics and Data Science Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [46]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, CA, USA. [47]The Vertebrate Genome Laboratory, The Rockefeller University, New York, NY, USA. [48]Bioinnovation and Genome Sciences, The Translational Genomics Research Institute (TGen), Phoenix, AZ, USA. [49]Human Technopole, Milan, Italy. [50]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [51]Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL, USA. [52]Canadian Center for Computational Genomics, McGill University, Montréal, Quebec, Canada. [53]Department of Human Genetics, McGill University, Montréal, Quebec, Canada. [54]Victor Phillip Dahdaleh Institute of Genomic Medicine, Montréal, Quebec, Canada. [55]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN, USA. [56]Google, Mountain View, CA, USA. [57]Department of Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT, USA. [58]Department of Biology, University of Florence, Florence, Italy. [59]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, CA, USA. [60]Consortium for Science, Policy & Outcomes, Arizona State University, Washington, DC, USA. [61]Department for Endocrinology and Diabetology at the Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. [62]Paul-Langerhans-Group Computational Diabetology, German Diabetes Center (DDZ) and Leibniz Institute for Diabetes Research, Düsseldorf, Germany. [63]Wellcome Sanger Institute, Hinxton, UK. [64]Department of Genetics, University of Cambridge, Cambridge, UK. [65]ISEM, Univ Montpellier, CNRS, IRD, Montpellier, France. [66]Institut Universitaire de France, Paris, France. [67]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA. [68]Department of Bioethics & Humanities, University of Washington School of Medicine, Seattle, WA, USA. [69]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA. [70]Department of Anthropology, University of Kansas, Lawrence, KS, USA. [71]School of Health Sciences, University of Manchester, Manchester, UK. [72]Traditional, Ancestral and Unceded Territory of the Gabrielino/Tongva Peoples, Institute for Society & Genetics, University of California, Los Angeles, Los Angeles, CA, USA. [73]Traditional, Ancestral and Unceded Territory of the Gabrielino/Tongva Peoples, Institute for Precision Health, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [74]Traditional, Ancestral and Unceded Territory of the Gabrielino/Tongva Peoples, Division of General Internal Medicine & Health Services Research, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [75]Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA, USA. [76]Medical and Population Genomics Lab, Sidra Medicine, Doha, Qatar. [77]Montreal Heart Institute, Montréal, Quebec, Canada. [78]Department of Pediatrics, University of California San Diego, La Jolla, CA, USA. [79]Center for Genomic Health, Yale University School of Medicine, New Haven, CT, USA. [80]Department of Genetics, Yale University School of Medicine, New Haven, CT, USA. [81]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [82]Department of Evolution and Ecology and the Center for Population Biology, University of California, Davis, CA, USA. [83]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. [84]Department of Pediatrics, Division of Genetics, School of Medicine, University of California, Irvine, CA, USA. [85]Sun Yat-sen University, Guangzhou, China. [86]Edison Family Center for Genome Sciences & Systems Biology, Washington University School of Medicine, St Louis, MO, USA. [87]Department of Biology and Center for Medical Genomics, Penn State University, University Park, PA, USA. [88]Division of Medical Genetics, Department of Medicine, University of Washington School of Medicine, Seattle, WA, USA. [89]Department of Biology, Penn State University, University Park, PA, USA. [90]Department of Biomedical Science, College of Health Sciences, Qatar University, Doha, Qatar. [91]Department of Genetic Medicine, Weill Cornell Medicine-Qatar, Doha, Qatar. [92]IRSD—Digestive Health Research Institute, University of Toulouse, INSERM, INRAE, ENVT, UPS, Toulouse, France. [93]MATCH Biosystems, Elche, Spain. [94]Universidad Miguel Hernández de Elche, Elche, Spain. [95]Department of Computational Biology and Medical Sciences, The University of Tokyo, Chiba, Japan. [96]Department of Computer Science, University of Pisa, Pisa, Italy. [97]Law School, University of Wisconsin-Madison, Madison, WI, USA. [98]Institute of Genetics and Biomedical Research, UoS of Milan, National Research Council, Milan, Italy. [99]Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. [100]Institute for Molecular Medicine Finland, Helsinki Institute of Life Science, University of Helsinki, Helsinki, Finland. [101]The Center for Bio- and Medical Technologies, Moscow, Russia. [102]Centre for Biomedical Research and Technology, HSE University, Moscow, Russia. [103]Department of Biology, Johns Hopkins University, Baltimore, MD, USA. [104]Coriell Institute for Medical Research, Camden, NJ, USA. [105]University of Amsterdam, Amsterdam, The Netherlands. [106]School of Clinical Medicine, University of Cambridge, Cambridge, UK. [107]Center for Genomic Discovery, Mohammed Bin Rashid University, Dubai Health, Dubai, UAE. [108]Dubai Health Genomic Medicine Center, Dubai Health, Dubai, UAE. [109]GenomeArc, Mississauga, Ontario, Canada. [110]Department of Biology and Biotechnologies "Charles Darwin", University of Rome "La Sapienza", Rome, Italy. [111]Center for Genomics, Loma Linda University School of Medicine, Loma Linda, CA, USA. [112]PacBio, Menlo Park, CA, USA. [113]The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an Jiaotong University, Xi'an, China.

Methods

Centromere mapping and annotation

To identify completely and accurately assembled centromeres in whole-genome assemblies from diverse humans, we developed a pipeline called CenMAP (v.0.4.3.1; https://github.com/logsdon-lab/CenMAP). In brief, CenMAP aligns each whole-genome assembly to the centromeric regions of the CHM13 reference genome[7] (v.2.0) using minimap2 (ref. 46) (v.2.28) with the following parameters: -x asm20 --secondary=no -s 25000 -K 15G. The resulting BAM files are then converted into BED files using rustybam (v.0.1.34; https://github.com/vollgerlab/rustybam) and the BED files are intersected with a BED file containing coordinates of unique regions within the pericentromeres that correspond to each chromosome using BEDTools[47] (v.2.31.1). Contigs spanning the centromeres are subsequently identified and renamed to indicate the chromosome they most accurately align to based on the greatest number of alignments to the pericentromeric region. To confirm that the centromeric contigs contain α-satellite sequences, CenMAP runs dna-brnn[48] (v.0.1-r65), which annotates α-satellite repeats within each centromeric contig. Neighbouring α-satellite repeat annotations are iteratively merged if they are within 10 kb of each other, yielding the coordinates of the centromeric α-satellite HOR array(s). The coordinates of the α-satellite HOR array(s) are extended on either side by approximately 500 kb, producing the coordinates of the centromeric regions within each contig.

Evaluation of contiguity and accuracy within centromeric regions. CenMAP evaluates the contiguity and accuracy of the centromeric regions by first extracting the regions using BEDTools and annotating them with RepeatMasker[49] (v.4.1.0) using the following parameters: -engine rmblast -species human -dir {output_directory} -pa {threads} {input_contig}. CenMAP uses these annotations along with a custom tool we developed, CenStats (v.0.0.10; https://github.com/logsdon-lab/CenStats), to check the completeness of the contigs and reorient misoriented contigs from p to q arm relative to the CHM13 reference genome (v.2.0). This is achieved with the subcommand status, which compares the repeat sequence edit distance and Jaccard index relative to the reference genome, selecting the best matching orientation. Completely assembled centromeric contigs are subsequently reoriented to match the orientation of centromeres within the CHM13 reference genome (v.2.0) and then evaluated for the presence of assembly errors. To detect assembly errors in the centromeric regions, we developed NucFlag (v.0.3.3), an upgraded version of NucFreq[50] that uses peak detection in the first- and second-highest read pileup nucleotide frequency to detect collapsed and misjoined sequences within the contigs. With NucFlag, PacBio HiFi reads are first aligned to the whole-genome assemblies using pbmm2 (v.1.13.1; https://github.com/PacificBiosciences/pbmm2) with the following parameters: --log-level DEBUG --preset SUBREAD --min-length 5000. Alignments are filtered using SAMtools[51] (v.1.22) and the filter flag 2308, to remove secondary, supplementary and low-quality alignments. Only regions annotated as 'ALR/Alpha' by RepeatMasker are evaluated, and regions with nucleotide frequencies that are <95% of the mean of the highest nucleotide frequency are considered candidate misjoins, while regions with >225% of the mean highest nucleotide frequency are considered collapses. If the heterozygous site ratio exceeds 20% of the mean second nucleotide frequency, these regions are considered smaller collapses. Only centromeric regions that are free of misjoins, collapses, or other assembly errors are considered accurately assembled and retained for downstream analyses.

Annotation of centromeric regions. To determine the sequence composition and organization of all centromeric regions, CenMAP (v.0.4.3.1) uses a custom tool that we developed (Snakemake-HumAS-SD v.0.1.0; https://github.com/logsdon-lab/Snakemake-HumAS-SD) that leverages the α-satellite monomer HMM libraries from HumAS-HMMER (https://github.com/fedorrik/HumAS-HMMER_for_AnVIL) with the rapid annotation of StringDecomposer[52] to efficiently annotate the centromeric regions. Snakemake-HumAS-SD first constructs a comprehensive α-satellite monomer library for each chromosome based on an existing HumAS-HMMER HMM profile (https://github.com/fedorrik/HumAS-HMMER_for_AnVIL/blob/main/AS-HORs-hmmer3.4-071024.hmm). Chromosomes with shared α-satellite monomers, such as chromosomes 1/5/19 as well as chromosomes 13/14/21/22, had their libraries merged. We developed an α-satellite HOR variation detection method that generates BED-format annotations of α-satellite HOR variants for each centromere based on the previous SV approach (https://github.com/fedorrik/stv). The method takes α-satellite monomer sequences as input and segments them based on three criteria: (1) adjacent monomers are separated by gaps >160 bp; (2) adjacent monomers are on opposite strands; or (3) monomer IDs are not monotonically increasing (on the forward strand) or decreasing (on the reverse strand). Compared with the original SV method, our approach can detect previously unresolvable α-satellite HOR structures, such as the newly identified 8-mer (S1C16H1L.2-9) expansion on chromosome 16, and annotates divergent α-satellite HORs that were not annotated using the previous SV method.

Estimation of the length of active α-satellite HOR arrays within centromeric regions. To estimate the length of active α-satellite HOR arrays, we ran CenStats (v.0.0.10) with the length subcommand, which leverages the annotations generated by Snakemake-HumAS-SD (described above) to output length statistics of each active α-satellite HOR array. CenStats first uses Snakemake-HumAS-SD α-satellite HOR SV BED files to identify the coordinates of active α-satellite HOR monomers (those with names containing 'L'). Adjacent active α-satellite HOR monomers are then merged into one block if they are within around 1.5 α-satellite monomer lengths (256 bp) of each other, and blocks containing fewer than two α-satellite monomers are removed. To account for interruptions of stretches of α-satellite HORs caused by MEIs, adjacent blocks of active α-satellite HORs separated by less than 8 kb are iteratively merged into a single active α-satellite HOR array. Finally, the coordinates of α-satellite HOR arrays that consist of at least five α-satellite HOR monomers and have at least 90% of their total length composed of active α-satellite HORs is outputted as a BED file. To determine the orientation of active α-satellite HOR arrays, we followed a similar procedure, but incorporated both strand orientation and distance to group active α-satellite HOR monomers into arrays.

Detection of CDRs within centromeric regions. In addition to annotating centromeric sequences, CenMAP (v.0.4.3.1) uses methylation information in native long-read sequencing data to annotate the CpG methylation profile across each centromeric region. CenMAP first aligns native ONT long-read sequencing data from each genome to the relevant whole-genome assembly using minimap2 (ref. 46) (v.2.28) with the following command: minimap2 -x lr:hqae -s 4000 -y -I 8G {ref} {ont_reads}. CenMAP filters the alignments using SAMtools[51] (v.1.9) -F 2308 to retain only primary alignments and remove partial and secondary alignments. Finally, CenMAP uses CDR-Finder[31] (v.1.0.4), a tool to detect hypomethylated regions termed CDRs[22], spanning from the first hypomethylated base to the last hypomethylated base and inclusive of the sequence in between. We manually curated the CDR calls by checking each region outputted in the BED file using IGV[53], removing false positives, recovering missed calls and merging calls that were within 150 kb of each other.

Pairwise sequence identity heat maps. To generate pairwise sequence identity heat maps of each centromeric region, we ran ModDotPlot[54] (v.0.8.4) with the following command: moddotplot static -f {fasta} -w 5000 -id 70.0. We normalized the colour scale across the ModDotPlots

by binning the percentage sequence identities and recolouring the datapoints according to the binning.

Visualization of centromeric regions. We also developed a Python-based tool (CenPlot v.0.1.4; https://github.com/logsdon-lab/CenPlot) for visualizing the various types of annotations produced by CenMAP. Each data type is formatted into a BED4 or BED9 file and plotted as tracks ordered in a configuration TOML file through the following command: cenplot draw -t {track_toml} -c {chrom} -d {outdir}. These tracks include RepeatMasker annotation, α-satellite HOR annotation produced by Snakemake-HumAS-SD (https://github.com/logsdon-lab/Snakemake-HumAS-SD), active α-satellite HOR array strand orientation from CenStats, CpG methylation frequencies produced by CDR-Finder[31] (v.1.0.4) as well as the one-dimensional or two-dimensional sequence identity heat maps produced by ModDotPlot[54].

Assembly repair within centromeric regions

With the ability to detect assembly errors (described above), we then developed a repair tool (AssemblyRepairer v.1.0; https://github.com/logsdon-lab/AssemblyRepairer) to repair erroneous regions within the Verkko centromere assemblies using correctly assembled regions from the hifiasm centromere assemblies. For each Verkko centromere containing errors in α-satellite sequences, we aligned the hifiasm assembly from the same individual using minimap2 (ref. 46) (v.2.28) with the following parameters: -x asm5 --eqx --cs --secondary=no -s 25000 -K 8G. Alignment records with mapping quality (MAPQ) > 20 were retained, adjacent alignments within 5 kb were merged and alignments with >30% overlap were removed. Resulting alignments <100 kb were excluded from downstream analysis. This step yielded one or more corresponding regions in the hifiasm assemblies for each erroneous Verkko centromere. NucFlag[50] (v.0.3.3) was then run to identify errors in the matched hifiasm regions. Unique *k*-mers of length 5 kb were generated from each Verkko centromere, and 1,000 were randomly selected as anchors. Anchors mapping to multiple sites in the hifiasm assemblies were discarded to ensure uniqueness. The remaining anchors were used to precisely locate erroneous regions across both assemblies. Erroneous Verkko regions were replaced with the corresponding correct hifiasm sequences, and NucFlag was run on the repaired assemblies to ensure that the regions were accurately repaired. Regions with remaining errors were excluded from downstream analysis. We finally merged the remaining centromeres that were completely and accurately assembled in hifiasm but not in Verkko with repaired Verkko centromeres and generated new whole-genome assemblies for all 65 samples. Each repaired whole-genome assembly was then validated with HMM-Flagger[19] (v.1.1.0), NucFlag[15] (v.0.3.3), GAVISUNK[20] (v.1.0) and Merqury[21] (v.1.3) using PacBio HiFi, ONT, Illumina or Element data. For HMM-Flagger, we first aligned all PacBio HiFi and ONT reads to the repaired assembly using minimap2[46] (v.2.28) with the following parameters: -y -a --eqx --cs -x lr:hqae -I8g. Each BAM file was then run through the default HMM-Flagger pipeline using the corresponding data type's alphaTsv file. We applied GAVISUNK, a singly unique *k*-mer (SUNK)-based error detection method (with parameter 30-mer), to validate α-satellite regions. Merqury (v.1.3) was used to obtain QV scores for each centromeric region with the following command: meryl k = 21 count {input_reads} output {output_db} & & merqury.sh {output_db} {input_asm}. Finally, we manually checked each centromeric region for uniform PacBio HiFi and ONT coverage, ensuring that they were free of large assembly errors.

MEI detection, annotation and curation

To detect MEIs within centromeric α-satellite HOR arrays, we ran RepeatMasker[49] (v.4.1.6) with the Dfam[55] (v.3.8) library and -s parameter to identify unique non-long terminal repeat (LTR) retroposons (*Alu*, LINE-1 and SVAs) within completely and accurately assembled centromeric regions. For each unique non-LTR retroposon, we extracted its sequence along with an additional 30 bp of flanking sequence on either side using Pysam[56,57] (v.0.22.0). We then reran RepeatMasker[49] (v.4.1.6) with the Dfam[55] (v.3.8) library and -s parameter on these FASTAs, generating an outfile. The RepeatMasker outfiles and FASTAs were inputted into L1ME-AID (v.1.2.0-beta) with the default settings for MEI identification, as previously described[13]. MEIs on the same centromeric α-satellite HOR array were aligned to one another using MUSCLE[58] (v.3.8.31), and alignments were manually curated to identify unique insertion events as well as their duplications. Finally, we intersected the MEI calls with coordinates of each centromeric α-satellite HOR array using BEDTools[47] (v.2.31.1) to derive the final callset within centromeric α-satellite HOR arrays.

Structural variation detection within centromeres

To detect large-scale variation between centromeres, we ran SVbyEye[59] (v.0.99.0). We first aligned contigs containing a centromeric region to each other using minimap2 (ref. 46) (v.2.28) and the following command: minimap2 -x asm20 --c --eqx --secondary=no {target.fasta} {query.fasta} > {output.paf}, generating a PAF file. We then visualized these alignments with R (v.4.4.2) using the plotting function: plotMiro(paf.table = paf.table, color.by = "direction").

Estimation of local sequence identity and entropy across centromeric regions

To determine the local sequence identity across each centromeric region, we averaged the pairwise sequence identity values derived from ModDotPlot[54] (v.0.8.4) BEDPE format files. For each region, genomic intervals were divided into 5 kb bins, and sequence identity values between bin pairs were used to construct a symmetric identity matrix. A sliding window of 25 kb (five consecutive bins) was applied along the diagonal of the matrix with a 5 kb step size. For each window, local sequence identity was calculated as the mean of all off-diagonal identity values, excluding those within a three-bin neighbourhood of the diagonal to reduce self-similarity bias. Average sequence identities of active α-satellite HOR arrays were calculated for arrays >50 kb. To remove the influence of outliers at array boundaries, we excluded one window from the end of the active α-satellite HOR arrays. To define regions of increased admixture within active α-satellite HOR arrays, we calculated the entropy of the entire α-satellite HOR array using the frequencies of the different types of HOR units in 10-unit windows with 1-unit slide with the following equation:

$$\text{Entropy} = -\sum (\text{frequency}_i * \log_2(\text{frequency}_i)).$$

Cell culture

The human cell lines HG00512, HG00513, HG00514, HG02769, HG03065 and GM20355 were purchased from the Coriell Institute for Medical Research and cultured in RPMI-1640 medium with 2 mM L-glutamine (Corning, MT10040CV) and supplemented with 15% FBS (Gibco, A5669701) and 1% penicillin–streptomycin (Thermo Fisher Scientific, 15140122). All of the cell lines were cultured in a humidity-controlled environment at 37 °C under 5% CO_2.

CENP-A DiMeLo-seq

DiMeLo-seq was performed as described previously[60] with some modifications for Oxford Nanopore Technologies (ONT) long-read sequencing. In brief, 7×10^6 cells were resuspended in 1 ml of Dig-Wash buffer (20 mM HEPES-KOH, pH 7.5, 150 mM NaCl, 0.5 mM Spermidine, 1 tablet of cOmplete, EDTA-free Protease Inhibitor Cocktail (11873580001, Roche) per 50 ml buffer, 0.1% (w/v) BSA and 0.02% Digitonin) and incubated on ice for 5 min to isolate nuclei. Nuclei were pelleted and resuspended in 200 µl of Tween-wash buffer (20 mM HEPES-KOH, pH 7.5, 150 mM NaCl, 0.5 mM Spermidine, 1 tablet of cOmplete, EDTA-free Protease Inhibitor Cocktail per 50 ml buffer, 0.1% (w/v) BSA and 0.1%

(w/v) Tween-20) containing a primary antibody at a 1:50 dilution. Primary antibodies used in this study were mouse anti-CENP-A antibody (ADI-KAM-CC006, Enzo) or rabbit anti-IgG antibody (ab171870, Abcam). The samples were rotated overnight on a nutator at 15 rpm and 4 °C to allow for antibody binding. A mouse nanobody-Hia5 (Nb-Hia5), gifted by N. Altemose, was prepared as a secondary antibody at 200 nM in 200 µl of Tween-wash buffer. After washing away unbound primary antibody with Tween-wash buffer, samples were resuspended in secondary antibody buffer and rotated for 2 h on a nutator at 15 rpm and 4 °C. The methyl donor *S*-adenosyl methionine (SAM; B9003S, NEB) was prepared at 800 µM in 100 µl of activation buffer (15 mM Tris, pH 8.0, 15 mM NaCl, 60 mM KCl, 1 mM EDTA, pH 8.0, 0.5 mM EGTA, pH 8.0, 0.05 mM Spermidine and 0.1% (w/v) BSA). After washing away unbound Nb-Hia5, the samples were resuspended with SAM activation buffer and incubated at 37 °C for 2 h to activate the DNA adenine methyltransferase activity of Hia5. After incubation, the samples were collected and resuspended in 400 µl of cold PBS for genomic DNA (gDNA) extraction using the Monarch HMW DNA Extraction Kit for Cells & Blood (T3050S, NEB). For gDNA extraction and ONT library preparation, the FindingNemo protocol[54] was performed with the following changes: (1) extracted gDNA was kept at room temperature for at least 48 h to improve DNA homogeneity; (2) the FRA (transposase fragmentation mix) in the Ultra-Long DNA Sequencing Kit V14 (SQK-ULK114, ONT) was adjusted by adding 8 µl for optimal DNA fragmentation; (3) library DNA was eluted in 100 µl of 10 mM Tris-HCl (pH 9.0) for a single library load. DNA libraries were loaded onto the PromethION 24 R10.4.1 flow cell (FLO-PRO114M, ONT) and sequenced with MinKNOW v.25.03.7 (ONT). To enrich centromeric sequences during sequencing, we used the deplete mode of adaptive sampling enabled in MinKNOW to exclude sequencing reads originating from outside centromeric regions. The depletion BED file was made based on centromere annotations of the CHM13 reference genome (v.2.0). Libraries were retrieved and reloaded after 20–24 h, depending on pore availability.

Both anti-CENP-A and anti-IgG DiMeLo-seq ONT data were basecalled with dorado v.0.9.6+0949eb8d using the following parameters and model: `sup,5mCG_5hmCG,6mA --min-qscore 10; basecall_model=dna_r10.4.1_e8.2_400bps_sup@v5.0.0,modbase_models=dna_r10.4.1_e8.2_400bps_sup@v5.0.0_5mCG_5hmCG@v3,dna_r10.4.1_e8.2_400bps_sup@v5.0.0_6mA@v3`. Reads with abnormally high N^6-methyladenine (6mA) methylation proportions (number of 6mA sites with a base modification probability greater that 80%/total number of AT sites) were filtered out to reduce noise.

We determined outlier reads by calculating the global mean and s.d. of the proportion of 6mA over all reads and calculating a standard score for each read. Any read with greater than two *z*-scores above the mean was filtered using SAMtools[51]. Filtered reads were then aligned to their respective assemblies using minimap2 (ref. 46) (v.2.28) with the following parameters: `-y -a --eqx --cs -x lr:hqae -I8g -s 4000`. To quantify 6mA methylation, we applied modkit (v.0.3.2, https://github.com/nanoporetech/modkit) using the pileup command with the following parameters: `modkit pileup {input.bam} {output.bed} --ref {input.fa} --motif AT 0 --filter-threshold 0.8`. Finally, using a custom pipeline, DiMeLo-seq-Viz (v.0.1.0, https://github.com/logsdon-lab/Snakemake-DiMeLo-seq-Viz), we binned each 6mA window into 5 kb non-overlapping windows, calculated the average 6mA signal and subtracted the average IgG signal from the average CENP-A signal to determine the normalized CENP-A signal per sample. The resulting BEDgraph file was visualized using CenPlot (v.0.1.4).

CENP-A CUT&RUN and ONT long-read sequencing

CUT&RUN was performed following the CUTANA protocol developed by EpiCypher with some modifications. For each reaction, 550,000 cells were collected and nuclei were isolated following the nucleus extraction protocol for CUTANA assays (EpiCypher). Nuclei were immobilized with activated Concanavalin A (ConA) beads provided in the CUTANA ChIC/CUT&RUN Kit (14-1048, EpiCypher). After binding to ConA beads, nuclei were incubated overnight at 4 °C in 50 µl of Antibody buffer (1× proteinase inhibitor, 0.5 mM Spermidine, 0.01% (w/v) Digitonin and 2 mM EDTA in pre-wash buffer) with either mouse anti-CENP-A antibody (ADI-KAM-CC006, Enzo) or a rabbit anti-IgG antibody (ab171870, Abcam) at a 1:50 dilution. An appropriate amount of K-MetSat Panel was added to the reaction before adding the IgG negative control. After overnight incubation at 4 °C, rotating on a nutator at 15 rpm, the samples were washed and incubated in cell permeabilization buffer (1× proteinase inhibitor, 0.5 mM Spermidine, and 0.01% (w/v) Digitonin in pre-wash buffer) with 2.5 µl pAG-MNase (15-1016, EpiCypher) at 1:20 dilution for 10 min at room temperature. In our experience, a reduction in pAG-MNase (for example, to 0.25 µl or 0.05 µl) can increase fragment lengths (up to 1.5-fold) but often also reduces sequencing yield (to around 20%). We therefore chose to use the original volume of pAG-MNase in our reactions to obtain long reads while reducing sequencing costs. After pAG-MNase was bound to the primary antibody, 100 mM $CaCl_2$ (EpiCypher) was added to activate chromatin-targeted digestion at 4 °C for 2 h. Digestion was stopped by adding stop buffer (EpiCypher), and the samples were transferred to low-bind tubes for DNA purification and extraction following the fragment release and cleanup in the protocol of ref. 61. Each sample was brought up to 300 µl with 0.1× TE (EpiCypher), followed by addition of 3 ul of 10% SDS and 2.5 µl of 20 mg ml^{-1} proteinase K (P8107S, NEB), then incubated at 70 °C for 10 min.

DNA extraction was performed using one phenol:chloroform:isoamyl alcohol (25:24:1 (v/v); 15593031, Invitrogen) extraction step, followed by transfer of the aqueous phase to a MaXtract High Density 2 ml tube (129056, Qiagen), then one chloroform (C2432, Sigma-Aldrich) extraction step, again transfer the aqueous phase to a MaXtract tube. DNA was precipitated with 750 µl of 100% ethanol and 2 µl of glycogen (AM9510, AMB) and stored overnight at −20 °C. CUT&RUN-enriched DNA was washed with 100% ethanol and eluted in 0.1× TE.

For ONT library preparation and sequencing, 0.5–5 ng of CUT&RUN-enriched DNA was amplified using CUTANA CUT&RUN Library Prep Kit (14-1001 and 14-1002, EpiCypher). Up to the indexing PCR step, the KOD Xtreme Hot Start DNA polymerase (71975-3, Sigma-Aldrich) was used for fragment enrichment. Each 50 µl of PCR reaction contains 2× KOD PCR buffer, 0.4 mM dNTPs, 1.5 µl of i7 primer, 1.5 µl of i5 primer and 1 µl of KOD polymerase. The PCR cycling conditions were as follows: 2 min initial denaturation at 94 °C; 20 cycles of 10 s denaturation at 98 °C, 30 s primer annealing at 55 °C and 5 min extension at 68 °C; followed by a final 5 min extension at 68 °C. Note that we did not perform a size-selection step during library preparation to preserve long DNA fragments. The sequencing library was prepared using the ONT SQK-NBD114.24 kit and sequenced on a PromethION 24 using R10.4.1 flow cells (FLO-PRO114M, ONT).

The resulting CUT&RUN ONT data were processed using a custom pipeline we developed called CUT&RUN-Viz (v.0.0.1; https://github.com/logsdon-lab/Snakemake-CutNRun). In this pipeline, adapters are first trimmed using cutadapt[62] (v.5.1) with the following parameters: `-e 0.1 -q 20 -O 5 -a AGATCGGAAGAGC {reads.fastq}`. Reads <100 bp or >10 kb are then removed using seqkit[63] (v.2.10.0) with the following command: `seq -m 100 -M 10000 {reads.fastq}`. The reads are aligned to the relevant whole-genome assembly using BWA-MEM[64] (v.0.7.19) with the following parameters: `-k 50 -c 1000000 {ref.fasta} {reads.fastq}`. Finally, CENP-A alignments are normalized to IgG alignments using deepTools[65] (v.3.5.6) with the following command: `bamCompare -b1 {bam_CENPA} -b2 {bam_IgG} --operation ratio --binSize 5000 --minMappingQuality 0`. The resulting BigWig is visualized with CenPlot (v.0.1.4, https://github.com/logsdon-lab/CenPlot).

FISH

To detect a hybrid chromosome 13/22 centromere in human GM20355 cells, we performed FISH on metaphase chromosome spreads as

previously described[7,66] with slight modifications. In brief, human GM20355 cells were treated with colcemid and swollen in a 0.56% potassium chloride (KCl) hypotonic solution at 37 °C. After 30 min, cells were fixed with methanol:acetic acid (3:1) and dropped onto previously cleaned slides. Slides were aged in a drying oven at 90 °C for 1.5 h, treated with pepsin (0.005% in 0.01 N HCl) at 37 °C for 30 min, hybridized with labelled FISH probes at 70 °C for 2 min and incubated overnight at 37 °C.

The FISH probes used to validate the hybrid chromosome 13/22 centromere were either derived from a pZ21A plasmid clone harbouring a portion of the *D21Z1* α-satellite HOR array[67] (Cy3; red) or generated by PCR using primers targeting the p arm portion of the S2C14/22H1L α-satellite HOR array of chromosomes 14 and 22 (FluorX; green) according to the following procedure. First, a PCR reaction was prepared in a total volume of 25 µl, containing 0.25 µM forward primer (5′-GGTCCTCAAACCGTCCGAAA-3′), 0.25 µM reverse primer (5′-CAGCTTTCAGGTCTATGGTGAG-3′), 5 ng µl^{-1} human genomic DNA, 50% DreamTaq PCR Master Mix 2× (DreamTaq DNA polymerase, 2× DreamTaq buffer, dATP, dCTP, dGTP and dTTP at 0.4 mM each, and 4 mM $MgCl_2$), and double-distilled water. PCR was performed using the following thermal cycling conditions: initial denaturation at 94 °C for 30 s, followed by 30 cycles of 94 °C for 30 s (denaturation), 65 °C for 1 min (annealing) and 72 °C for 1 min (extension). After the last cycle, a final extension was carried out at 72 °C for 10 min, and the reaction was held at 4 °C. The resulting PCR product was subsequently labelled using PCR in a 25 µl reaction containing 2 µl PCR product, 10% Taq buffer with $(NH_4)_2SO_4$ (10×), 2 mM $MgCl_2$, 0.2 µM forward primer, 0.2 µM reverse primer, 0.2 mM dACG, 0.1 mM fluorescein-dUTP, 0.4% BSA, 0.06 U µl^{-1} Taq DNA polymerase and double-distilled H_2O. PCR amplification was performed as described above.

The slides were washed with high stringency three times in 0.1× SSC at 60 °C for 5 min each before mounting in Antifade and DAPI-staining. The slides were imaged on an epifluorescence microscope (Leica DM RXA2) equipped with a charge-coupled device camera (CoolSNAP HQ2) and a ×100/1.6–0.6 NA objective lens. DAPI, Cy3 and fluorescein fluorescence signals, detected through specific filters, were individually collected as greyscale images using the iVision software (v.4.5). Pseudocolouring and merging of the images were performed using the Adobe Photoshop software (v.25.11.0).

To independently validate these findings, we performed a second round of FISH using different fluorescent probes specific to the chromosome 13/21 and 14/22 α-satellite DNA. We first prepared metaphase chromosome spreads from GM20355 cells by adding KaryoMAX Colcemid Solution (0.1 µg ml^{-1}, Life Technologies) to the medium for 6–7 h and collecting cells by centrifugation at about 250*g* for 5 min. Collected cells were incubated in hypotonic 0.4% KCl solution for 12 min and prefixed by addition of methanol:acetic acid (3:1) fixative solution (1% total volume). Prefixed cells were spun down and then fixed in methanol:acetic acid (3:1). Cells were dropped on a glass slide and incubated at 65 °C overnight. Before hybridization, the slides were dehydrated in a 70%, 80% and 100% ethanol series for 2 min each. The slides were denatured in 70% deionized formamide/2× SSC solution preheated to 72 °C for 1.5 min. Denaturation was stopped by immersing the slides in a 70%, 80% and 100% ethanol series chilled to −20 °C. Labelled DNA probes were denatured separately in a hybridization buffer by heating to 80 °C for 7 min before applying to denatured slides. Fluorescently labelled human CenSat probes for *D13Z1*/*D21Z1* (LPE013R) and *D14Z1*/*D22Z1* (LPE014G) were purchased from CytoCell. The specimens were hybridized to the probes under a HybriSlip hybridization cover (GRACE Biolabs) and sealed with rubber cement or Cytobond (SciGene) in a humidified chamber at 37 °C for 48–72 h. After hybridization, the slides were washed in 50% formamide/2× SSC three times for 5 min per wash at 45 °C, then in 1× SSC solution at 45 °C for 5 min twice and at room temperature once. The slides were mounted in Vectashield containing DAPI (Vector Laboratories). Confocal *z*-stack images were acquired on the Nikon TiE microscope equipped with a PlanApo ×100 oil-immersion objective NA 1.45, Yokogawa CSU-W1 spinning disk, ORCA-Fusion BT sCMOS camera (Hamamatsu) and NIS Elements software.

To detect the inner-kinetochore protein CENP-C on the recombined chromosome 13/22, freshly prepared chromosome spreads were dropped on a glass slide, washed with PBS/0.1% Triton X-100 and blocked with 5% BSA in PBS/0.1% Triton X-100. Primary antibody (rabbit polyclonal anti-CENP-C, ABE1957, Millipore) and secondary antibody (goat anti-rabbit Alexa Fluor Plus 647, A32733TR, Thermo Fisher Scientific) were diluted 1:100 and 1:500, respectively, in 5% BSA/PBS/0.1% Triton X-100. The samples were incubated with primary antibody overnight, washed three times for 5 min, incubated with secondary antibodies for 2–4 h and washed again three times for 5 min. All of the washes were performed with PBS/0.1% Triton X-100. After antibody incubation, spreads were post-fixed in 2% paraformaldehyde diluted in PBS for 15 min, washed in PBS, and processed for CenSat FISH as described above, starting with ethanol dehydration series. DNA was stained with 1.5 µg ml^{-1} DAPI. For structured illumination super-resolution microscopy (SIM) performed on CENP-C immuno-FISH, slides were rinsed in double-distilled H_2O, air-dried in the dark, mounted in ProLong Glass antifade mountant (Thermo Fisher Scientific) and allowed to cure for at least 24 h before imaging. *z*-stack images were acquired on an Elyra 7 Lattice SIM^2 microscope (Zeiss) equipped with two PCO.edge 4.2 sCMOS cameras, four high-power continuous wave lasers (405, 488, 561 and 642 nm) and a Zeiss PlanApo ×63 oil-immersion objective NA 1.4. The illumination pattern was set to 15 phases, and the *z*-stack spacing was set at 100 nm. Raw SIM images were reconstructed using the ZEN Black software (Zeiss). For the CENP-C channel, sharpness adjustment of 10.5 with best-fit settings was used; for the fluorescence probe channels, standard sharpness with best-fit settings were sued, and the DAPI channel was processed in the widefield mode. All image processing was performed using ImageJ/FIJI. Maximum-intensity projections of FISH images from spinning-disk confocal *z*-stacks were generated. Reconstructed SIM images were also maximum-intensity projected from the best focal planes, except for the DAPI channel, for which the slice of highest contrast was selected. Chromosomes of interest were manually segmented, oriented vertically and shown in the composite mode.

Phylogenetic reconstruction of centromeres

To reconstruct the evolutionary history of centromeres, we first identified orthologous sequences between human and chimpanzee by aligning the CHM13 (v.2.0) genome to the mPanTro3 (v.2.1) genome[37] using minimap2 (ref. 46) (v.2.28) with the following parameters: `-ax asm20 --eqx -Y --secondary=no`. From this set of orthologous sequences, we selected 20 kb regions in the monomeric α-satellite or unique sequences flanking the α-satellite HOR arrays in both the p and q arms, prioritizing regions closest to the α-satellite HOR arrays. For acrocentric chromosomes (13–15, 21 and 22), only q arm regions were included owing to the high abundance of satellites and segmental duplications on the p arm as well as the known homologous recombination between acrocentric chromosomes in the p arms that prevent accurate phylogenetic reconstruction[23]. We aligned the 20 kb CHM13 sequences to the 65 repaired human genome assemblies as well as the mPanTro3 (v.2.1) genome[37] using the same minimap2 parameters described above. We limited our analyses to only those regions with one-to-one unambiguous mapping, and we excluded sequences with alignment lengths >10% of the aligned region (2 kb). A multiple-sequence alignment was performed using MAFFT[68] (v.7.526) with the `--auto` setting. Maximum-likelihood phylogenies were inferred using IQ-TREE[69] (v2.3.6) with automatic model selection (`--auto`) and 1,000 ultrafast bootstrap replicates, using the chimpanzee sequence as an outgroup. Divergence times were estimated under a molecular clock model by scaling branch lengths, assuming a human–chimpanzee divergence of 6.2 million

years[37]. Phylogenetic trees were visualized using iTOL[70] (v.7.2.2; https://itol.embl.de/).

Identification of major centromere haplotypes

To identify major centromere haplotypes, we first compared the phylogenetic trees from both p and q arms and selected the tree that exhibited higher topological congruence with the distribution and architecture of the active α-satellite HOR array as well as higher bootstrap values. We then partitioned the tree based on a divergence time of 100 ka. The 100 ka threshold was determined following a comprehensive inspection of all centromere phylogenies, providing an optimal balance between intraclade sequence homogeneity and population-level stratification. Finally, we manually reviewed each clade to identify haplotypes that were distinct from those in other clades. One haplotype from each clade was selected as a representative of that clade and included in our list of major centromere haplotypes.

Analysis of archaic centromere introgression

We applied ARCkmerFinder[38] (v.1.1; https://github.com/hsiehphLab/ARCkmerFinder), a *k*-mer-based method, to whole-genome assemblies for introgression detection. In brief, we generated *k*-mer databases for each of the four archaic genomes[39–42] and a joint *k*-mer database from African genomes in the 1000 Genomes short-read collection[71] using Meryl[21] (v.1.4.1; *k*-mer size = 15 bp). As low-frequency *k*-mers are more likely to represent sequencing errors, we modelled the observed *k*-mer frequency distribution for each short-read genome using a mixture of negative binomials to determine appropriate cut-off thresholds[38]. Archaic-specific *k*-mers were identified by subtracting the joint African *k*-mers from each archaic database, leveraging the limited archaic introgression in Africans to reduce ancestral polymorphism effects. For each assembly, we scanned contig sequences for the presence of archaic-specific *k*-mers and binned the signal within 2 kb windows to aggregate local evidence of archaic sequence. Genome-wide cut-offs for detecting archaic introgression were then determined from whole-genome estimates of archaic sequence content.

Pan-centromeric mutation rate estimation

To estimate the mutation rates of active α-satellite HOR arrays and flanking sequences for each centromere haplogroup, we first identified clades within the phylogenetic tree that contain at least four centromeres, from which one was selected as a reference and the remaining as queries. We then aligned the reference's active α-satellite HOR array, along with 2 Mb of the adjacent pericentromeric region, to each query in the same clade using minimap2 (ref. 46) (v.2.28) with the following parameters: `-x asm20 -K 8G --eqx -Y -r 500000,500000`. Unmapped, supplementary and partial alignments were removed using SAMtools[51] (v.1.22) flag 260. We then generated a BED file of 10 kb windows located within each reference centromere. We used subseq (https://github.com/EichlerLab/subseq) to extract query sequences. For each window, the corresponding query sequence was required to be between 8,000 and 12,000 bp in length and exhibit a one-to-one, unambiguous mapping. For sequences in each window, a multiple-sequence alignment was performed using MAFFT[68] (v.7.526) with the following parameters: `--maxiterate 1000 --globalpair`. The mutation rate per segment was estimated based on Kimura's model of neutral evolution[72]. In brief, we modelled the estimated divergence (*D*) as a result of between-species substitutions and within-species polymorphisms, assuming demographic equilibrium and equal mutation rates across sexes, that is,

$$D = 2\mu t + cN_e\mu$$

where μ is the mutation rate, t is the divergence time estimated from our phylogenetic trees, c is the chromosome ratio (4 for autosomes, 3 for chromosome X and 1 for chromosome Y) and N_e is the ancestral human effective population size. We assume a generation time of 20 years and a human–chimpanzee divergence time of 6.2 million years. To convert the genetic unit to a physical unit, our computation also assumes $N_e = 10{,}000$ and uniformly drawn values for the generation and divergence times. The estimated divergence (*D*) is calculated from the Tamura–Nei genetic distance using the libtn93 library (https://github.com/sdwfrost/libtn93?tab=readme-ov-file#api) with a match mode of RESOLVE and a minimum overlap of 100 characters.

Detection of de novo SNVs and SVs among centromeres across generations

To detect SNVs and SVs within centromeric regions across generations (in the CEPH 1463 pedigree), we first aligned each child's centromere to its corresponding parental centromere using minimap2 (ref. 46) (v.2.28) with the following parameters: `-ax asm5 --eqx --cs --secondary=no -s 25000 -K 8G`. Only high-confidence, uniquely mapped alignments were retained for downstream analysis. SNVs and SVs were then detected by parsing the CIGAR string. SNVs were identified from 'X' operations in the CIGAR string, where each 'X' represents a single base mismatch. SVs were identified from insertions (I) or deletions (D) of at least 50 bp in the CIGAR string. To validate the SNVs and SVs, we manually inspected each event using IGV plots generated from both PacBio HiFi and ONT read alignments for the child and parent, enabling us to exclude potential artifacts owing to mosaicism or alignment errors.

Multigenerational centromeric mutation rate estimation

To estimate the mutation rate among centromeric regions (in the CEPH 1463 pedigree), we first annotated the sequence of each centromeric region using RepeatMasker[49] (v.4.1.0) and Snakemake-HumAS-SD (https://github.com/logsdon-lab/Snakemake-HumAS-SD; v0.1.0), Tandem repeats finder[73] (TRF; v.4.09), and CDR-Finder[31] (v.1.0.4). We then intersected the validated SNVs identified from minimap2 alignments (described above) with each of these regions using BEDTools[47] (v.2.31.1). We divided the number of validated SNVs by the total number of base pairs in the parent for each region per generation to derive the estimated mutation rate, defined as the number of mutations per base pair per generation.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All datasets generated and/or used in this study are listed in Supplementary Table 24 with their BioProject accession number (if available) and/or URL. For convenience, we also list their BioProject accession numbers and/or URLs here: 65 original human whole-genome assemblies generated by the HGSVC (PRJEB76276 and PRJEB83624); 65 human whole-genome assemblies with repaired centromeric contigs (PRJNA1309774); PacBio HiFi and ONT data from 65 humans generated by the HGSVC (PRJEB58376, PRJEB75216, PRJEB77558, PRJEB75190, PRJNA698480, PRJEB75739, PRJEB36100, PRJNA988114, PRJNA339722, PRJEB41778, ERP159775); Illumina data from 65 humans generated by the 1000 Genomes Project, Genome in a Bottle Consortium and the International HapMap Consortium (PRJNA200694, ERP114329, ERP120144); CENP-A and IgG DiMeLo-seq data from 6 humans (SRR35424754, SRR35424755, SRR35269316, SRR35269317, SRR35269314, SRR35269315, SRR35144393, SRR35144394, SRR35424752, SRR35424753, SRR35048265, SRR35048266), CENP-A and IgG CUT&RUN ONT data from GM20355 (SRR35048263, SRR35048264); 227 original human whole-genome assemblies generated by the HPRC (PRJNA730822); PacBio HiFi data from 227 humans generated by the HPRC (PRJNA701308, PRJNA731524); ONT data from 227 humans generated by the HPRC (PRJNA701308, PRJNA731524); 28 human whole-genome assemblies, PacBio HiFi data and ONT data from the CEPH 1463 pedigree (PRJEB86317 or dbGaP

accession: phs003793.v1.p1); Element data from the CEPH 1463 pedigree (PRJEB3246); CHM13 whole-genome assembly (PRJNA559484); CHM1 whole-genome assembly (PRJNA975207); Altai Neanderthal (PRJEB1265); Vindija Neanderthal (PRJEB21157); Chagyrskaya Neanderthal (http://ftp.eva.mpg.de/neandertal/Chagyrskaya/); Denisovan (PRJEB3092); and chimpanzee (AG18354) whole-genome assembly (PRJNA916736). Moreover, centromere sequences and annotations for the 65 whole-genome assemblies containing repaired centromeres (HGSVC Phase 3) as well as 227 whole-genome assemblies (HPRC Release 2) and 17 CEPH 1463 whole-genome assemblies containing repaired centromeres consented for public release are available at Figshare[74] (https://doi.org/10.6084/m9.figshare.29921990), and original imaging data underlying this manuscript can be accessed from the Stowers Original Data Repository (https://www.stowers.org/research/publications/libpb-2621). Finally, the centromeres described in this study can be visualized at www.cencyclopedia.com—an interactive and dynamic centromere browser.

Code availability

All custom code used in this project are publicly available at the following GitHub repositories: AssemblyRepairer (v.1.0; https://github.com/logsdon-lab/AssemblyRepairer); CenMAP (v.0.4.3.1; https://github.com/logsdon-lab/CenMAP); CenStats (v.0.0.10; https://github.com/logsdon-lab/CenStats); NucFlag (v.0.3.3; https://github.com/logsdon-lab/NucFlag); CenPlot (v.0.1.4; https://github.com/logsdon-lab/CenPlot); Snakemake-HumAS-SD (v.0.1.0; https://github.com/logsdon-lab/Snakemake-HumAS-SD); DiMeLo-seq-Viz (v.0.1.0; https://github.com/logsdon-lab/Snakemake-DiMeLo-seq-Viz); CUT&RUN-Viz (v.0.0.1; https://github.com/logsdon-lab/Snakemake-CutNRun); Pan-centromeric mutation rate estimation (v.0.1.0; https://github.com/logsdon-lab/Snakemake-MutationRate); and Cencyclopedia (v.0.0.1; https://github.com/logsdon-lab/Cencyclopedia).

46. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
47. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
48. Li, H. Identifying centromeric satellites with dna-brnn. *Bioinformatics* **35**, 4408–4410 (2019).
49. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0 (2013).
50. Vollger, M. R. et al. Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88 (2019).
51. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
52. Dvorkina, T., Bzikadze, A. V. & Pevzner, P. A. The string decomposition problem and its applications to centromere analysis and assembly. *Bioinformatics* **36**, i93–i101 (2020).
53. Robinson, J. T. et al. Integrative Genomics Viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
54. Sweeten, A. P., Schatz, M. C. & Phillippy, A. M. ModDotPlot—rapid and interactive visualization of tandem repeats. *Bioinformatics* **40**, btae493 (2024).
55. Hubley, R. et al. The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**, D81–D89 (2016).
56. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
57. Bonfield, J. K. et al. HTSlib: C library for reading/writing high-throughput sequencing data. *Gigascience* **10**, giab007 (2021).
58. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
59. Porubsky, D. et al. SVbyEye: a visual tool to characterize structural variation among whole-genome assemblies. *Bioinformatics* **41**, btaf332 (2025).
60. Altemose, N. et al. DiMeLo-seq: a long-read, single-molecule method for mapping protein–DNA interactions genome wide. *Nat. Methods* **19**, 711–723 (2022).
61. Hainer, S. J. & Fazzio, T. G. High-resolution chromatin profiling using CUT&RUN. *Curr. Protoc. Mol. Biol.* **126**, e85 (2019).
62. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
63. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE* **11**, e0163962 (2016).
64. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at http://arxiv.org/abs/1303.3997 (2013).
65. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
66. Lichter, P. et al. High-resolution mapping of human chromosome 11 by in situ hybridization with cosmid clones. *Science* **247**, 64–69 (1990).
67. Archidiacono, N. et al. Comparative mapping of human alphoid sequences in great apes using fluorescence in situ hybridization. *Genomics* **25**, 477–484 (1995).
68. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
69. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
70. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128 (2007).
71. Byrska-Bishop, M. et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–34409 (2022).
72. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, 1983).
73. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
74. Shenghan, G. et al. Data for 'A global view of human centromere variation and evolution'. *Figshare* https://doi.org/10.6084/m9.figshare.29921990 (2026).
75. Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026–1028 (2007).

Acknowledgements We thank E. E. Eichler for advice, discussions and guidance; Y. Nechemia-Arbely, M. A. Mahlke, G. A. Hartley, R. J. O'Neill, Z. Liu, P. W. Hook, W. Timp and N. Altemose for reagents, protocols and advice; and the members of the Microscopy Technology Center at the Stowers Institute for enabling many of our experiments.

Author contributions S.G., K.K.O. and G.A.L. conceived the project. S.-C.C., T.A.P., A.M. and M.V. performed experiments. S.G., K.K.O., M.L., T.A.P., D.S.G., Z.Y. and G.A.L. analysed the data. Y.M., P.H., J.L.G., M.K.K., M.V. and G.A.L. supervised the project. S.G., K.K.O. and G.A.L. wrote, edited and revised the manuscript. All of the authors have read and approved the manuscript. The Members of the HGSVC and HPRC selected samples, produced sequencing data and generated and evaluated the genome assemblies.

Funding This research was supported, in part, by funding from the National Institutes of Health (NIH) National Institute for General Medical Sciences (NIGMS) R00GM147352 (to G.A.L.) and 1P20GM139769 (to M.K.K. and M.L.). This work was also supported, in part, by funding from the NIH National Cancer Institute (NCI) R01CA266339 (to J.L.G.) and R50CA305001 (to T.A.P.) and the NIH National Human Genome Research Institute (NHGRI) 5R00HG011041 (to P.H.) and U41HG010972, U01HG010971, U01HG013760, U01HG013755, U01HG013748, U01HG013744 and R01HG011274 (to the Human Pangenome Reference Consortium; BioProject: PRJNA730823). This work was also supported by the Searle Scholars Program (to G.A.L.) and the Stowers Institute for Medical Research (to J.L.G.).

Competing interests The other authors declare no competing interests.

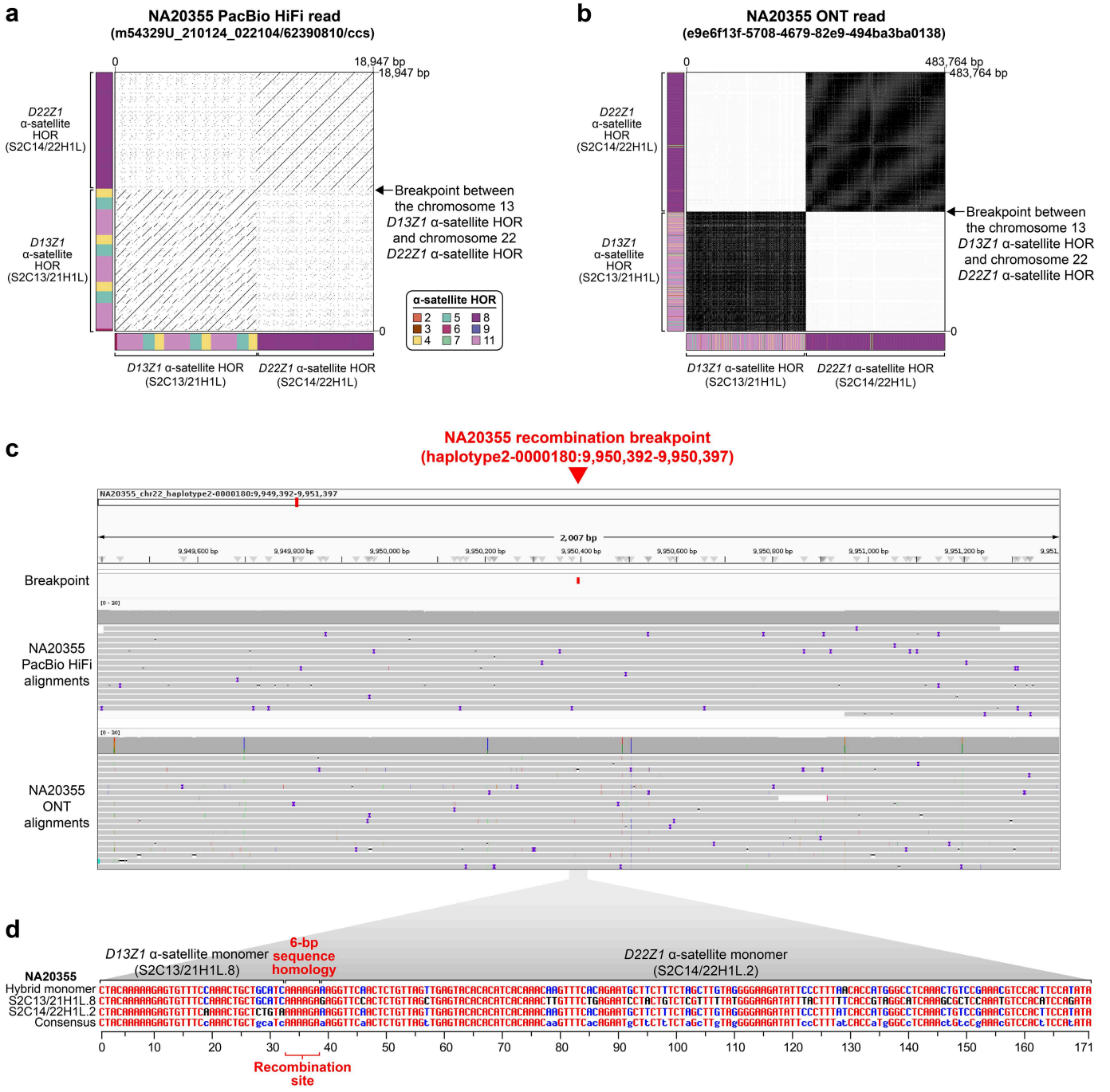
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-026-10841-9.

Correspondence and requests for materials should be addressed to Glennis A. Logsdon.

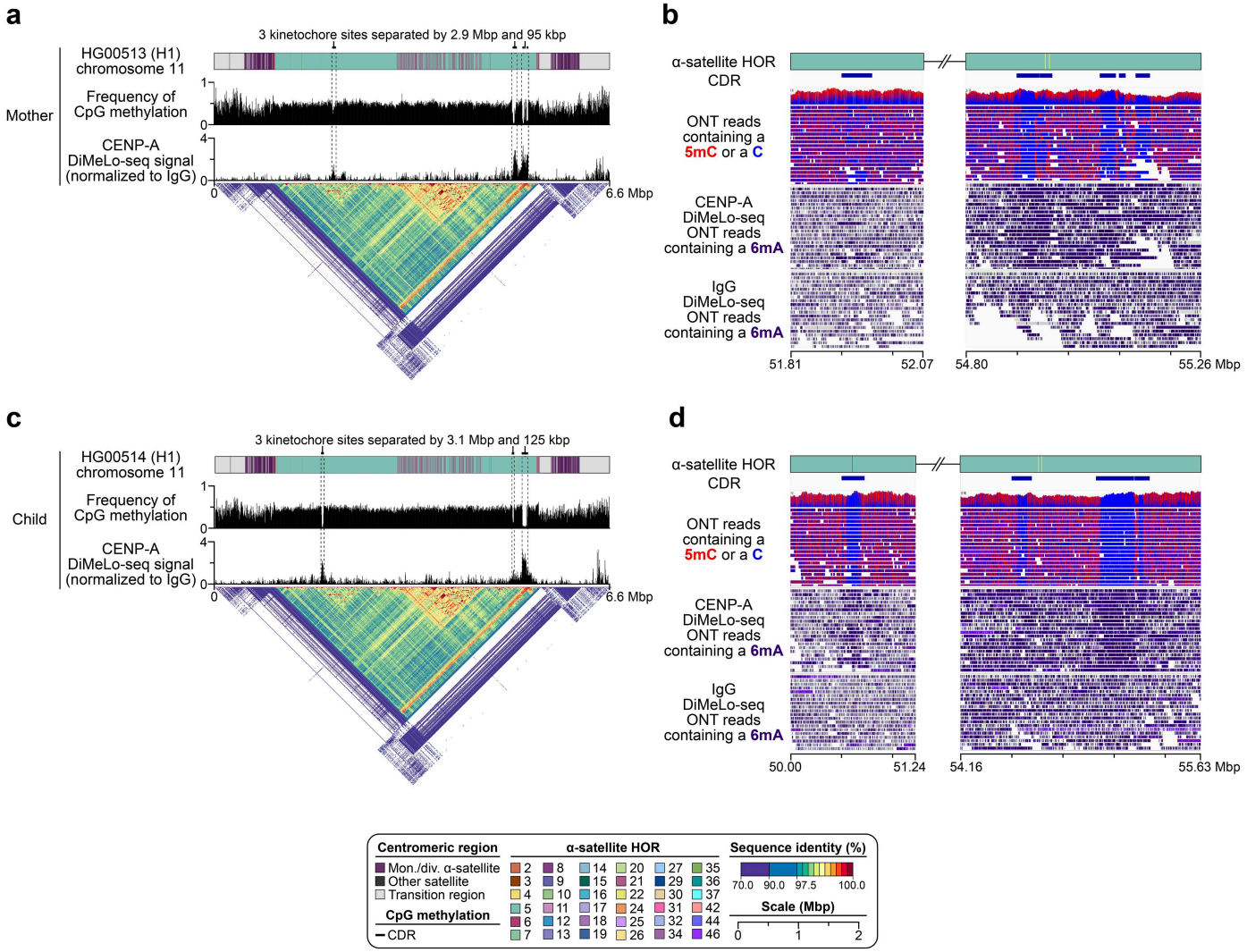
Peer review information *Nature* thanks Gautam Dey, Daniele Fachinetti and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at http://www.nature.com/reprints.

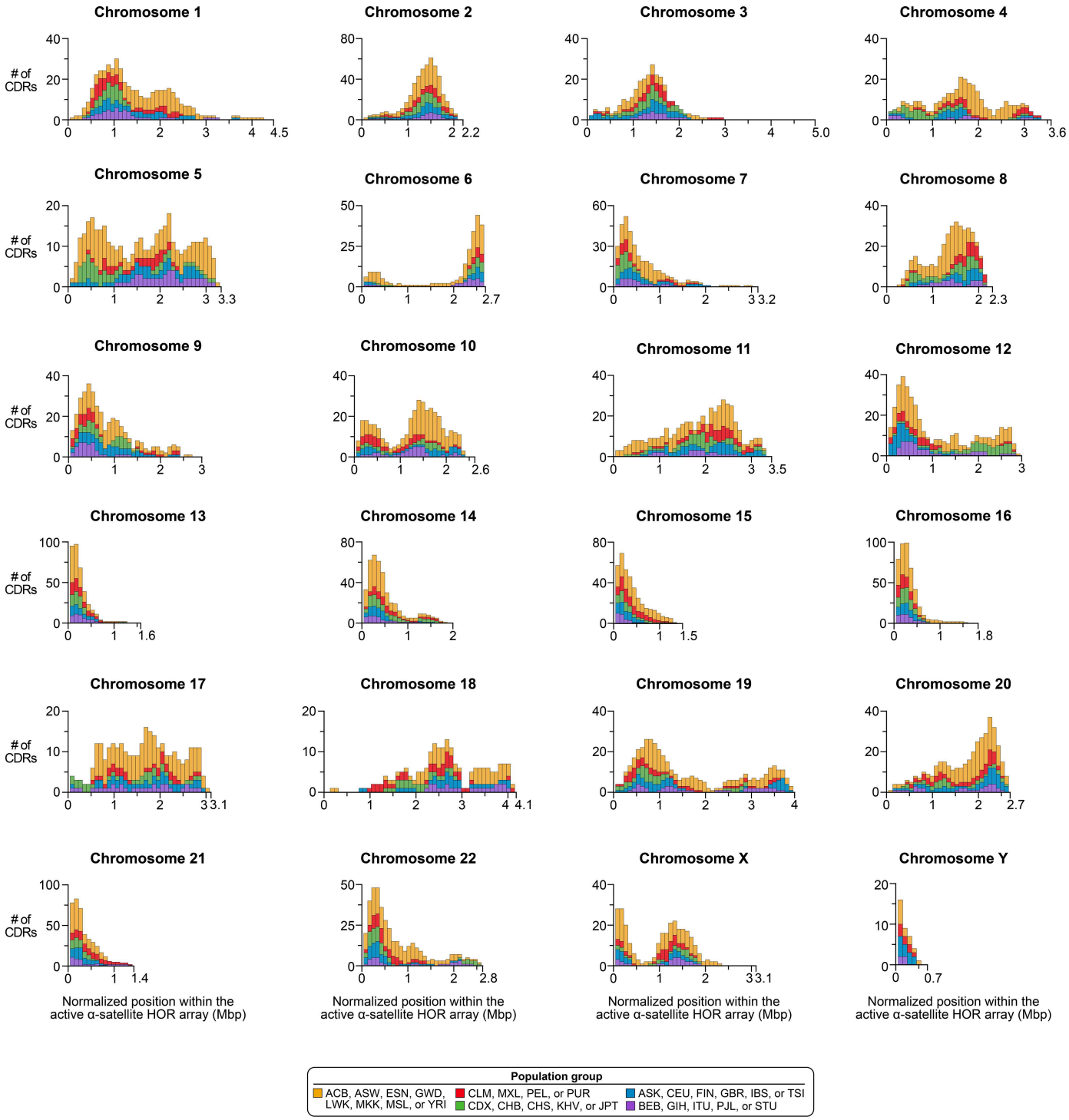


Extended Data Fig. 1 | See next page for caption.

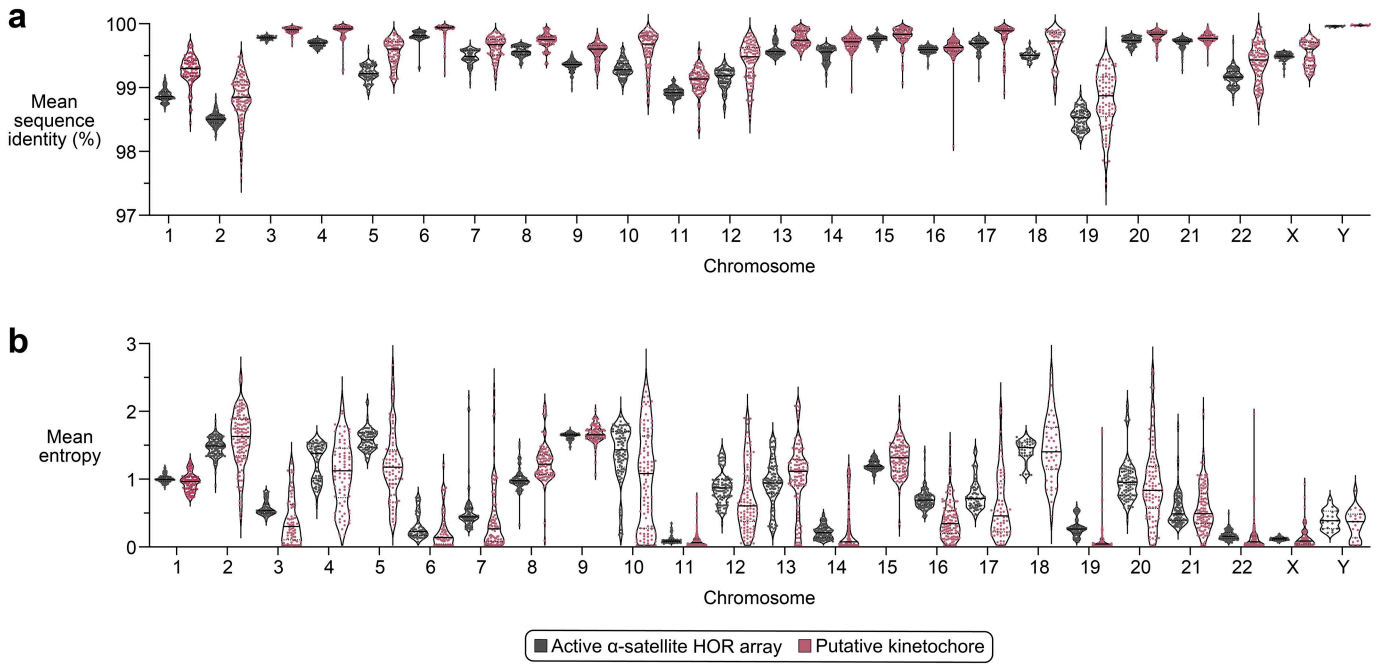
Extended Data Fig. 1 | Evidence of a recombination event between chromosome 13 and chromosome 22 centromeres in NA20355, giving rise to a hybrid 13/22 chromosome. a,**b**) Dot plots of **a**) a PacBio HiFi read and **b**) an ONT read showing the transition between chromosome 13 *D13Z1* and chromosome 22 *D22Z1* α-satellite HORs in NA20355. Word length, 20 bp for the PacBio HiFi read, 50 bp for the ONT read, generated with Gepard[75]. **c**) Detection of the breakpoint within NA20355 and support for the junction with raw NA20355 PacBio HiFi and ONT reads. **d**) Comparison of the NA20355 hybrid monomer to the S2C13/21H1L.8 and S2C144/22H1L.2 monomers up and downstream of the breakpoint. A 6-bp homology sequence (AAAAGA) is found among all three and likely contributes to microhomology-mediated recombination, leading to the hybrid 13/22 chromosome. **e**) Fluorescent in situ hybridization (FISH) on a GM20355 metaphase chromosome spread showing the recombined chromosome 13/22. Fluorescent probes specific to chromosome 13/21 α-satellite (red) and 14/22 α-satellite (green) colocalize on the recombined chromosome 13/22 in 100% of cells ($n = 20$). Scale bar (chromosome spread), 10 µm. Insets are magnified 6.5×. **f**) Immunolabeling combined with FISH (immuno-FISH) on a GM20355 metaphase chromosome spread showing CENP-C localization on chromosome 13/21 α-satellite DNA (red) of the recombined chromosome 13/22 ($n = 30$). Scale bar (chromosome spread), 10 µm; scale bar (individual chromosome), 1 µm.



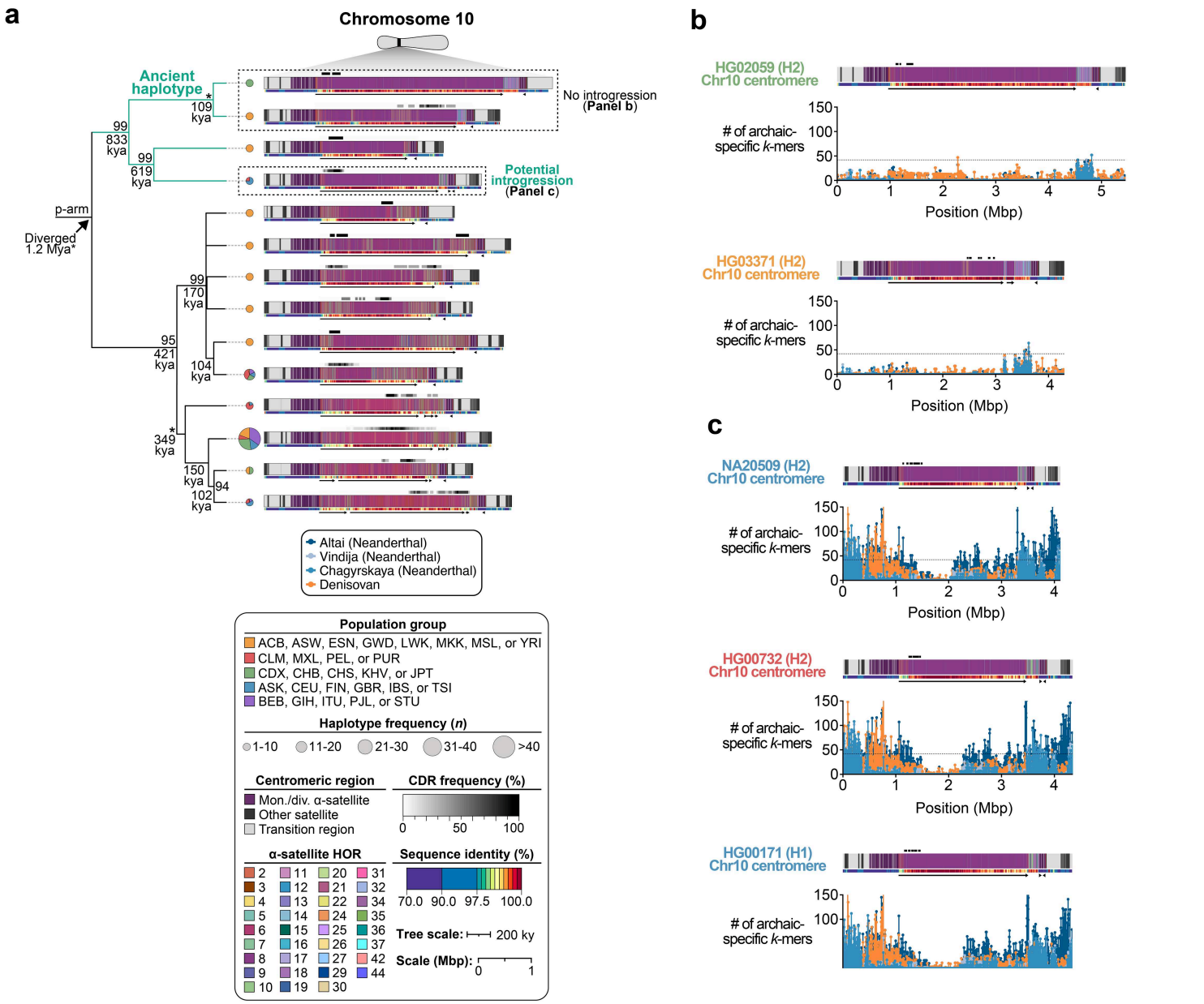
Extended Data Fig. 2 | Di-kinetochores separated by 2.9 Mbp of sequence are stably transmitted across generations. a) Plot showing multiple kinetochore sites within the HG00513 chromosome 11 centromere (H1), with two of them separated by ~2.9 Mbp of sequence and the third separated by 95 kbp of sequence. **b**) Raw ONT sequencing reads showing hypomethylated regions enriched with CENP-A within the HG00513 chromosome 11 centromere (H1), as determined with DiMeLo-seq. **c**) Plot showing that multiple kinetochores within the chromosome 11 centromere are transmitted to the child, HG00514, with slight changes in position. **d**) Raw ONT sequencing reads showing hypomethylated regions enriched with CENP-A within the HG00514 chromosome 11 centromere (H1), as determined with DiMeLo-seq.



Extended Data Fig. 3 | The position of the putative kinetochore is closely associated with the underlying sequence and structure of the centromere. The position of the CDR across active α-satellite HOR arrays for chromosomes 1–22, X and Y. For 11 out of 24 chromosomes (chromosomes 1, 3, 7, 9, 13–16, 21, 22, and Y), CDRs are mainly located on the p arm-proximal side of the array. For 2 out of 24 chromosomes (chromosomes 2 and 20), CDRs are mainly located on the q arm-proximal side of the array. For the remaining 11 out of 24 chromosomes (chromosomes 4–6, 8, 10–12, 17–19, and X), CDRs are located on different regions of the α-satellite HOR array depending on their haplotype structure.



Extended Data Fig. 4 | Sequences underlying the putative kinetochore have a higher mean sequence identity and lower entropy than the rest of the active α-satellite HOR array. a,b) Plot showing the **a**) mean sequence identity or **b**) mean entropy for the active α-satellite HOR array vs. the putative kinetochore site for each centromere in 65 diverse human genomes. All chromosomes have a higher mean sequence identity for the putative kinetochore site compared to the rest of the α-satellite HOR array. Almost all chromosomes, except for chromosomes 8 and 13, have a lower mean entropy for the putative kinetochore site compared to the rest of the α-satellite HOR array.



Extended Data Fig. 5 | Evidence of archaic introgression in a subset of centromeres from chromosome 10. a) Maximum-likelihood phylogenetic tree depicting the p-arm topology for chromosome 10 centromere haplogroups along with estimated divergence times and haplotype frequencies. Asterisks indicate nodes with 100% bootstrap support, and nodes with 90–99% bootstrap support are indicated numerically. Nodes without an asterisk or number have bootstrap support <90%. **b,c**) Distribution of archaic-specific DNA *k*-mers across 20 kbp windows of the **b**) HG02059 (H2), HG03371 (H2) or **c**) NA20509 (H2), HG00732 (H2), or HG00171 (H1) chromosome 10 centromeric regions using archaic short-read data from four genomes: Altai (Neanderthal), Vindija (Neanderthal), Chagryskaya (Neanderthal), and Denisovan. The dashed line indicates the genome-wide *k*-mer count cutoff, 42, which corresponds to the genome-wide archaic introgression fraction[38]. While the two ancient haplotypes shown in Panel b do not have an enrichment of archaic DNA *k*-mers, the haplotypes in Panel c do have an enrichment, suggesting this haplotype may have resulted, at least in part, due to introgression with Neanderthals and Denisovans.

nature portfolio

Corresponding author(s): Glennis A. Logsdon

Last updated by author(s): Jun 12, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Oxford Nanopore Technologies MinKNOW (v25.03.7) and iVision (v4.5).
Data analysis	All custom code used in this project are publicly available at the following GitHub repositories: AssemblyRepairer (v1.0; https://github.com/logsdon-lab/AssemblyRepairer); CenMAP (v0.4.3.1; https://github.com/logsdon-lab/CenMAP); CenStats (v0.0.10; https://github.com/logsdon-lab/CenStats); NucFlag (v0.3.3; https://github.com/logsdon-lab/NucFlag); CenPlot (v0.1.4; https://github.com/logsdon-lab/CenPlot); Snakemake-HumAS-SD (v0.1.0; https://github.com/logsdon-lab/Snakemake-HumAS-SD); DiMeLo-seq-Viz (v0.1.0; https://github.com/logsdon-lab/Snakemake-DiMeLo-seq-Viz); CUT&RUN-Viz (v0.0.1; https://github.com/logsdon-lab/Snakemake-CutNRun); Pan-centromeric mutation rate estimation (v0.1.0; https://github.com/logsdon-lab/Snakemake-MutationRate); and Cencyclopedia (v0.0.1; https://github.com/logsdon-lab/Cencyclopedia). Other publicly available software used in this study include: minimap2 (v2.28), pbmm2 (v1.13.1), BEDTools (v2.31.1), SAMtools (v1.22 and v1.9), RepeatMasker (v4.1.0 and v4.1.6), Dfam (v3.8), Pysam (v0.22.0), MUSCLE (v3.8.31), L1ME-AID (v1.2.0-beta), dna-brnn (v0.1-r65), trf (v4.09), StringDecomposer (no version specified), CDR-Finder (v1.0.4), IGV (v2.19.5), ModDotPlot (v0.8.4), HMM-Flagger (v1.1.0), GAVISUNK (v1.0), Meryl (v1.4.1), Merqury (v1.3), SVbyEye (v0.99.0), R (v4.4.2), Gepard (v2.1.0), modkit (v0.3.2), cutadapt (v5.1), seqkit (v2.10.0), bwa mem (v0.7.19), deepTools (v3.5.6), MAFFT (v.7.526), IQ-TREE (v2.3.6), iTOL (v7.2.2), subseq (no version specified), ARCkmerFinder (v1.1) and Adobe Photoshop (v25.11.0).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All datasets generated and/or used in this study are listed in Supplementary Table 24 with their BioProject, accession # (if available), and/or URL. For convenience, we also list their BioProjects and/or URLs here: 65 original human whole-genome assemblies generated by the HGSVC (PRJEB76276 and PRJEB83624); 65 human whole-genome assemblies with repaired centromeric contigs (PRJNA1309774); PacBio HiFi and ONT data from 65 humans generated by the HGSVC (PRJEB58376, PRJEB75216, PRJEB77558, PRJEB75190, PRJNA698480, PRJEB75739, PRJEB36100, PRJNA988114, PRJNA339722, PRJEB41778, ERP159775); Illumina data from 65 humans generated by the HGSVC (PRJNA200694, ERP114329, ERP120144); CENP-A and IgG DiMeLo-seq data from 6 humans (SRR35424754, SRR35424755, SRR35269316, SRR35269317, SRR35269314, SRR35269315, SRR35144393, SRR35144394, SRR35424752, SRR35424753, SRR35048265, SRR35048266), CENP-A and IgG CUT&RUN ONT data from GM20355 (SRR35048263, SRR35048264); 227 original human whole-genome assemblies generated by the HPRC (PRJNA730822); PacBio HiFi data from 227 humans generated by the HPRC (PRJNA701308, PRJNA731524); ONT data from 227 humans generated by the HPRC (PRJNA701308, PRJNA731524); 28 human whole-genome assemblies, PacBio HiFi data, and ONT data from the CEPH 1463 pedigree (PRJEB86317 or dbGaP accession: phs003793.v1.p1); Element data from the CEPH 1463 pedigree (PRJEB3246); CHM13 whole-genome assembly (PRJNA559484); CHM1 whole-genome assembly (PRJNA975207); Altai Neanderthal (PRJEB1265); Vindija Neanderthal (PRJEB21157); Chagyrskaya Neanderthal (http://ftp.eva.mpg.de/neandertal/Chagyrskaya/); Denisovan (PRJEB3092); and chimpanzee (AG18354) whole-genome assembly (PRJNA916736). In addition, centromere sequences and annotations for the 65 whole-genome assemblies containing repaired centromeres (HGSVC Phase 3) as well as 227 whole-genome assemblies (HPRC Release 2) and 17 CEPH 1463 whole-genome assemblies containing repaired centromeres consented for public release are available from figshare (https://doi.org/10.6084/m9.figshare.29921990). Finally, the centromeres described in this study can be visualized on www.cencyclopedia.com, an interactive and dynamic centromere browser.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We analyzed 320 human samples, including 159 males (46,XY) and 161 females (46,XX).
Reporting on race, ethnicity, or other socially relevant groupings	We report analyses of publicly available human genome assemblies generated by the Human Genome Structural Variation Consortium (HGSVC), Human Pangenome Reference Consortium (HPRC), and the Platinum Pedigree Consortium and their associated population groups, as established and described by the 1000 Genomes Project (https://www.internationalgenome.org/category/population/).
Population characteristics	The individuals included in the study represent 28 human populations, included 26 defined by the 1000 Genomes Project (ACB, ASW, ESN, GWD, LWK, MSL, YRI, CLM, MXL, PEL, PUR, CDX, CHB, CHS, KHV, JPT, CEU, FIN, GRB, IBS, TSI, BEB, GIH, ITU, PJL, and STU) and 2 defined by the International HapMap Consortium and Human Genome Structural Variation Consortium (MKK and ASK).
Recruitment	Individuals included in this study belong to the following population groups: African Caribbean in Barbados (ACB); African Ancestry in Southwest USA (ASW); Esan in Nigeria (ESN); Gambian in Western Division – Mandinka (GWD); Luhya in Webuye, Kenya (LWK); Maasai in Kinyawa, Kenya (MKK); Mende in Sierra Leone (MSL); Yoruba in Ibadan, Nigeria (YRI); Colombian in Medellin, Colombia (CLM); Mexican Ancestry in Los Angeles, California, USA (MXL); Peruvian in Lima, Peru (PEL); Puerto Rican in Puerto Rico (PUR); Chinese Dai in Xishuangbanna (CDX); Han Chinese in Beijing, China (CHB); Han Chinese South, China (CHS); Kinh in Ho Chi Minh City, Vietnam (KHV); Japanese in Tokyo, Japan (JPT); Ashkenazi (ASK); Utah residents (CEPH) with Northern and Western European ancestry (CEU); Finnish in Finland (FIN); British from England and Scotland (GBR); Iberian Populations in Spain (IBS); Toscani in Italia (TSI); Bengali in Bangladesh (BEB); Gujarati Indians in Houston, Texas, USA (GIH); Indian Telugu in the UK (ITU); Punjabi in Lahore, Pakistan (PJL); and Sri Lankan Tamil in the UK (STU).
Ethics oversight	The HPRC and HGSVC lymphoblastoid cell lines are available from the Coriell Institute for Medical Research for research purposes and are covered by the appropriate ethics approvals by the Coriell Institute. The Platinum Pedigree cells lines are approved and overseen by the Institutional Review Board of the University of Utah under IRB_00065564.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We analyzed 65 human whole-genome assemblies generated by the HGSVC, 227 human whole-genome assemblies generated by the HPRC, 28 human whole-genome assemblies belonging to the Platinum Pedigree, and 1 chimpanzee whole-genome assembly generated by the Telomere-to-Telomere (T2T) Consortium. Sample size was defined based on the availability of consented participants in the respective research studies..
Data exclusions	No data were excluded.
Replication	Computational experiments are deterministic and are, therefore, reproducible. Each wet-lab experiment was performed at least two independent times successfully.
Randomization	N/A: this was not an interventional trial.
Blinding	N/A: this was not an interventional trial.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Mouse monoclonal anti-CENP-A (3-19) antibody (ADI-KAM-CC006, Enzo; 1:50 dilution) Rabbit polyclonal anti-IgG antibody (ab171870, Abcam; 1:50 dilution) Rabbit polyclonal anti-CENP-C antibody (ABE1957, Millipore; 1:100 dilution) Goat anti-rabbit Alexa Fluor Plus 647 antibody (A32733TR, ThermoFisher Scientific; 1:500 dilution)
Validation	The anti-CENP-A antibody was generated against a synthetic peptide consisting of amino acids 3-19 of CENP-A, and mutation of this epitope in human cells prevents antibody binding (Logsdon et al., JCB, 2015). The anti-IgG antibody lacks specificity for known cellular antigens and has been validated by the manufacturer for ChIP applications via Western blot. The anti-CENP-C antibody was tested in both wild-type and CENP-C knockdown HeLa cells and showed time-dependent loss following CENP-C knockdown (Falk et al., Science, 2015).

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	Human cell lines HG00512, HG00513, HG00514, HG02769, HG03065, and GM20355 were purchased from the Coriell Institute for Medical Research (Camden, NJ).
Authentication	Human cell lines HG00512, HG00513, HG00514, HG02769, HG03065, and GM20355 are part of the NHGRI Sample Repository for Human Genetic Research at the Coriell Institute for Medical Research (Camden, NJ) and was authenticated using a multiplex PCR assay with six autosomal microsatellite markers.
Mycoplasma contamination	Human cell lines HG00512, HG00513, HG00514, HG02769, HG03065, and GM20355 are negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used in the study.

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.