Nature Biotechnology: Letter

Linear Assembly of a Human Y Centromere using Nanopore Long Reads

Miten Jain[1,§], Hugh E. Olsen[1,§], Daniel J. Turner[2], David Stoddart[2], Kira V. Bulazel[3], Benedict Paten[1], David Haussler[1], Huntington F. Willard[3], Mark Akeson[1], and Karen H. Miga[1,3*]

§ These authors contributed equally to this work.
* Author for correspondence khmiga@soe.ucsc.edu

Affiliations:

1. UC Santa Cruz Genomics Institute, University of California, Santa Cruz, CA, USA
2. Oxford Nanopore Technologies, Oxford, UK
3. Duke Institute for Genome Sciences and Policy, Duke University, Durham, North Carolina, USA.

The human genome reference sequence remains incomplete due to the challenge of assembling long tracts of near-identical tandem repeats, or satellite DNAs, that are highly enriched in centromeric regions [1]. Efforts to resolve these regions capitalize on a small number of sparsely arranged sequence variants that offer unique markers to break the repeat monotony and ensure proper overlap-layout-consensus assembly DNAs [2–4]. Identifying and spanning sequence variants that may be spaced hundreds of kilobases away within a given array requires long and highly accurate sequence reads. Achieving this requires an advancement in standard single-molecule sequencing, which to date has been error-prone and offers a low throughput of sufficiently long-reads (100 kb+)[5,6]. Here we present a strategy that generates long-reads capable of spanning the complete sequence insert of bacterial artificial chromosomes (BACs) that are hundreds of kilobases in length (~100-300kb). We demonstrate that these reads are sufficient to resolve the linear ordering of repeats within a single satellite array on the Y chromosome, allowing the first complete sequence characterization of a human centromere.

Centromeres are specialized loci that facilitate spindle attachment and ensure proper chromosome segregation during cell division. Normal human centromeric regions are defined by

the enrichment of a divergent, AT-rich ~171-bp tandem repeat, known as alpha satellite DNA [6,7]. The majority of alpha satellite DNAs are organized into higher order repeats (HORs), where a chromosome-specific subset of alpha satellite repeat units, or monomers, is reiterated as a single repeat structure hundreds or thousands of times with high (>99%) sequence conservation to form extensive arrays [8,9]. The sequence composition of individual HOR structures and the extent of repeat variation within the context of each chromosome-assigned HOR array are important to establish kinetochore assembly and ensure centromere identity [1,10,11]. Yet, despite the functional significance of the genomic organization and structure, sequences within human centromeric regions remain absent from even the most complete chromosome assemblies [10,12,13]. To date, no sequencing technology or collection of sequencing technologies has been capable of assembling through centromeric regions due to the requirements for extremely high-quality, long reads to confidently traverse informative, low-copy sequence variants within a given array [3,14]. To this end, we have implemented a nanopore long-read sequencing strategy to generate high-quality reads capable of spanning hundreds of kilobases of highly repetitive DNAs. We have focused on the haploid satellite array that spans the Y centromere (DYZ3) as it is particularly suitable for assembly due to its tractable array size, well-characterized HOR structure and previous physical mapping data [15–19].

We employed a transposase-based method ('1D Longboard Strategy') to generate high-read coverage of full-length BAC DNA with nanopore sequencing (MinION sequencing device, Oxford Nanopore Technologies). This method is designed to linearize the BAC with a single cut-site, followed by addition of the necessary sequencing adaptors (as described in Fig. 1a and Online Methods). The BAC DNA is then read in its entirety through the pore, resulting in complete, end-to-end read coverage of the BAC insert sequence. Plots of read length versus megabase yield revealed enrichment for full length BAC DNA sequences (Fig. 1b and Supplementary Fig. 1). In total, we generated over >3500 full-length 1D reads that span the entirety of 10 BACs (one control BAC from Xq24 [4,5] and nine BACs that mapped to the DYZ3 locus [20]) with MinION sequencing (Supplementary Table 1).

BAC-based assembly across the DYZ3 locus requires overlap among a few informative sequence variants, thus placing great importance on the accuracy of base-calls. However, individual reads (MinION R9.4 chemistry, 1D reads) provide inadequate sequence identity to ensure proper assembly [4,5]. In our experiments, we observed a median alignment identity of

84.8% for individual reads obtained from a control BAC (Xq24; RP11-482A22). Further, we determined insertions, deletions, and mismatches rates of 3.6%, 4.6%, and 3.4% respectively, which are consistent with previous genome-wide estimates [5]. To improve overall base quality, we derived a consensus from multiple alignments of 1D reads that span the full insert length for each BAC (Online Methods). We found that we were able to improve the consensus quality with modest coverage increase and sampling (multiple alignments from 60 randomly sampled full-length reads, with 10 iterations) (Fig. 1c). Additional polishing steps were performed using re-alignment of all full-length nanopore reads for each BAC to improve consensus sequence base quality (99.2% observed for control BAC, RP11-482A22; and an observed range of 99.4 - 99.8% for vector sequences in DYZ3-containing BACs; Online Methods).

To validate satellite sequence variants and to evaluate inherent nanopore sequence biases, we performed Illumina high-coverage BAC resequencing (with coverage median range: 419-1984). We compared counts of 5-mers between corresponding Illumina and nanopore sequence libraries derived from each BAC. Although the 5-mer frequency profiles between the two datasets were largely concordant (Supplemental Figure 2a,b), we found that poly(dA) and poly(dT) homopolymers were overrepresented in our initial nanopore read datasets, a finding that is consistent with genome-wide observations [5]. These poly(dA) and poly(dT) over-representations were reduced in our quality corrected consensus sequences especially for 6mers and 7mers (Supplemental Figure 2c,d). In addition to detecting sequence biases, we used Illumina to validate single-copy sequence variants. In doing so, we used Illumina sequence coverage the BAC cloning vector in each dataset, a region on the BAC expected to be single-copy, to confidently identify single-copy sites within the DYZ3 array that are useful for overlap methods of assembly (illustrated for RP11-718M18; Supplemental Fig. 3b,c; Online Methods). After eliminating reads with multiple best alignments, Illumina base coverage profiles were used to determine informative HOR sequence variants. Further, using a k-mer strategy (where k=21 bp) that identified exact matches between the Illumina and each BAC consensus sequence, we observed an average positive prediction value of 95.8. This allowed us to identify and mask all sites not supported by Illumina reads as false positive variants. Finally, standard quality polishing with pilon [21] was applied strictly to unique (that is, non-satellite DNA) sequences on the proximal p and q arms to improve final quality. Alignment of polished consensus sequences from our control BAC from Xq24 (RP11-482A22) and non-satellite DNA in the p-arm adjacent to the centromere (Yp11.2, RP11-531P03), revealed base-quality

improvement to >99% identity. Using this strategy, we generated nine BAC sequences with high-quality, illumina sequence validated variants and long-range repeat structure, (e.g. 217 kb for RP11-718M18, Fig. 1d) to guide the ordered assembly of BACs from p-arm to q-arm, spanning an entire Y centromere.

We ordered the DYZ3-containing BACs using 38 single-copy, Illumina-validated variants, resulting in 365 kb of assembled alpha satellite DNA (Fig 2, with variant overlap between BACs spanning the p-arm and q-arm shown in Supplemental Fig. 4). The majority of the centromeric locus is defined by a 301 kb array that is comprised entirely of the DYZ3 higher-order repeat (HOR), with a 5.8 kb consensus sequence, repeated in an uninterrupted head-to-tail orientation without repeat inversions or transposable element interruptions [15,19,22]. The assembled length of the RP11 DYZ3 array is consistent with estimates for 96 individuals from the same Y haplogroup (R1b) (Supplemental Fig. 5; mean: 315 kb; median: 350 kb) [23,24]. This finding is in general agreement with pulse-field gel electrophoresis (PFGE) DYZ3 size estimates presented in previous physical mapping of the Y centromeric region [15,16,18]. Using a Y-haplogroup matched cell line [25], we demonstrate concordant PFGE array size estimates across six restriction digests with our RP11 Y centromere length measurement (Supplemental Fig 6).

Pairwise comparisons among the 52 HOR repeats in the assembled DYZ3 array reveal limited sequence divergence between copies (mean 99.7% pairwise identity), as expected for highly homogenized HORs [9,14,17]. Further, in agreement with previous assessment of sequence variation within the DYZ3 array [15,19], we detected instances of a 6.0 kb HOR structural variant and provide evidence for nine copies within the RP11 DYZ3 array that are, in all but one instance, found in tandem [15]. The variant 6.0 kb DYZ3 units are present in two clusters that are separated by 110 kb, as roughly predicted by previous restriction map estimates [17]. Sequence characterization of the DYZ3 array revealed nine HOR haplotypes, defined by linkage between variant bases that are frequent in the array (Supplemental Table 2; Supplemental Fig. 7). These HOR-haplotypes are organized into three local blocks that are enriched for distinct haplotype groups, consistent with previous demonstrations of short-range homogenization of satellite DNA sequence variants [14,15,26].

Directly adjacent to the 301 kb HOR array, we identified a brief transition zones on both p-arm (6.1 kb) and q-arm (4.9 kb) where we observed interspersed monomers with high sequence identity (~98-100%) to the canonical DYZ3 HOR unit, yet do not have the same

multi-monomeric repeat structure (Supplemental Figure 8). Distal from the HOR array, sequence identity with DYZ3 markedly decreases (80-85%), and we observe a shift in monomer orientation before the satellite junction with the p-arm (Supplemental Figure 8b). We observe accumulation of transposable elements in the divergent satellite at the ends of the HOR array, in support of the accretion model for HOR array turnover [1,27]. The DYZ3 HOR sequence and chromosomal location of the active centromere on the human chromosome Y is not shared among closely related great apes (Supplemental Figure 9a) [28,29]. However, previous evolutionary dating of specific transposable element subfamilies (notably, L1PA3 9.2–15.8 MYA [30]) within the divergent satellite DNAs, as well as shared synteny of 11.9 kb of alpha satellite DNA in the chimpanzee genome Yq assembly indicate that the locus was present in the last common ancestor with chimpanzee (Supplemental Figure 9b)[28,31].

Functional centromeres are defined by the presence of inner centromere proteins that epigenetically mark the site of kinetochore assembly [32–34]. To define the genomic position of the functional centromere on the Y chromosome we studied the enrichment profiles of inner kinetochore centromere protein A (CENP-A), a histone H3 variant that replaces histone H3 in centromeric nucleosomes, using a Y-haplogroup matched cell line that offers a similar DYZ3 array sequence (Figure 2, Supplemental Figure 10 a) [10,24,25,34,35]. CENP-A enrichment is predominantly restricted to the canonical DYZ3 HOR array, although we do identify reduced centromere protein enrichment extending up to 20 kb into flanking divergent alpha satellite on both the p-arm and q-arm side (Supplemental Figure 10 b-d). Thus, we have provided a complete genomic definition of a human centromere, critical to advance sequence-based studies of centromere identity and function.

In conclusion, we have implemented a long-read strategy to advance sequence characterization of tandemly repeated satellite DNAs. Despite their repetitive content, our analysis provides the necessary directed genomic approach to map, sequence and assemble centromere regions. In doing so, we report the array repeat organization and structure of a human centromere on chromosome Y. Complete, haploid resolved linear assembly of centromeric regions, as shown in our analysis, are expected to have evolutionary and functional implications. We expect that this work will be applicable to ongoing efforts to complete of the human genome.

Acknowledgements

This work was supported by grants to M.A from NHGRI [HG007827] and D.H., B.P. [DT06172015] from the Keck Foundation.

Contributions

K.M and H.W. conceived the project. K.M., M.J., D.T., D.S., H.O. and M.A. designed the experiments; M.J. and H.O. were involved with BAC sample preparation; M.J. and H.O. performed MinION sequencing and base-calling; M.J and K.M analyzed the BAC sequencing data and validation analyses; K.M. performed the pulse-field gel electrophoresis array length estimates; K.B. contributed FISH analysis; K.M., M.J. and H.O. contributed to analysis and figure generation; M.A., D.T., D.S., H.W., B.P., and D.H. provided technical advice; all authors contributed to the writing, editing and completion of the manuscript.

Competing financial interests

M.A. is a consultant to Oxford Nanopore Technologies. D.T. and D.S. are employed by Oxford Nanopore Technologies.

Corresponding authors

Correspondence to: Karen H. Miga (khmiga@soe.ucsc.edu)

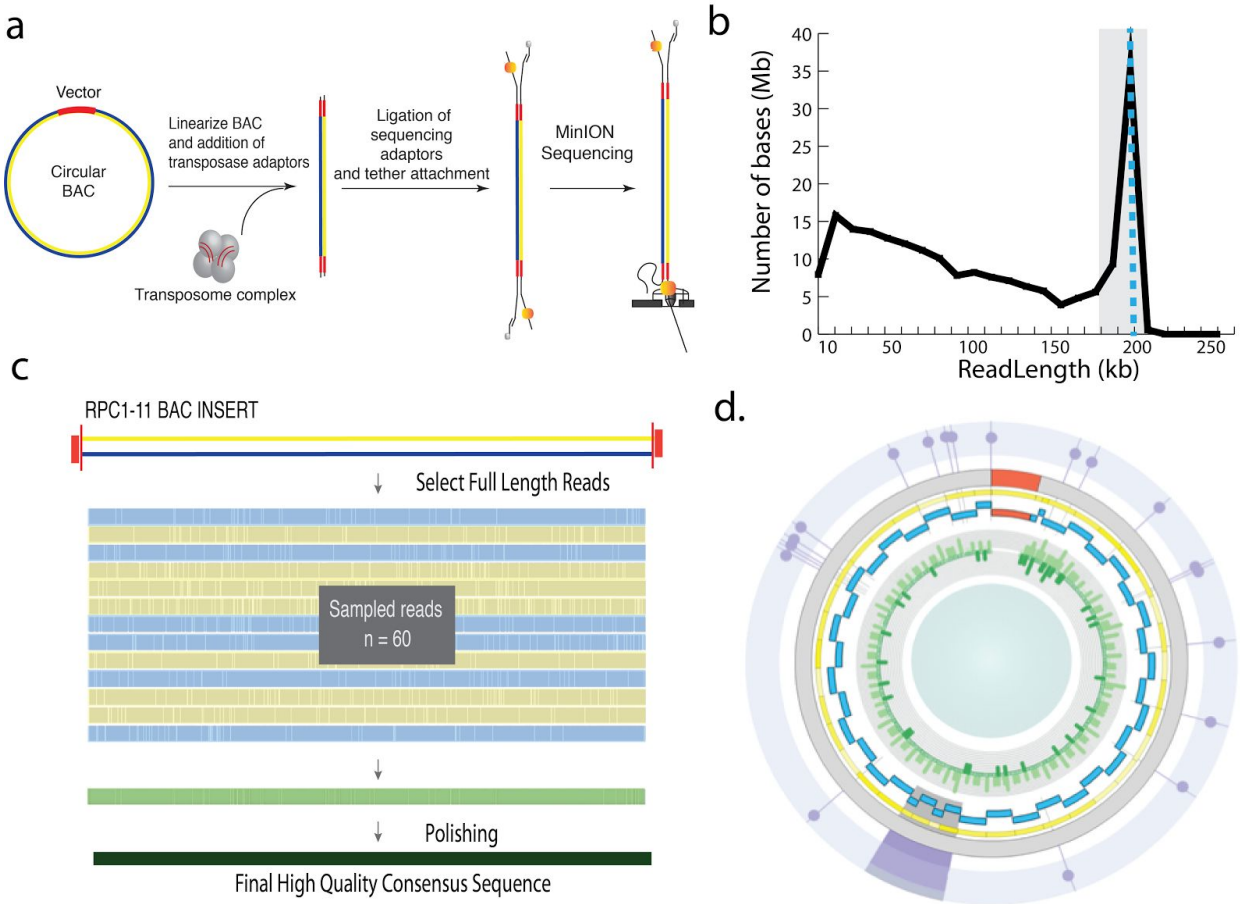

Figure 1: BAC-based 1D Longboard nanopore sequencing strategy on the MinION. (a) Optimized strategy to cut each circular BAC once with transposase, resulting in a linear and complete DNA fragment of the BAC. After ligation of sequencing adaptors we perform MinION sequencing. (b) Yield plots of BAC DNA (RP11-648J18) provide enrichment, or peaks, supporting BAC lengths. Shading demonstrates the selection of a narrow range of read lengths used in deriving the consensus, the blue dotted line reveals the median value within the selected region providing the closest estimate of insert size. (c) To generate the high quality consensus sequence for each BAC we performed multiple alignment of 60 full length 1D reads (shown as blue and yellow for both orientations) sampled at random with 10 iterations, followed by polishing steps (green) with the entire nanopore long read data and Illumina data. (d) A Circos representation of the polished RP11-718M18 BAC consensus sequence (insert shown in grey: 217 kb, vector in red: 8.8 kb). Blue boxes indicate the position of each DYZ3 HOR found in a head-to-tail orientation. Purple shading indicates Illumina-validated low copy variants and the site of the DYZ3 repeat structural variants (6 kb) in tandem. Illumina read mapping to

support the variants are shown as the green histogram, with exact match 21-mers (at least 2 overlapping) below in dark green.

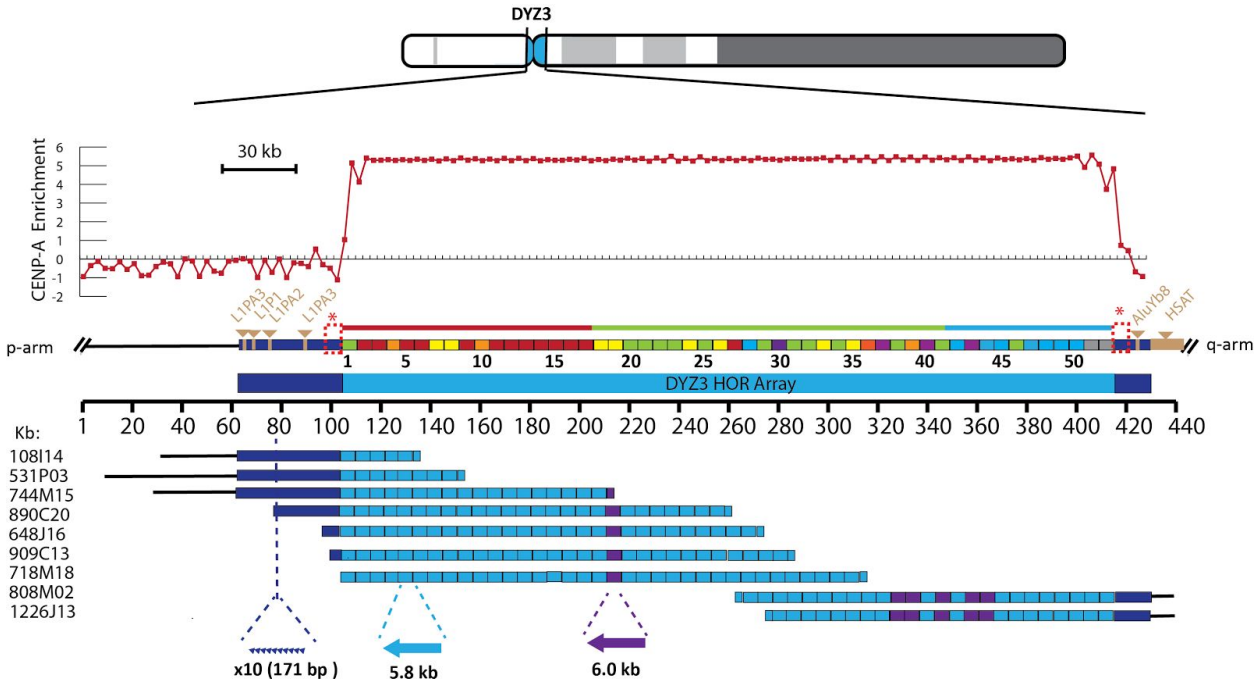

Figure 2: Linear assembly of the RP11 Y centromere. Ordering of nine DYZ3-containing BACs spanning from proximal p-arm to proximal q-arm provided evidence for a 354,250 bp region enriched in alpha satellite DNA. Highly divergent monomeric alpha satellite is indicated in dark blue, with brown arrows indicating sites of transposable element insertion. Transition regions are boxed in red and marked with an asterisk. The majority of the centromeric locus is defined by the DYZ3 conical 5.8 kb higher-order repeat (HOR) (light blue), that is observed in a head to tail orientation from p-arm to q-arm, for a total of 301 kb. Nine HOR variants (6.0 kb, shown in purple) have been identified, with all but one identified in tandem. DYZ3 HORs were classified into nine haplotypes using four frequent satellite DNA variants in the array (haplotype (H)1 red, H2 orange, H3 yellow, H4 green, H5 blue, H6 dark orange; H7 purple, H8 dark purple, H9 grey). We identified three predominant blocks: H1 proximal to the p-arm (red), H4 in the middle of the array (green), and H5 adjacent to the q-arm (blue). The genomic location of the functional Y centromere is defined by the enrichment of centromere protein A (CENP-A), where enrichment (~5-6x) is attributed predominantly to the DYZ3 HOR array.

References

1. Schueler, M. G., Higgins, A. W., Rudd, M. K., Gustashaw, K. & Willard, H. F. Genomic and genetic definition of a functional human centromere. *Science* **294,** 109–115 (2001).
2. Khost, D. E., Eickbush, D. G. & Larracuente, A. M. Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in Drosophila melanogaster. *Genome Res.* **27,** 709–721 (2017).
3. Miga, K. H. Completing the human genome: the progress and challenge of satellite DNA assembly. *Chromosome Res.* **23,** 421–426 (2015).
4. Jain, M. *et al.* Improved data analysis for the MinION nanopore sequencer. *Nat. Methods* **12,** 351–356 (2015).
5. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *bioRxiv* 128835 (2017). doi:10.1101/128835
6. Manuelidis, L. Chromosomal localization of complex and simple repeated human DNAs. *Chromosoma* **66,** 23–32 (1978).
7. Mitchell, A. R., Gosden, J. R. & Miller, D. A. A cloned sequence, p82H, of the alphoid repeated DNA family found at the centromeres of all human chromosomes. *Chromosoma* **92,** 369–377 (1985).
8. Willard, H. F. Chromosome-specific organization of human alpha satellite DNA. *Am. J. Hum. Genet.* **37,** 524–532 (1985).
9. Willard, H. F. & Waye, J. S. Chromosome-specific subsets of human alpha satellite DNA: analysis of sequence divergence within and between chromosomal subsets and evidence for an ancestral pentameric repeat. *J. Mol. Evol.* **25,** 207–214 (1987).
10. Hayden, K. E. *et al.* Sequences associated with centromere competency in the human genome. *Mol. Cell. Biol.* **33,** 763–772 (2013).
11. Maloney, K. A. *et al.* Functional epialleles at an endogenous human centromere. *Proc. Natl.*

Acad. Sci. U. S. A. **109,** 13704–13709 (2012).

12. Rudd, M. K. & Willard, H. F. Analysis of the centromeric regions of the human genome assembly. *Trends Genet.* **20,** 529–533 (2004).
13. 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).
14. Durfy, S. J. & Willard, H. F. Patterns of intra- and interarray sequence variation in alpha satellite from the human X chromosome: evidence for short-range homogenization of tandemly repeated DNA sequences. *Genomics* **5,** 810–821 (1989).
15. Tyler-Smith, C. & Brown, W. R. Structure of the major block of alphoid satellite DNA on the human Y chromosome. *J. Mol. Biol.* **195,** 457–470 (1987).
16. Oakey, R. & Tyler-Smith, C. Y chromosome DNA haplotyping suggests that most European and Asian men are descended from one of two males. *Genomics* **7,** 325–330 (1990).
17. Tyler-Smith, C. Structure of repeated sequences in the centromeric region of the human Y chromosome. *Development* **101 Suppl,** 93–100 (1987).
18. Wevrick, R. & Willard, H. F. Long-range organization of tandem arrays of alpha satellite DNA at the centromeres of human chromosomes: high-frequency array-length polymorphism and meiotic stability. *Proceedings of the National Academy of Sciences* **86,** 9394–9398 (1989).
19. Wolfe, J. *et al.* Isolation and characterization of an alphoid centromeric repeat family from the human Y chromosome. *J. Mol. Biol.* **182,** 477–485 (1985).
20. Tilford, C. A. *et al.* A physical map of the human Y chromosome. *Nature* **409,** 943–945 (2001).
21. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9,** e112963 (2014).

22. Cooper, K. F., Fisher, R. B. & Tyler-Smith, C. The major centromeric array of alphoid satellite DNA on the human Y chromosome is non-palindromic. *Hum. Mol. Genet.* **2,** 1267–1270 (1993).
23. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
24. Miga, K. H. *et al.* Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res.* **24,** 697–707 (2014).
25. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5,** e254 (2007).
26. Warburton, P. E. & Willard, H. F. Interhomologue sequence variation of alpha satellite DNA from human chromosome 17: evidence for concerted evolution along haplotypic lineages. *J. Mol. Evol.* **41,** 1006–1015 (1995).
27. McAllister, B. F. & Werren, J. H. Evolution of tandemly repeated sequences: What happens at the end of an array? *J. Mol. Evol.* **48,** 469–481 (1999).
28. Hughes, J. F. *et al.* Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463,** 536–539 (2010).
29. Archidiacono, N. *et al.* Evolution of chromosome Y in primates. *Chromosoma* **107,** 241–246 (1998).
30. Khan, H., Smit, A. & Boissinot, S. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res.* **16,** 78–87 (2006).
31. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
32. Karpen, G. H. & Allshire, R. C. The case for epigenetic effects on centromere identity and function. *Trends Genet.* **13,** 489–496 (1997).

33. Black, B. E. & Cleveland, D. W. Epigenetic centromere propagation and the nature of CENP-a nucleosomes. *Cell* **144,** 471–479 (2011).

34. 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).

35. Henikoff, J. G., Thakur, J., Kasinathan, S. & Henikoff, S. A unique chromatin complex occupies young α-satellite arrays of human centromeres. *Sci Adv* **1,** (2015).