

The structure, function, and evolution of a complete human chromosome 8

Evan Eichler (✉ eee@gs.washington.edu)

University of Washington School of Medicine <https://orcid.org/0000-0002-8246-4014>

Glennis Logsdon

University of Washington <https://orcid.org/0000-0003-2396-0656>

Mitchell Vollger

University of Washington <https://orcid.org/0000-0002-8651-1615>

PingHsun Hsieh

University of Washington

Yafei Mao

University of Washington

Mikhail Liskovych

National Cancer Institute

Sergey Koren

Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute <https://orcid.org/0000-0002-1472-8962>

Sergey Nurk

National Human Genome Research Institute

Ludovica Mercuri

University of Bari

Philip Dishuck

University of Washington School of Medicine

Arang Rhie

National Institutes of Health

Leonardo de Lima

Stowers Institute for Medical Research

Tatiana Dvorkina

Saint Petersburg State University

David Porubsky

<https://orcid.org/0000-0001-8414-8966>

Alla Mikheenko

Saint Petersburg State University

Andrey Bzikadze

Graduate Program in Bioinformatics and Systems Biology, University of California, San Diego, CA, USA

Milinn Kremitzki

McDonnell Genome Institute at Washington University, St. Louis Mo, USA

Tina Lindsay

McDonnell Genome Institute at Washington University, St. Louis Mo, USA

Chirag Jain

National Human Genome Research Institute

Kendra Hoekzema

University of Washington

Shwetha Murali

University of Washington

Katherine Munson

University of Washington <https://orcid.org/0000-0001-8413-6498>

Carl Baker

University of Washington

Melanie Sorensen

Department of Genome Sciences, University of Washington School of Medicine

Alexandra Lewis

University of Washington School of Medicine

Urvashi Surti

University of Pittsburgh

Jennifer Gerton

Stowers Institute for Medical Research <https://orcid.org/0000-0003-0743-3637>

Vladimir Larionov

National Cancer Institute

Mario Ventura

University of Bari <https://orcid.org/0000-0001-7762-8777>

Karen Miga

UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, United States.

<https://orcid.org/0000-0002-3670-4507>

Adam Phillippy

National Human Genome Research Institute <https://orcid.org/0000-0003-2983-8934>

Biological Sciences - Article

Keywords: Long-read Sequencing, Linear Assembly, Centromeric Satellite DNA

Posted Date: September 22nd, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-72559/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Nature on April 7th, 2021. See the published version at <https://doi.org/10.1038/s41586-021-03420-7>.

1 The structure, function, and evolution of a complete human chromosome 8

2
3
4
5
6
7
8
9

Glennis A. Logsdon¹, Mitchell R. Vollger¹, PingHsun Hsieh¹, Yafei Mao¹, Mikhail A. Liskovykh², Sergey Koren³, Sergey Nurk³, Ludovica Mercuri⁴, Philip C. Dishuck¹, Arang Rhie³, Leonardo G. de Lima⁵, Tatiana Dvorkina⁶, David Porubsky¹, Alla Mikheenko⁶, Andrey V. Bzikadze⁷, Milinn Kremitzki⁸, Tina A. Graves-Lindsay⁸, Chirag Jain³, Kendra Hoekzema¹, Shwetha C. Murali^{1,9}, Katherine M. Munson¹, Carl Baker¹, Melanie Sorensen¹, Alexandra M. Lewis¹, Urvashi Surti¹⁰, Jennifer L. Gerton⁵, Vladimir Larionov², Mario Ventura⁴, Karen H. Miga¹¹, Adam M. Phillippy³, Evan E. Eichler^{1,9}

- 10 1. Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA
11 98195, USA
12 2. Developmental Therapeutics Branch, National Cancer Institute, Bethesda, MD 20892, USA
13 3. Genome Informatics Section, Computational and Statistical Genomics Branch, National Human
14 Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA
15 4. Department of Biology, University of Bari, Aldo Moro, Bari 70121, Italy
16 5. Stowers Institute for Medical Research, Kansas City, MO 64110, USA
17 6. Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, Saint Petersburg
18 State University, Saint Petersburg 199034, Russia
19 7. Graduate Program in Bioinformatics and Systems Biology, University of California, San Diego,
20 CA 92093, USA
21 8. McDonnell Genome Institute, Department of Genetics, Washington University School of
22 Medicine, St. Louis, MO 63108, USA
23 9. Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA
24 10. Department of Pathology, University of Pittsburgh, Pittsburgh, PA 15213, USA
25 11. Center for Biomolecular Science and Engineering, University of California, Santa Cruz, Santa
26 Cruz, CA 95064, USA
27

28 Correspondence to:

29 Evan E. Eichler, Ph.D.
30 Department of Genome Sciences
31 University of Washington School of Medicine
32 3720 15th Ave NE, S413A
33 Seattle, WA 98195-5065
34 Phone: 1-206-543-9526
35 E-mail: eee@gs.washington.edu

36 **ABSTRACT**

37 The complete assembly of each human chromosome is essential for understanding human biology and
38 evolution. Using complementary long-read sequencing technologies, we complete the first linear
39 assembly of a human autosome, chromosome 8. Our assembly resolves the sequence of five
40 previously long-standing gaps, including a 2.08 Mbp centromeric α -satellite array, a 644 kbp defensin
41 copy number polymorphism important for disease risk, and an 863 kbp variable number tandem repeat
42 at chromosome 8q21.2 that can function as a neocentromere. We show that the centromeric α -satellite
43 array is generally methylated except for a 73 kbp hypomethylated region of diverse higher-order α -
44 satellite enriched with CENP-A nucleosomes, consistent with the location of the kinetochore. Using a
45 dual long-read sequencing approach, we complete the assembly of the orthologous chromosome 8
46 centromeric regions in chimpanzee, orangutan, and macaque for the first time to reconstruct its
47 evolutionary history. Comparative and phylogenetic analyses show that the higher-order α -satellite
48 structure evolved specifically in the great ape ancestor, and the centromeric region evolved with a
49 layered symmetry, with more ancient higher-order repeats located at the periphery adjacent to
50 monomeric α -satellites. We estimate that the mutation rate of centromeric satellite DNA is accelerated
51 at least 2.2-fold, and this acceleration extends beyond the higher-order α -satellite into the flanking
52 sequence.

53

54 **INTRODUCTION**

55 Since the announcement of the sequencing of the human genome 20 years ago^{1,2}, human
56 chromosomes have remained unfinished due to large regions of highly identical repeats located within
57 centromeres, segmental duplication, and the acrocentric short arms of chromosomes. The presence of
58 large swaths (>100 kbp) of highly identical repeats that are themselves copy number polymorphic has
59 meant that such regions have persisted as gaps, limiting our understanding of human genetic variation
60 and evolution^{3,4}. In the case of centromeres, for example, the AT-rich, 171 bp repeat, known as α -
61 satellite, is organized in tandem to form hundreds to thousands of higher-order repeats (HORs) that
62 span mega-base pairs of human DNA and are variable in copy number between homologous
63 chromosomes⁵⁻⁸. Such repetitive structures have complicated cloning and assembly of these and other
64 regions of the human genome and, as a result, the sequences have either remained as gaps or are
65 presented as decoys of predicted sequence to improve mapping against the human reference^{9,10}.

66

67 The advent of long-read sequencing technologies and associated algorithms have now made it
68 possible to systematically assemble these regions from native DNA for the first time¹¹⁻¹³. In addition, the
69 use of DNA from complete hydatidiform moles (CHMs) to serve as reference genomes has greatly
70 simplified sequence resolution of these complex regions. Most CHMs carry only the paternal
71 complement of human chromosomes due to an aberrant fertilization event in which a single sperm
72 duplicates to give rise to two identical haploid sets of chromosomes. As a result, there is no allelic
73 variation, permitting the assembly of a single haplotype without interference from a second haplotype¹⁴.
74 The use of long reads from CHM DNA created the first comprehensive map of human structural
75 variation¹⁵ and the first report of a completely sequenced human X chromosome, where the centromere
76 was fully resolved¹⁶.

77

78 Here, we present the first complete linear assembly of a human autosomal chromosome not only to
79 permit the study of human biology and evolution but to serve as a benchmark for the completion of
80 other chromosomes and future diploid genomes. We chose human chromosome 8 because it carries a
81 modestly sized centromere (approximately 1.5-2.2 Mbp)^{8,17}, where the α -satellite repeats are organized

82 into a well-defined HOR array. The chromosome, however, also contains one of the most structurally
83 dynamic regions in the human genome—the β -defensin gene cluster located at 8p23.1^{18–20}—as well as
84 a neocentromere located at 8q21.2, which have been largely unresolved for the last 20 years. We use
85 the finished chromosome 8 sequence to perform the first comparative sequence analyses of complete
86 centromeres across the great ape phylogeny and show how this information enables new insights into
87 the structure, function, and evolution of our genome.

88

89 RESULTS

90 **Telomere-to-telomere assembly of chromosome 8.** To resolve the gaps in human chromosome 8
91 (**Fig. 1a**), we developed a targeted assembly method that leverages the complementary strengths of
92 Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) long-read sequencing (**Fig.**
93 **1b; Methods**). We reasoned that ultra-long (>100 kbp) ONT reads harbor sufficient sequence variation
94 to permit the assembly of complex regions, generating an initial sequence scaffold that could be
95 replaced with highly accurate PacBio high-fidelity (HiFi) contigs to improve the overall base accuracy.
96 To this end, we generated 20-fold sequence coverage of ultra-long ONT data and 32.4-fold coverage of
97 PacBio HiFi data from a CHM (CHM13hTERT; abbr. CHM13; **Extended Data Fig. 1; Methods**). Over
98 half of the ultra-long ONT data is composed of reads exceeding 139.8 kbp in length, with the longest
99 mapped read 1.538 Mbp long (**Extended Data Fig. 1a**). More than half of the PacBio HiFi data is
100 contained in reads greater than 17.8 kbp, with a median accuracy exceeding 99.9% (**Extended Data**
101 **Fig. 1b**). We assembled complex regions in chromosome 8 by first creating a library of singly unique
102 nucleotide k -mers (SUNKs)²¹, or sequences of length k that occur approximately once per haploid
103 genome (here, $k = 20$), from CHM13 PacBio HiFi data (**Methods**). These SUNKs were validated with
104 Illumina data generated from the same genome and used to barcode ultra-long ONT reads (**Fig. 1b;**
105 **Methods**). Ultra-long ONT reads sharing highly similar barcodes were assembled into an initial
106 sequence scaffold that traverses each gap and complex genomic region within chromosome 8 (**Fig. 1b;**
107 **Methods**). We improved the base-pair accuracy of the sequence scaffolds by replacing the raw ONT
108 sequence with several concordant PacBio HiFi contigs and integrating them into a linear assembly of
109 human chromosome 8 from Nurk and colleagues¹¹ (**Fig. 1b; Methods**).

110

111 The complete telomere-to-telomere sequence of human chromosome 8 is 146,259,671 bases long and
112 encompasses 3,334,256 additional bases missing from the current reference genome (GRCh38). Most
113 of the additions reside within distinct chromosomal regions: a ~644 kbp copy number polymorphic β -
114 defensin gene cluster mapping to chromosome 8p23.1 (**Fig. 1c**); the complete centromere
115 corresponding to 2.08 Mbp of α -satellite DNA (**Fig. 2a**); a 863 kbp 8q21.2 variable number tandem
116 repeat (VNTR) (**Fig. 3a**); and both telomeric regions ending with the canonical TTAGGG repeat
117 sequence (**Extended Data Fig. 2**). We validated the organization and accuracy of the chromosome 8
118 assembly via a suite of orthogonal technologies, including optical mapping (Bionano Genomics),
119 Strand-seq^{22,23}, and comparisons to finished BAC sequence as well as whole-genome sequence
120 Illumina data derived from the same source genome (**Methods**). Our analyses show that the CHM13
121 chromosome 8 assembly is free of assembly errors, false joins, and misorientations (**Extended Data**
122 **Fig. 3**). We estimate the overall base accuracy to be between 99.9915% and 99.9999% (quality value
123 (QV) score between 40.70 and 63.19, as determined from sequenced BACs and mapped k -mers²⁴,
124 respectively). An analysis of 24 million human full-length transcripts generated from Iso-Seq data
125 identifies 61 protein-coding and 33 noncoding loci that map better to this finished chromosome 8
126 sequence than to GRCh38, including the discovery of novel genes mapping to copy number
127 polymorphic regions (see below; **Fig. 1d, Extended Data Fig. 4**).

128 Our targeted assembly method is particularly useful for traversing large complex regions of highly
129 identical duplications. A case in point is the β -defensin gene cluster¹⁸, which we resolved into a single
130 7.06 Mbp locus—substantially larger than the 4.56 Mbp region in the current human reference genome
131 (GRCh38), which is flanked by two 50 kbp gaps (**Fig. 1c**). To assemble this locus, we initially barcoded
132 26 ultra-long ONT reads (averaging 378.7 kbp in length) with SUNKs and assembled them to generate
133 a 7.06 Mbp sequence scaffold. PacBio HiFi contigs concordant with the ONT-based scaffold replaced
134 99.9934% of the sequence (7,058,731 out of 7,059,195 bp), increasing the overall base accuracy of the
135 assembly to 99.9911% (QV score of 40.48; estimated with mapped BACs; **Extended Data Fig. 5a**).
136 Our analysis shows that the β -defensin assembly is free from structural errors and misassemblies
137 (**Extended Data Fig. 5a**). Additionally, our analysis reveals a more complex haplotype than GRCh38,
138 consistent with previously published reports of structural variation associated with the chromosome
139 8p23.1 β -defensin gene cluster^{18,20}. We resolve the breakpoints of one of the largest common inversion
140 polymorphisms in the human genome (3.89 Mbp in length) and show that the breakpoints map within
141 large, highly identical duplications that are copy number polymorphic in the human population (**Fig. 1d**).
142 In contrast to the human reference, which carries two such segmental duplications (SDs), there are
143 three SDs in CHM13: a 544 kbp SD on the distal end and two 693 and 644 kbp SDs on the proximal
144 end, respectively (**Fig. 1c**). Each SD cassette carries at least five β -defensin genes and, as a result, we
145 identify five additional β -defensin genes that are virtually identical at the amino acid level to the
146 reference (**Fig. 1c, Extended Data Table 1**). Because ONT data allows methylation signals to be
147 assessed²⁵, we inferred the methylation status of cytosines across the entire β -defensin locus. All three
148 SDs harbor a 151-163 kbp methylated region residing in the LTR-rich region of the duplication, while
149 the remainder of the SD, including the β -defensin gene cluster, is largely unmethylated (**Fig. 1c**),
150 consistent with its transcription. Complete sequence resolution of this alternate haplotype is important
151 because the inverted haplotype preferentially predisposes to recurrent microdeletions associated with
152 developmental delay, microcephaly, and congenital heart defects^{26,27}, and copy number polymorphism
153 of the five β -defensin genes has been associated with immune-related phenotypes such as psoriasis
154 and Crohn's disease^{19,28}.

155
156 **Sequence resolution of the chromosome 8 centromere.** Prior studies have estimated the length of
157 the chromosome 8 centromere to be between 1.5 and 2.2 Mbp, based on analysis of the HOR α -
158 satellite array^{8,17}. While various HORs of different lengths are thought to comprise the centromere, the
159 predominant species has a unit length of 11 monomers, resulting in a tandem repeat of 1881 bp^{8,17}.
160 Using our targeted assembly method, we spanned the chromosome 8 centromere with 11 ultra-long
161 ONT reads (mean length 389.4 kbp), which were replaced with PacBio HiFi contigs based on SUNK
162 barcoding. Unlike the ONT assembly, the HiFi assembly was not completely continuous but was more
163 accurate, allowing it to be anchored uniquely into the ONT sequence scaffold. Our assembled CHM13
164 chromosome 8 centromere consists of a 2.08 Mbp D8Z2 α -satellite HOR array flanked by blocks of
165 monomeric α -satellite on the p- (392 kbp) and q- (588 kbp) arms (**Fig. 2a**). Both monomeric α -satellite
166 blocks are interspersed with LINES, SINEs, and LTRs, with tracts of γ -satellite specific to the q-arm. We
167 validated the sequence, structure, and organization of the chromosome 8 centromere using five
168 orthogonal methods. First, long-read sequence read-depth analysis from two orthogonal native DNA
169 sequencing platforms shows uniform coverage, suggesting that the assembly is free from large
170 structural errors (**Extended Data Fig. 6a**). Fluorescent *in situ* hybridization (FISH) on stretched
171 chromosomes confirms the long-range order and organization of the centromere (**Extended Data Fig.**
172 **6a,b**). Droplet digital PCR shows that there are 1344 +/- 142 D8Z2 HORs within the α -satellite array,
173 consistent with our estimates (**Extended Data Fig. 6c; Methods**). Pulsed-field gel electrophoresis
174 Southern blots on CHM13 DNA digested with two different restriction enzymes recapitulates the

175 banding pattern predicted from the assembly (**Fig. 2a,b**). Finally, applying our assembly approach to
176 ONT and HiFi data available for a diploid human genome generates two additional chromosome 8
177 centromere haplotypes, replicating the overall organization with only subtle differences in overall length
178 of HOR arrays (**Extended Data Fig. 7, Extended Data Table 2**).

179
180 Using the assembled centromere sequence, we investigated its genetic and epigenetic organization.
181 On a genetic level, we found that the chromosome 8 centromeric HOR array is primarily composed of
182 four distinct HOR types represented by 4, 7, 8, or 11 α -satellite monomer cassettes (**Fig. 2a, Extended**
183 **Data Fig. 8**). While the 11-mer predominates (36%), the other HORs are also abundant (19-23%) and
184 are all derivatives of the 11-mer (**Extended Data Fig. 8b,c**). Interestingly, we find that HORs are
185 differentially distributed regionally across the centromere. While most regions are admixed with different
186 HOR types, we also identify regions of homogeneity, such as clusters of 11-mers mapping to the
187 periphery of the HOR array (92 and 158 kbp in length) and a 177 kbp region in the center that is
188 composed solely of 7-mer HORs. To investigate the epigenetic organization, we mapped methylated
189 cytosines along the centromeric region and found that most of the α -satellite HOR array is methylated,
190 except for a small, 73 kbp hypomethylated region (**Fig. 2a**). To determine if this hypomethylated region
191 is the site of the epigenetic centromere (marked by the presence of nucleosomes containing the histone
192 H3 variant, CENP-A), we mapped CENP-A ChIP-seq data from diploid human cells and found that
193 CENP-A was primarily located within a 632 kbp stretch encompassing the hypomethylated region (**Fig.**
194 **2a, Extended Data Fig. 9**). Subsequent chromatin fiber FISH revealed that CENP-A maps to the
195 hypomethylated region within the α -satellite HOR array (**Fig. 2c**). Remarkably, the hypomethylated
196 region shows some of the greatest HOR admixture, suggesting a potential optimization of HOR
197 subtypes associated with the active kinetochore (mean entropy over the 73 kbp region = 1.91;
198 **Extended Data Fig. 8a, Methods**).

199
200 To better understand the long-range organization and evolution of the centromere, we generated a
201 pairwise sequence identity heat map (**Methods**), which compares the sequence identity of 5 kbp
202 fragments along the length of the centromere (**Fig. 2a**). We find that the centromere consists of five
203 major evolutionary layers that show mirror symmetry. The outermost layer resides in the monomeric α -
204 satellite, where sequences are highly divergent from the rest of the centromere but are more similar to
205 each other (Arrow 1). The second layer defines the monomeric-to-HOR transition and is a short (57-60
206 kbp) region. The p and q regions are 87-92% identical with each other but only 78% or less with other
207 centromeric satellites (Arrow 2). The third layer is completely composed of HORs. The p and q regions
208 are 92 and 149 kbp in length, respectively, and share more than 96% sequence identity with each other
209 (Arrow 3) but less than that with the rest of the centromere. This layer is composed largely of
210 homogenous 11-mers and defines the transition from unmethylated to methylated DNA. The fourth
211 layer is the largest and defines the bulk of the HOR α -satellite (1.42 Mbp in total). It shows the greatest
212 admixture of different HOR subtypes and, once again, the p and q blocks share identity with each other
213 but are more divergent from the rest of the layers (Arrow 4). Both blocks are highly methylated with the
214 exception of the 73 kbp hypomethylated region mapping to the q-arm. Finally, the fifth layer
215 encompasses the centermost 416 kbp of the HOR array, a region of near-perfect sequence identity that
216 is divergent from the rest of the centromere (Arrow 5).

217
218 **Sequence resolution of the chromosome 8q21.2 VNTR.** The layered and mirrored nature of the
219 chromosome 8 centromere is reminiscent of another gap region located at chromosome 8q21.2, which
220 we resolved for the first time (**Fig. 3**). This region is a cytogenetically recognizable euchromatic
221 variant²⁹ thought to contain one of the largest VNTRs in the human genome²⁹. The 12.192 kbp

222 repeating unit encodes the *GOR1/REXO1L1* pseudogene and is highly copy number polymorphic
223 among humans^{29,30}. This VNTR is of biological interest because it is the site of a recurrent
224 neocentromere, where a functional centromere devoid of α -satellite has been observed in multiple
225 unrelated individuals^{31,32}. The complete genetic and epigenetic composition of the 8q21.2 VNTR has
226 not yet been resolved because the region largely corresponds to a gap in the human reference genome
227 (GRCh38). Using our approach, we successfully assembled the VNTR into an 863.5 kbp sequence
228 composed of ~71 repeating units (67 complete and 7 partial units) (**Fig. 3a**). A pulsed-field gel Southern
229 blot of digested CHM13 DNA confirms the length and structure of the VNTR (**Fig. 3a,b**). Chromatin
230 fiber FISH estimates that the array is composed of 67 +/- 5.2 repeats, consistent with the assembly
231 (**Extended Data Fig. 10, Methods**). Mapping of long-read data reveals uniform coverage along the
232 entire assembly (**Extended Data Fig. 10a**), indicating a lack of large structural errors. We estimate that
233 the 12.192 kbp repeat unit varies from 53 to 158 copies in the human population, creating tandem
234 repeat arrays ranging from 652 kbp to 1.94 Mbp (**Fig. 3c**). We identify a higher-order structure of the
235 VNTR consisting of five distinct domains that alternate in orientation (**Fig. 3a**), and each domain
236 contains 5 to 23 complete repeat units that are more than 98.5% identical to each other (**Fig. 3a**).
237 Mapping of methylated cytosines to the array shows that each 12.192 kbp repeat is primarily
238 methylated in the 3 kbp region corresponding to *GOR1/REXO1L1*, while the rest of the repeat unit is
239 largely unmethylated (**Fig. 3a**). Mapping of centromeric chromatin from a cell line harboring an 8q21.2
240 neocentromere³² shows that the CENP-A nucleosomes map to the unmethylated region of the repeat
241 unit in the CHM13 assembly (**Fig. 3a**). While this is consistent with the VNTR being the potential site of
242 the functional kinetochore of the neocentromere, sequence and assembly of the neocentromere-
243 containing cell line will be critically important.

244
245 **Centromere evolutionary reconstruction.** We used the complete sequence of the centromeric region
246 of chromosome 8 to comparatively target the orthologous regions in other primate species in an effort
247 to fully reconstruct the evolutionary history of the centromere over the last 25 million years. We first
248 assembled reference genomes corresponding to chimpanzee, orangutan, and macaque genomes.
249 Each diploid genome assembly was sequenced to 25- to 40-fold coverage using PacBio HiFi sequence
250 data, with which we generated assemblies ranging from 6.02 to 6.12 Gbp in size, consistent with
251 assemblies where both haplotypes were assembled (**Extended Data Table 3**). Focusing on the
252 centromere, we also generated ONT datasets for the same references, which were simultaneously
253 used to construct an initial sequence scaffold of each orthologous region corresponding to the human
254 chromosome 8 centromere. Once the scaffold assembly was established and barcoded with SUNKs,
255 we used these SUNKs to replace the ONT scaffold with overlapping high-accuracy PacBio HiFi contigs.
256 We successfully generated two contiguous assemblies of the chimpanzee chromosome 8 centromere
257 (one for each haplotype), one haplotype assembly from the orangutan chromosome 8 centromere, and
258 one complete haplotype from the macaque chromosome 8 centromere (**Fig. 4**). Mapping of long-read
259 data to each assembly shows uniform coverage, indicating a lack of large structural errors (**Extended**
260 **Data Figs. 11,12**). Analysis of each nonhuman primate (NHP) chromosome 8 centromere reveals
261 distinct HOR patterns ranging in size from 1.69 Mbp in chimpanzee to 10.92 Mbp in macaque,
262 consistent with estimates from short-read sequence data and cytogenetic analyses^{33,34} (**Fig. 4**).

263
264 Similar to human, we constructed a pairwise sequence identity map of each NHP centromere. The
265 data, once again, reveal a mirrored and layered organization, with the chimpanzee organization being
266 most similar to human (**Figs. 2a, 4**). In general, each NHP chromosome 8 centromere is composed of
267 four or five distinct layers, with the outermost layer showing the lowest degree of sequence identity (73-
268 78% in chimpanzee and orangutan; 90-92% in macaque) and the innermost layer showing the highest

269 sequence identity (90-100% in chimpanzee and orangutan; 94-100% in macaque). The orangutan
270 structure is striking in that there appears to be very little admixture of HOR units between the layers, in
271 contrast to other apes where the different HOR cassettes are derived from a major HOR structure. The
272 blocks of orangutan HORs (with the exception of layer 3) show a reduced degree of sequence identity.
273 This suggests that the orangutan centromere evolved as a mosaic of independent HOR units. In
274 contrast to all apes, the macaque lacks HORs and, instead, harbors a basic dimeric repeat structure³³,
275 which is much more homogenous and highly identical (>90%) across the nearly 11 Mbp of assembled
276 centromeric array.

277

278 We assessed the phylogenetic relationship between higher-order and monomeric α -satellites from each
279 primate centromere using a maximum-likelihood framework, taking advantage of the positional
280 information from the completed sequence to define orthologous locations between the species (**Fig. 5a**).
281 We find that all great ape higher-order α -satellite sequences (corresponding to layers 2-5) cluster into a
282 single clade, while the monomeric α -satellite (layer 1) split into two clades separated by tens of millions
283 of years. The proximal clade contains monomeric α -satellite from both the p- and q-arms, while the
284 more divergent clade shares monomeric α -satellite solely from the q-arm, and specifically, the α -
285 satellite nestled between clusters of γ -satellite (**Extended Data Fig. 13**). Unlike great apes, both
286 monomeric and dimeric repeat structures from the macaque group together and are sister clades to the
287 monomeric ape clades, suggesting a common ancient origin restricted to these flanking pericentromeric
288 regions.

289

290 Because we independently assembled the centromere for each primate and successfully transitioned
291 from α -satellite to unique sequence for both the p- and q-arms, we used this orthology to understand
292 how rapidly sequences decay over the course of evolution. Anchored in orthologous sequence, we
293 assessed divergence based on 10 kbp windows of pairwise alignments in the \sim 2 Mbp flanking the α -
294 satellite HOR array (**Fig. 5b**). We find that the mean divergence increases more than threefold as the
295 sequence transitions from unique to monomeric α -satellite. Such increases are rare in the genome
296 based on sampling of at least 19,926 random loci, where only 1.27-1.99% of loci show comparable
297 levels of divergence (**Fig. 5c**). Using evolutionary models (**Methods**), we estimate a minimal mutation
298 rate of the chromosome 8 centromeric region of \sim 4.8 \times 10⁻⁸ and \sim 8.4 \times 10⁻⁸ mutations per base pair per
299 generation on the p- and q-arms, respectively, which is 2.2- to 3.8-fold higher than the basal mean
300 mutation rate (\sim 2.2 \times 10⁻⁸) (**Extended Data Table 4**). These analyses provide the first complete
301 comparative sequence analysis of a primate centromere for an orthologous chromosome and a
302 framework for future studies of genetic variation and evolution of these regions across the genome.

303

304

DISCUSSION

305

306

307

308

309

310

311

312

313

314

315

Chromosome 8 is the first human autosome to be sequenced and assembled from telomere to telomere and contains only the third completed human centromere to date^{16,35}. The assembly of chromosome 8 was achieved via a dual-technology assembly method that leverages the scaffolding capability of ONT with the high-accuracy of PacBio HiFi long reads to traverse complex regions of our genome that have remained as gaps since human genomes were first assembled^{1,2}. The result is a whole-chromosome assembly with an estimated base-pair accuracy exceeding 99.99%. We also successfully applied this hybrid strategy to reconstruct centromeric regions from diploid organisms. We generated complete draft assemblies of both chromosome 8 centromere haplotypes, for example, from a chimpanzee and a human sample. In contrast, only one haplotype was contiguously assembled in macaque and orangutan with the second haplotype remaining incomplete. It should be noted that both the human and chimpanzee diploid samples are genetically admixed, and it is possible that this

316 heterogeneity facilitated the partitioning of reads and the reconstruction of both haplotypes from these
317 samples.

318
319 Comparison of the centromeric regions between human chromosomes 8 and X reveals similarities and
320 differences in their epigenetic status and organization¹⁶. Both chromosomes harbor a pocket of
321 hypomethylation (~61-73 kbp in length), and we show that this hypomethylated region is enriched for
322 the centromeric histone, CENP-A—although CENP-A enrichment extends over a broader swath (632
323 kbp) with its peak centered over the hypomethylated region. These data strongly suggest we have
324 identified the functional kinetochore binding site^{36,37}. In contrast to the X chromosome HOR array, which
325 is primarily comprised of one type of HOR¹⁶, the chromosome 8 centromere shows a mixture of
326 primarily four different types of HORs that are present in near-equal abundance and organized into
327 layers of differing sequence identity. Although the HOR units are derived from the original 11-mer
328 repeat, the degree of admixture and purity varies considerably across the centromere, suggesting a
329 more complex model of evolution. While this layered HOR organization is evident for both
330 chromosomes X and 8, the mirror symmetry is only observed for chromosome 8. Other centromeres
331 will need to be sequenced and assembled to determine the generality of this feature (**Extended Data**
332 **Fig. 14**). Importantly, the X chromosome HOR array is less than half as diverse when compared to
333 chromosome 8¹⁶, likely due to the slower rate of mutation of the X chromosome compared to human
334 autosomes.

335
336 The layered and mirrored organization of the centromere is consistent with rapid evolutionary turnover
337 of centromeric repeats³⁸⁻⁴⁰, wherein highly identical repeats undergo unequal crossover and
338 homogenization, pushing older, more divergent repeats to the edges in an assembly-line fashion (**Fig.**
339 **5d**). The chromosome 8 centromere reveals five such layers, with the evolutionarily youngest layer in
340 the center of the α -satellite HOR array and more ancient layers flanking it. The two additional human
341 chromosome 8 centromeres, as well as each primate centromere, show a similar gradient of
342 divergence as one proceeds towards the periphery, with some of the most identical tracts mapping
343 centrally. The location and purity of the 7-mer HOR units in the human chromosome 8 centromeric
344 array, for example, are consistent with the Smith model of rapid unequal crossing over and
345 homogenization. Surprisingly, the hypomethylated regions, which we predict define the active
346 kinetochore, do not map to the most active site of homogenization as postulated by the library
347 hypothesis⁴¹. In contrast, the 73 kbp hypomethylated region maps to a segment showing some of the
348 greatest admixture, suggesting that these different HOR subtypes may be important for defining the
349 functional centromere in contrast to the most identical HOR tract. The “mosaic” architecture of the
350 orangutan HOR may be the result of a recent or even ongoing arms race to define the most competent
351 HOR associated with the kinetochore in that species. The assembled sequence allows such
352 hypotheses to be functionally tested in the future.

353
354 In addition to the centromere, we resolved other complex loci involving copy number variable SDs and
355 VNTRs. The new copies are predicted to encode novel duplicate genes (e.g., new copies of defensin
356 genes) and, as such, the additional sequence adds to our understanding of human gene annotation
357 and, subsequently, access to the underlying genetic variation therein. The completion of these
358 sequences further enhances functional annotation of the genome, showing, for example, that the
359 12.192 kbp tandem repeat defines the site of kinetochore attachment in individuals with a chromosome
360 8 neocentromere. In addition, analysis of the sequence structure provides potential insights into the
361 ontogeny of centromeres. For instance, the complete sequence of this 863 kbp VNTR shows that it
362 possesses a higher-order repetitive structure where adjacent segments flip between a direct and

363 indirect orientation in blocks ranging from 72 to 313 kbp units (**Fig. 3a**). Several studies of
364 neocentromeres devoid of satellite sequences have suggested that such inversions are commonly
365 shared features among some ectopic centromeres^{42,43}, thus providing further support that such inverted
366 configurations may be a key feature for new centromere formation.

367
368 Our targeted assembly method also facilitated draft assemblies of centromeres and their
369 heterochromatic flanking sequences from closely related NHPs. This allowed orthologous relationships
370 to be established and phylogenetic relationships among satellite repeats to be determined with respect
371 to their location. We confirm that HOR structures evolved after apes diverged from Old World monkeys
372 (OWM; <25 million years ago)^{33,44,45} but also distinguish different classes of monomeric repeats that
373 share an ancient origin with the OWM. One ape monomeric clade present only in the q-arm clearly
374 groups with the macaque's (**Extended Data Fig. 13**). We hypothesize that this ~70 kbp segment
375 present in chimpanzee and human, but absent in orangutan, represents the remnants of the ancestral
376 centromere of apes and OWM now residing only at the pericentromeric periphery of chromosome 8q.
377 The other shared monomeric clade (mapping to layer 1) continues to have diverged after apes split
378 from OWM and likely represents the origin of ape HORs. This observation supports the emergence of a
379 new class of monomers in great apes with greater potential to form centromeres⁴⁰. In contrast to apes,
380 OWM show a much simpler trajectory of continual decay of a basic dimeric HOR⁴⁶ from a 4 Mbp core of
381 near-perfect sequence identity, creating a satellite more than double in size when compared to ape
382 counterparts.

383
384 Using orthologous sequence alignments for the heterochromatin transition regions, we estimate that
385 mutation rates increase by two to fourfold in proximity to the HOR, likely due to the action of concerted
386 evolution, unequal crossing-over, and saltatory amplification^{33,39,40}. This acceleration includes
387 monomeric satellites with some evidence that it extends beyond the satellites themselves up to ~30 kbp
388 and ~170 kbp into unique regions on the p- and q-arms, respectively. It should be noted that our current
389 centromere mutation rate estimates represent a lower bound for the overall centromere, as they are
390 calculated from the monomeric α -satellite rather than the HOR array itself. Even between the human
391 and chimpanzee lineage separated by six million years, the α -satellite HOR array is too divergent to
392 generate a simple pairwise alignment that would permit the computation of a mutation rate over these
393 sequences. Novel satellite repeats are homogenized and swept to fixation within a population through
394 mechanisms such as unequal crossing over and gene conversion⁴⁷. Rapid evolution of centromeres
395 has been described in multiple species as a driving force for speciation due to the accumulation of
396 sequence differences that result in highly divergent HOR sequences, and subsequently, causes
397 incompatibility and reproductive barrier in hybrids between closely related species⁴⁸⁻⁵⁰. In this light, it is
398 interesting that sequence comparisons among three human centromere 8 haplotypes predict regions of
399 excess allelic variation and structural divergence (**Extended Data Fig. 7c-e**), although the locations
400 within the HOR differ among haplotypes (**Extended Data Fig. 7**). Now that complex regions such as
401 these can be sequenced and assembled, it will be important to extend these analyses to other
402 centromeres, multiple individuals, and additional species to understand their full impact with respect to
403 genetic variation and evolution.

404

405 **METHODS**

406

407 **Cell culture**

408 CHM13hTERT (abbr. CHM13) cells were cultured in complete AmnioMax C-100 Basal Medium
409 (Thermo Fisher Scientific, 17001082) supplemented with 15% AmnioMax C-100 Supplement (Thermo
410 Fisher Scientific, 12556015) and 1% penicillin-streptomycin (Thermo Fisher Scientific, 15140122).
411 Chimpanzee (*Pan troglodytes*; Clint; S006007) and macaque (*Macaque mulatta*; AG07107) cells were
412 cultured in MEM α containing ribonucleosides, deoxyribonucleosides, and L-glutamine (Thermo Fisher
413 Scientific, 12571063) supplemented with 12% FBS (Thermo Fisher Scientific, 16000-044) and 1%
414 penicillin-streptomycin (Thermo Fisher Scientific, 15140122). Orangutan (*Pongo abelii*; Susie;
415 PR01109) cells were cultured in MEM α containing ribonucleosides, deoxyribonucleosides, and L-
416 glutamine (Thermo Fisher Scientific, 12571063) supplemented with 15% FBS (Thermo Fisher
417 Scientific, 16000-044) and 1% penicillin-streptomycin (Thermo Fisher Scientific, 15140122). All cells
418 were cultured in a humidity-controlled environment at 37°C with 95% O₂ and 5% CO₂.

419

420 **DNA extraction, library preparation, and sequencing**

421 PacBio HiFi data were generated from the chimpanzee, orangutan, and macaque genomes as
422 previously described⁵¹ with modifications. Briefly, high-molecular-weight (HMW) DNA was extracted
423 from cells using a modified Qiagen Gentra Puregene Cell Kit protocol⁵². HMW DNA was used to
424 generate HiFi libraries via the SMRTbell Express Template Prep Kit v2 and SMRTbell Enzyme Clean
425 Up kits (PacBio). Size selection was performed with SageELF (Sage Science), and fractions sized 11,
426 14, 18, 22, or 25 kbp (as determined by FEMTO Pulse (Agilent)) were chosen for sequencing. Libraries
427 were sequenced on the Sequel II platform with three to seven SMRT Cells 8M (PacBio) using either
428 Sequel II Sequencing Chemistry 1.0 and 12-hour pre-extension or Sequel II Sequencing Chemistry 2.0
429 and 3- or 4-hour pre-extension, both with 30-hour movies, aiming for a minimum estimated coverage of
430 25X in HiFi reads (assuming a genome size of 3.2 Gbp). Raw data was processed using the CCS
431 algorithm (v3.4.1 or v4.0.0) with the following parameters: –minPasses 3 –minPredictedAccuracy 0.99
432 –maxLength 21000 or 50000.

433

434 Ultra-long ONT data were generated from the CHM13, chimpanzee, and orangutan genomes according
435 to a previously published protocol⁵³. Briefly, 5 x 10⁷ cells were lysed in a buffer containing 10 mM Tris-
436 Cl (pH 8.0), 0.1 M EDTA (pH 8.0), 0.5% w/v SDS, and 20 ug/mL RNase A for 1 hour at 37C. Proteinase
437 K (200 ug/mL) was added, and the solution was incubated at 50C for 2 hours. DNA was purified via two
438 rounds of 25:24:1 phenol-chloroform-isoamyl alcohol extraction followed by ethanol precipitation.
439 Precipitated DNA was solubilized in 10 mM Tris (pH 8) containing 0.02% Triton X-100 at 4C for two
440 days. Libraries were constructed using the Rapid Sequencing Kit (SQK-RAD004) from ONT with
441 modifications to the manufacturer's protocol. Specifically, 2-3 ug of DNA was resuspended in a total
442 volume of 18 ul with 16.6% FRA buffer. FRA enzyme was diluted 2- to 12-fold into FRA buffer, and 1.5
443 uL of diluted FRA was added to the DNA solution. The DNA solution was incubated at 30C for 1.5 min,
444 followed by 8C for 1 min to inactivate the enzyme. RAP enzyme was diluted 2- to 12-fold into RAP
445 buffer, and 0.5 uL of diluted RAP was added to the DNA solution. The DNA solution was incubated at
446 room temperature (RT) for 2 hours before loading onto a primed FLO-MIN106 R9.4.1 flow cell for
447 sequencing.

448

449 Additional ONT data was generated from the CHM13, chimpanzee, orangutan, and macaque genomes.
450 Briefly, HMW DNA was extracted from cells using a modified Qiagen Gentra Puregene Cell Kit
451 protocol⁵². HMW DNA was prepared into libraries with the Ligation Sequencing Kit (SQK-LSK109) from

452 ONT and loaded onto primed FLO-MIN106 or FLO-PRO002 R9.4.1 flow cells for sequencing on the
453 GridION or PromethION, respectively. All ONT data were base called with Guppy 3.6.0 or 4.11.0 with
454 the HAC model.

455

456 **PacBio HiFi whole-genome assembly**

457 Chimpanzee, orangutan, and macaque genomes were assembled from PacBio HiFi data (**Extended**
458 **Data Table 2**) using HiCanu¹¹ (v2.0). The CHM13 genome was previously assembled with HiCanu and
459 described by Nurk and colleagues¹¹. Contigs from each primate assembly were used to replace the
460 ONT-based sequence scaffolds in targeted regions (described below).

461

462 **Targeted sequence assembly**

463 Gapped regions within human chromosome 8 were targeted for assembly via a SUNK-based method
464 that combines both PacBio HiFi and ONT data. Specifically, CHM13 PacBio HiFi data was used to
465 generate a library of SUNKs ($k = 20$; total = 2,062,629,432) via Jellyfish (v2.2.4) based on the
466 sequencing coverage of the HiFi dataset. 99.88% (2,060,229,331) of the CHM13 PacBio HiFi SUNKs
467 were validated with CHM13 Illumina data (SRR3189741). A subset of CHM13 ultra-long ONT reads
468 aligning to the CHM1 β -defensin patch (GenBank: KZ208915.1) or select regions within the GRCh38
469 chromosome 8 reference sequence (chr8:42,881,543-47,029,467 for the centromere and
470 chr8:85,562,829-85,848,463 for the 8q21.2 locus) were barcoded with Illumina-validated SUNKs.
471 Reads sharing at least 50 SUNKs were selected for inspection to determine if their SUNK barcodes
472 overlapped. SUNK barcodes can be composed of “valid” and “invalid” SUNKs. Valid SUNKs are those
473 that occur once in the genome and are located at the exact position on the read. In contrast, invalid
474 SUNKs are those that occur once in the genome but are falsely located at the position on the read, and
475 this may be due to a sequencing or base-calling error, for example. Valid SUNKs were identified within
476 the barcode as those that share pairwise distances with at least ten other SUNKs on the same read.
477 Reads that shared a SUNK barcode containing at least three valid SUNKs and their corresponding
478 pairwise distances ($\pm 1\%$ of the read length) were assembled into a tile. The process was repeated
479 using the tile and subsetted ultra-long ONT reads several times until a sequence scaffold spanning the
480 gapped region was generated. Validation of the scaffold organization was carried out via three
481 independent methods. First, the sequence scaffold and underlying ONT reads were subjected to
482 RepeatMasker (v3.3.0) to ensure that read overlaps were concordant in repeat structure. Second, the
483 centromeric scaffold and underlying ONT reads were subjected to StringDecomposer⁵⁴ to validate the
484 HOR organization in overlapping reads. Finally, the sequence scaffold for each target region was
485 incorporated into the CHM13 chromosome 8 assembly¹¹, thereby filling the gaps in the chromosome 8
486 assembly. CHM13 PacBio HiFi and ONT data were aligned to the entire chromosome 8 assembly via
487 pbmm2 (v1.1.0) (for PacBio data; <https://github.com/PacificBiosciences/pbmm2>) or Winnowmap⁵⁵
488 (v1.0) (for ONT data) to identify large collapses or misassemblies. Although the ONT-based scaffolds
489 are structurally accurate, they are only 87-98% accurate at the base level due to base-calling errors in
490 the raw ONT reads¹³. Therefore, we sought to improve the base accuracy of the sequence scaffolds by
491 replacing the ONT sequences with PacBio HiFi contigs assembled from the CHM13 genome¹¹, which
492 have a consensus accuracy greater than 99.99%¹¹. Therefore, we aligned CHM13 PacBio HiFi contigs
493 generated via HiCanu¹¹ to the chromosome 8 assembly via minimap2⁵⁶ (v2.17-r941; parameters:
494 minimap2 -t 8 -l 8G -a --eqx -x asm20 -s 5000) to identify contigs that share high sequence identity with
495 the ONT-based sequence scaffolds. A typical scaffold had multiple PacBio HiFi contigs that aligned to
496 regions within it. Therefore, the scaffold was used to order and orient the PacBio HiFi contigs and
497 bridge gaps between them when necessary. PacBio HiFi contigs with high sequence identity replaced
498 almost all regions of the ONT-based scaffolds: ultimately, the chromosome 8 assembly is comprised of

499 146,254,195 bp of PacBio HiFi contigs and only 5,490 bp of ONT sequence scaffolds (99.9963%
500 PacBio HiFi contigs and 0.0037% ONT scaffold). The chromosome 8 assembly was incorporated into a
501 whole-genome assembly of CHM13¹¹ for validation via orthogonal methods (detailed below). The
502 chimpanzee, orangutan, and macaque chromosome 8 centromeres were assembled via the same
503 SUNK-based method.

504

505 **Accuracy estimation**

506 The accuracy of the CHM13 chromosome 8 assembly was estimated from mapped k-mers using
507 Merqury²⁴. Briefly, Merqury (v1.1) was run on the chromosome 8 assembly with the following
508 command: eval/qv.sh CHM13.k21.meryl chr8.fasta chr8_v9.

509

510 CHM13 Illumina data (SRR1997411, SRR3189741, SRR3189742, SRR3189743) was used to identify
511 k-mers with $k = 21$. In Merqury, every k-mer in the assembly is evaluated for its presence in the Illumina
512 k-mer database, with any k-mer missing in the Illumina set counted as base-level 'error'. We detected
513 1,474 k-mers found only in the assembly out of 146,259,650, resulting in a QV score of 63.19,
514 estimated as follows:

515

$$516 \quad -10 * \log(1 - (1 - 1474 / 146259650)^{(1/21)}) = 63.19$$

517

518 The accuracy percentage for chromosome 8 was estimated from this QV score as:

519

$$520 \quad 100 - (10^{(63.19/-10)}) * 100 = 99.999952$$

521 The accuracy of the CHM13 chromosome 8 assembly and β -defensin locus were also estimated from
522 sequenced BACs. Briefly, 66 BACs from the CHM13 chromosome 8 (BAC library VMRC59) were
523 aligned to the chromosome 8 assembly via minimap2⁵⁶ (v2.17-r941) with the following parameters: -l
524 8G -2K 1500m --secondary=no -a --eqx -Y -x asm20 -s 200000 -z 10000,1000 -r 50000 -O 5,56 -E 4,1 -
525 B 5. QV was then estimated using the CIGAR string in the resulting BAM, counting alignment
526 differences as errors according to the following formula:

527

$$528 \quad QV = -10 * \log_{10}[1 - (\text{matches} / (\text{mismatches} + \text{matches} + \text{insertions} + \\ 529 \quad \text{deletions}))]$$

530

531 The median QV was 40.6988 for the entire chromosome 8 assembly and 40.4769 for the β -defensin
532 locus (chr8:6300000-13300000; estimated from 47 individual BACs; see **Extended Data Fig. 4** for
533 more details), which falls within the 95% confidence interval for the whole chromosome. This QV score
534 was used to estimate the base accuracy⁵¹ as follows:

535

$$536 \quad 100 - (10^{(40.6988/-10)}) * 100 = 99.9915$$

537

$$100 - (10^{(40.4769/-10)}) * 100 = 99.9910$$

538

539 The BAC QV estimation should be considered a lower bound, since differences between the BACs and
540 the assembly may originate from errors in the BAC sequences themselves. Vollger and colleagues
541 showed that BACs can occasionally contain sequencing errors that are not supported by the underlying
542 PacBio HiFi reads⁵¹. Additionally, the upper bound for the estimated BAC QV is limited to ~53, since
543 BACs are typically ≤ 200 kbp and, as a result, the maximum calculable QV is 1 error in 200 kbp (QV

544 53). We also note that the QV of the centromeric region could not be estimated from BACs due to
545 biases in BAC library preparation, which preclude centromeric sequences in BAC clones.

546

547 **Strand-seq analysis**

548 We evaluated the directional and structural contiguity of CHM13 chromosome 8 assembly, including the
549 centromere, using Strand-seq data. First, all Strand-seq libraries produced from the CHM13 genome⁵¹
550 were aligned to the CHM13 assembly, including chromosome 8 using BWA-MEM⁵⁷ (v0.7.17-r1188) with
551 default parameters for paired-end mapping. Next, duplicate reads were marked by sambamba⁵⁸
552 (v0.6.8) and removed before subsequent analyses. We used SAMtools⁵⁹ (v1.9) to sort and index the
553 final BAM file for each Strand-seq library. To detect putative misassembly breakpoints in the
554 chromosome 8 assembly, we ran breakpointR⁶⁰ on all BAM files to detect strand-state breakpoints.
555 Misassemblies are visible as recurrent changes in strand state across multiple Strand-seq libraries⁶¹.
556 To increase our sensitivity of misassembly detection, we created a 'composite file' that groups
557 directional reads across all available Strand-seq libraries^{62,63}. Next, we ran breakpointR on the
558 'composite reads file' using the function 'runBreakpointR' to detect regions that are homozygous ('ww';
559 'HOM' - all reads mapped in minus orientation) or heterozygous inverted ('wc', 'HET' - approximately
560 equal number of reads mapped in minus and plus orientation). To further detect any putative chimerism
561 in the chromosome 8 assembly, we applied Strand-seq to assign 200 kbp long chunks of the
562 chromosome 8 assembly to unique groups corresponding to individual chromosomal homologues using
563 SaaRclust^{61,64}. For this, we used the SaaRclust function 'scaffoldDenovoAssembly' on all BAM files.

564

565 **Bionano analysis**

566 Bionano Genomics data was generated from the CHM13 genome¹⁶. Long DNA molecules labeled with
567 Bionano's Direct Labeling Enzyme were collected on a Bionano Saphyr Instrument to a coverage of
568 130X. The molecules were assembled with the Bionano assembly pipeline Solve3.4, using the
569 nonhaplotype-aware parameters and GRCh38 as the reference. The resulting data produced 261
570 genome maps with a total length of 2921.6 Mbp and a genome map N50 of 69.02 Mbp.

571

572 The molecule set and the nonhaplotype-aware map were aligned to the CHM13 draft assembly and the
573 GRCh38 assembly, and discrepancies were identified between the Bionano maps and the sequence
574 references using scripts in the Bionano software package -- runCharacterize.py, runSV.py, and
575 align_bnx_to_cmap.py.

576

577 A second version of the map was assembled using the haplotype-aware parameters. This map was
578 also aligned to GRCh38 and the final CHM13 assembly to verify heterozygous locations. These regions
579 were then examined further.

580

581 Analysis of Bionano alignments revealed three heterozygous sites within chromosome 8 located at
582 approximately chr8:21,025,201, chr8:80,044,843, and chr8:121,388,618 (**Extended Data Table 5**). The
583 structure with the greatest ONT read support was selected for inclusion in the chromosome 8 assembly
584 (**Extended Data Table 5**).

585

586 **TandemMapper and TandemQUAST analysis of the centromeric HOR array**

587 We assessed the structure of the CHM13 and NHP centromeric HOR arrays by applying
588 TandemMapper and TandemQUAST⁶⁵, which can detect large structural assembly errors in repeat
589 arrays. For the CHM13 centromere, we first aligned ONT reads longer than 50 kbp to the CHM13
590 assembly containing the contiguous chromosome 8 with Winnowmap⁵⁵ (v1.0) and extracted reads

591 aligning to the centromeric HOR repeat array (chr8:44243868-46323885). We then inputted these
592 reads in the following TandemQUAST command: tandemquast.py -t 24 --nano {ont_reads.fa} -o
593 {out_dir} chr8.fa. For the NHP centromeres, we aligned ONT reads to the whole-genome assemblies
594 containing the contiguous chromosome 8 centromeres with Winnowmap⁵⁵ (v1.0) and extracted reads
595 aligning to the centromeric HOR repeat arrays. We then inputted these reads in the following
596 TandemQUAST command: tandemquast.py -t 24 --nano {ont_reads.fa} -o {out_dir} chr8.fa.

597

598 **Methylation analysis**

599 Nanopolish²⁵ (v0.12.5) was used to measure CpG methylation from raw ONT reads (>50 kbp in length
600 for CHM13) aligned to whole-genome assemblies via Winnowmap⁵⁵ (v1.0). Nanopolish distinguishes 5-
601 methylcytosine from unmethylated cytosine via a Hidden Markov Model (HMM) on the raw nanopore
602 current signal. The methylation caller generates a log-likelihood value for the ratio of probability of
603 methylated to unmethylated CpGs at a specific k-mer. We filtered methylation calls using the
604 nanopore_methylation_utilities tool (<https://github.com/isaclee/nanopore-methylation-utilities>)⁶⁶, which
605 uses a log-likelihood ratio of 2.5 as a threshold for calling methylation. CpG sites with log-likelihood
606 ratios greater than 2.5 (methylated) or less than -2.5 (unmethylated) are considered high quality and
607 included in the analysis. Reads that do not have any high-quality CpG sites are filtered from the BAM
608 for subsequent methylation analysis. Nanopore_methylation_utilities integrates methylation information
609 into the BAM file for viewing in IGV's⁶⁷ bisulfite mode, which was used to visualize CpG methylation.

610

611 **Iso-Seq data generation and sequence analyses**

612 RNA was purified from approximately 1×10^7 CHM13 cells using an RNeasy kit (Qiagen; 74104) and
613 prepared into Iso-Seq libraries following a standard protocol⁶⁸. Libraries were loaded on two SMRT
614 Cells 8M and sequenced on the Sequel II. The data were processed via isoseq3 (v8.0), ultimately
615 generating 3,576,198 full-length non-chimeric (FLNC) reads. Poly-A trimmed transcripts were aligned to
616 this CHM13 chr8 assembly and to GRCh38 with minimap2⁵⁶ (v2.17-r941) with the following parameters:
617 -ax splice -f 1000 --sam-hit-only --secondary=no --eqx. Transcripts were assigned to genes using
618 featureCounts⁶⁹ with GENCODE⁷⁰ (v34) annotations, supplemented with CHESSE v2.2⁷¹ for any
619 transcripts unannotated in GENCODE. Each transcript was scored for percent identity of its alignment
620 to each assembly, requiring 90% of the length of each transcript to align to the assembly for it to count
621 as aligned. For each gene, non-CHM13 transcripts' percent identity to GRCh38 was compared to the
622 CHM13 chromosome 8 assembly. Genes with an improved representation in the CHM13 assembly
623 were identified with a cutoff of 20 improved reads per gene, with at least 0.2% average improvement in
624 percent identity. GENCODE (v34) transcripts were lifted over to the CHM13 chr8 assembly using
625 LiftOff⁷² to compare the GRCh38 annotations to this assembly and Iso-Seq transcripts.

626

627 We combined the 3.6 million full-length transcript data (above) with 20,937,742 FLNC publicly available
628 human Iso-Seq data (**Extended Data Table 6**). In total, we compared the alignment of 24,513,940
629 FLNC reads from 13 tissue and cell line sources to both the completed CHM13 chromosome 8
630 assembly and the current human reference genome, GRCh38. Of the 848,048 non-CHM13 cell line
631 transcripts that align to chromosome 8, 93,495 (11.02%) align with at least 0.1% greater percent
632 identity to the CHM13 assembly, and 52,821 (6.23%) to GRCh38. This metric suggests that the
633 chromosome 8 reference improves human gene annotation by ~4.79% even though most of those
634 changes are subtle in nature. Overall, 61 protein-coding and 33 noncoding loci have improved
635 alignments to the CHM13 assembly compared to GRCh38, with >0.2% average percent identity
636 improvement, and at least 20 supporting transcripts (**Extended Data Fig. 4, Extended Data Table 7**).

637 As an example, *WDYHV1 (NTAQ1)* has four amino acid replacements, with 13 transcripts sharing the
638 identical open reading frame to CHM13 (**Extended Data Fig. 16**).
639

640 **Pairwise sequence identity heat maps**

641 To generate pairwise sequence identity heat maps, we fragmented the centromere assemblies into
642 5 kbp fragments (e.g., 1-5000, 5001-10000, etc.) and made all possible pairwise alignments between
643 the fragments using the following minimap2⁵⁶ (v2.17-r941) command: `minimap2 -f 0.0001 -t 32 -X --eqx`
644 `-ax ava-ont`. The sequence identity was determined from the CIGAR string of the alignments and then
645 visualized using `ggplot2 (geom_raster)` in R (v1.1.383)⁷³. The color of each segment was determined by
646 sorting the data by identity and then creating 10 equally-sized bins, each of which received a distinct
647 color from the spectral pallet. The choice of a 5 kbp window came after testing a variety of window
648 sizes. Ultimately, we found 5 kbp to be a good balance between resolution of the figure (since each 5
649 kbp fragment is plotted as a pixel) and sensitivity of minimap2 (fragments less than 5 kbp often missed
650 alignments with the `ava-ont` preset).
651

652 **Analysis of α -satellite organization**

653 To determine the organization of the chromosome 8 centromeric region, we employed two independent
654 approaches. First, we subjected the CHM13 centromere assembly to an *in silico* restriction enzyme
655 digestion wherein a set of restriction enzyme recognition sites were identified within the assembly. In
656 agreement with previous findings that *XbaI* digestion can generate a pattern of HORs within the
657 chromosome 8 HOR array¹⁷, we found that each α -satellite HOR could be extracted via *XbaI* digestion.
658 The *in silico* digestion analysis indicates that the chromosome 8 centromeric HOR array is comprised of
659 1462 HOR units: 283 4-mers, 4 5-mers, 13 6-mers, 356 7-mers, 295 8-mers, and 511 11-mers. As an
660 alternative approach, we subjected the centromere assembly to *StringDecomposer*⁵⁴ using a set of 11
661 α -satellite monomers derived from a chromosome 8 11-mer HOR unit. The sequence of the α -satellite
662 monomers used are as follows: A:

663 AGCATTCTCAGAAACACCTTCGTGATGTTTGCAATCAAGTCACAGAGTTGAACCTTCCGTTTCATAG
664 AGCAGTTGGAAACACTCTTATTGTAGTATCTGGAAGTGGACATTTGGAGCGCTTTCAGGCCTATG
665 GTGAAAAGGAAATATCTTCCATAAAAACGACATAGA; B:

666 AGCTATCTCAGGAACCTTGTATGATGCATCTAATCAACTAACAGTGTTGAACCTTTGTAAGTGCAG
667 AGCACTTTGAAACACTCTTTTTTGGAACTGCAAGTGGATATTTGGATCGCTTTGAGGATTTGCTTG
668 GAAACGGGATGCAATATAAACGTACACAGC; C:

669 AGCATACTCAGAAAATACTTTGCCATATTTCCATTCAAGTCACAGAGTGGAAACATTCCCATTTCATAG
670 AGCAGTTGGAAACACTCTTTTTGGAGTATCTGGAAGTGGACATTTGGAGCGCTTCTGAACTATG
671 GTGAAAAGGAAATATCTTCCAATGAAAACAAGACAGA; D:

672 AGCATTCTGAGAACTTATTTGTGATGTGTGTCCTCAACAAACGGACTTGAACCTTTGTTTCATGC
673 AGTACTTCTGGAACACTCTTTTTGAAGATTCTGCATGCGGATATTTGGATAGCTTTGAGGATTTGCT
674 TGAAACGGGCTTACATGTAAAATTAGACAGC; E:

675 AGCATTCTCAGAACTTCTTTGTGGTGTCTGCATTCAAGTCACAGAATTGAACTTCTCCTCACATAG
676 AGCAGTTGTGCAGCACTCTATTTGTAGTATCTGGAAGTGGACATTTGGAGGGCTTTGTAGCCTATC
677 TGGAAAAGGAAATATCTTCCCATGAATGCGAGATAGA; F:

678 AGTAATCTCAGAAACATGTTTATGCTGTATCTACTCAACTAACTGTGCTGAACATTTCTATTGATAGA
679 GCAGTTTTGAGACCCTCTTCTTTTGGAACTGCAAGTGGATATTTGGATAGATTTGAGGATTTGCTT
680 GGAAACGGGATTATATAAAAAGTAGACAGC; G:

681 AGCATTCTCAGAACTTCTTTGTGATGTTTGCATCCAGCTCTCAGAGTTGAACATTCCCTTTTCATAG
682 AGTAGTTTGAACCTCTTTTTATAGTGTCTGGAAGCGGGCATTGGAGCGCTTTCAGGCCTATG
683 CTGAAAAGGAAATATCTACATATAGAACTAGACAGA; H:

684 AGCATTCTGAGAATCAAGTTTGTGATGTGGGTACTCAACTAACAGTGTTGATCCATTCTTTTGATAC
685 AGCAGTTTTGAACCACACTTTTTGTAGAATCTGCAAGTGGATATTTGGATAGCTGTGAGGATTTTCGT
686 TGGAAACGGGAATGTCTTCATAGAAAATTTAGACAGA; I:
687 AGCATTCTCAGAACCTTGATTGTGATGTGTGTTCTCCACTAACAGAGTTGAACCTTTCTTTTGACAG
688 AACTGTTCTGAAACATTCTTTTTATAGAATCTGGAAGTGGATATTTGGAAAGCTTTGAGGATTTTCGT
689 TGGAAACGGGAATATCTTCAAATAAAATCTAGCCAGA; J:
690 AGCATTCTAAGAAACATCTTAGGGATGTTTACATTCAAGTCACAGAGTTGAACATTCCCTTTACAG
691 AGCAGGTTTGAACAATCTTCTCGTACTATCTGGCAGTGGACATTTTGAGCTCTTTGGGGCCTATG
692 CTGAAAAGGAAATATCTTCCGACAAAAACTAGTCAGA; K:
693 AGCATTCGCAGAATCCCGTTTGTGATGTGTGCACTCAACTGTCAGAATTGAACCTTGGTTTGGAGA
694 GAGCACTTTTGAACACACTTTTTGTAGAATCTGCAGGTGGATATTTGGCTAGCTTTGAGGATTTTCG
695 TTGGAAACGGTAATGTCTTCAAAGAAAATCTAGACAGA.

696
697 This analysis indicated that the chromosome 8 centromeric HOR array is comprised of 1512 HOR units:
698 283 4-mers, 12 6-mers, 366 7-mers, 303 8-mers, 3 10-mers, 539 11-mers, 2 12-mers, 2 13-mers, 1 17-
699 mer, and 1 18-mer, which is concordant with the *in silico* restriction enzyme digestion results. The
700 predominant HOR types from StringDecomposer are presented in **Extended Data Fig. 8**.

701

702 **Copy number estimation**

703 To estimate the copy number for the 8q21.2 VNTR and *DEFB* loci in human lineages, we applied a
704 read-depth based copy number genotyper²¹ to a collection of 1,112 published high-coverage
705 genomes⁷⁴⁻⁷⁹. Briefly, sequencing reads were divided into multiples of 36-mer, which were then
706 mapped to a repeat-masked human reference genome (GRCh38) using mrsFAST⁸⁰ (v3.4.1). To
707 increase the mapping sensitivity, we allowed up to two mismatches per 36-mer. The read depth of
708 mappable sequences across the genome was corrected for underlying GC content, and copy number
709 estimate for the locus of interest was computed by summarizing over all mappable bases for each
710 sample.

711

712 **Entropy calculation**

713 To define regions of increased admixture within the centromeric HOR array, we calculated the entropy
714 using the frequencies of the different HOR units in 10-unit windows (1 unit slide) over the entire array.
715 The formula for entropy is:

716

$$717 \text{Entropy} = -\sum(\text{frequency}_i * \log_2(\text{frequency}_i))$$

718

719 where frequency is (# of HORs) / (total # of HORs) in a 10-unit window. The analysis is analogous to
720 that performed by Gymrek and colleagues⁸¹.

721

722 **Droplet digital PCR**

723 Droplet digital PCR was performed on CHM13 genomic DNA to estimate the number of D8Z2 α -
724 satellite HORs, as was previously done for the DXZ1 α -satellite HORs¹⁶. Briefly, genomic DNA was
725 isolated from CHM13 cells using the DNeasy Blood & Tissue Kit (Qiagen). DNA was quantified using a
726 Qubit Fluorometer and the Qubit dsDNA HS Assay (Invitrogen). 20 μ L reactions were prepared with 0.1
727 ng of gDNA for the D8Z2 assay or 1 ng of gDNA for the *MTUS1* single-copy gene (as a control).
728 EvaGreen droplet digital PCR (Bio-Rad) master mixes were simultaneously prepared for the D8Z2 and
729 *MTUS1* reactions, which were then incubated for 15 minutes to allow for restriction digest, according to
730 the manufacturer's protocol.

731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776

Pulsed-field gel electrophoresis and Southern blot

CHM13 genomic DNA was prepared in agarose plugs and digested with either BamHI or MfeI (to characterize the chromosome 8 centromeric region) or BmgBI (to characterize the chromosome 8q21.2 region) in the buffer recommended by the manufacturer. The digested DNA was separated with the CHEF Mapper system (Bio-Rad; autoprogram, 5-850 kbp range, 16 hr run), transferred to a membrane (Amersham Hybond-N+) and blot-hybridized with a 156 bp probe specific to the chromosome 8 centromeric α -satellite or 8q21.2 region. The probe was labeled with P^{32} by PCR-amplifying a synthetic DNA template #233: 5'-TTTGTGGAAGTGGACATTTTCGCTTTGTAGCCTATCTGGAAAAAGGAAATATCTTCCCATGAATGCGAGATAGAAGTAATCTCAGAAACATGTTTATGCTGTATCTACTCAACTAACTGTGCTGAACATTTCTATTGTAATAAATAGACAGAAGCATT-3' (for the centromere of chromosome 8); #264: 5'-TTTGTGGAAGTGGACATTTTCGCCCCGAGGGGCGCGGCAGGGATTCCGGGGGACCGGGAGTGGGGGGTTGGGGTTACTCTTGGCTTTTTGCCCCTCTCCTGCCGCCGGCTGCTCCAGTTTCTTTTCGCTTTGCGGCGAGGTGGTAAAAATAGACAGAAGCATT-3' (for the organization of the chromosome 8q21.2 locus) with PCR primers #129: 5'-TTTGTGGAAGTGGACATTTTC-3' and #130: 5'-AATGCTTCTGTCTATTTTTTA-3'. The blot was incubated for 2 hr at 65°C for pre-hybridization in Church's buffer (0.5 M Na-phosphate buffer containing 7% SDS and 100 μ g/ml of unlabeled salmon sperm carrier DNA). The labeled probe was heat denatured in a boiling water bath for 5 min and snap-cooled on ice. The probe was added to the hybridization Church's buffer and allowed to hybridize for 48 hr at 65°C. The blot was washed twice in 2 \times SSC (300 mM NaCl, 30 mM sodium citrate, pH 7.0), 0.05% SDS for 10 min at room temperature, twice in 2 \times SSC, 0.05% SDS for 5 min at 60°C, twice in 0.5 \times SSC, 0.05% SDS for 5 min at 60°C, and twice in 0.25 \times SSC, 0.05% SDS for 5 min at 60°C. The blot was exposed to X-ray film for 16 hr at -80°C.

Immunofluorescence (IF) and fluorescence *in situ* hybridization (FISH) on chromatin fibers

To determine the location of CENP-A relative to methylated DNA (specifically, 5-methylcytosines), we performed IF on stretched CHM13 chromatin fibers as previously described with modifications^{82,83}. Briefly, CHM13 cells were swollen in a hypotonic buffer consisting of a 1:1:1 ratio of 75 mM KCl, 0.8% NaCitrate, and dH₂O for 5 min. 3.5×10^4 cells were cytospun onto an ethanol-washed glass slide at 800 rpm for 4 min with high acceleration and allowed to adhere for 1 min before immersing in a salt-detergent-urea lysis buffer (25 mM Tris pH 7.5, 0.5 M NaCl, 1% Triton X-100, and 0.3 M urea) for 15 min at room temperature. The slide was slowly removed from the lysis buffer over a time period of 38 s and subsequently washed in PBS, incubated in 4% formaldehyde in PBS for 10 min, and washed with PBS and 0.1% Triton X-100. The slide was rinsed in PBS and 0.05% Tween-20 (PBST) for 3 min, blocked for 30 min with IF block (2% FBS, 2% BSA, 0.1% Tween-20, and 0.02% NaN₂), and then incubated with a mouse monoclonal anti-CENP-A antibody (1:200, Enzo, ADI-KAM-CC006-E) and rabbit monoclonal anti-5-methylcytosine antibody (1:200, RevMAb, RM231) for 3 h at room temperature. Cells were washed 3x for 5 min each in PBST and then incubated with Alexa Fluor 488 goat anti-rabbit (1:200, Thermo Fisher Scientific, A-11034) and Alexa Fluor 594 conjugated to goat anti-mouse (1:200, Thermo Fisher Scientific, A-11005) for 1.5 h. Cells were washed 3 \times for 5 min each in PBST, fixed for 10 min in 4% formaldehyde, and washed 3 \times for 1 min each in dH₂O before mounting in Vectashield containing 5 μ g/ml DAPI. Slides were imaged on an inverted fluorescence microscope (Leica DMI6000) equipped with a charge-coupled device camera (Leica DFC365 FX) and a 40 \times 1.4 NA objective lens.

777 To assess the repeat organization of the 8q21 neocentromere, we performed FISH⁸⁴ on CHM13
778 chromatin fibers. DNA fibers were obtained following Henry H. Q. Heng's protocol with minor
779 modifications⁸⁵. Briefly, chromosomes were fixed with methanol:acetic acid (3:1) and dropped on
780 previously clean slides. Chromosomes were dropped onto slides and soaked in PBS 1x. Manual
781 elongation was performed by coverslip and NaOH:ethanol (5:2) solution. We quantified the number and
782 intensity of the probe signals on a set of CHM13 chromatin fibers using ImageJ's Gel Analysis tool
783 (v1.51) and found that there were 63 +/- 7.55 green signals and 67 +/- 5.20 red signals (n=3
784 independent experiments), consistent with the 67 full and 7 partial repeats in the CHM13 8q21.2 VNTR.
785

786 To validate the organization of the chromosome 8 centromere, we performed FISH on CHM13
787 cytopun metaphase chromosome spreads in order to increase the chromosome length and improve
788 the resolution of the experiments. We followed the Haaf and Ward protocol⁸⁶ with slight modifications.
789 Briefly, cells were treated with colcemid and resuspended in HCM buffer (10 mM HEPES pH7.3, 30 mM
790 glycerol, 1 mM CaCl₂, 0.8 mM MgCl₂) and after 10 minutes, cytopun on silanized slides. Incubation
791 overnight in cold methanol was required to fully fix the chromosomes.
792

793 The probes used in the FISH experiments were picked from the human large-insert clone fosmid library
794 ABC10. ABC10 end sequences were mapped using MEGABLAST (similarity=0.99, parameters: -D 2 -v
795 7 -b 7 -e 1e-40 -p 80 -s 90 -W 12 -t 21 -F F) to a repeat-masked CHM13 genome assembly containing
796 the complete chromosome 8 (parameters: -e wublast -xsmall -no_is -s -species Homo sapiens).
797 Expected insert size for fosmids was set to (min) 32 kbp and (max) 48 kbp. Resulting clone alignments
798 were grouped into the following categories based on uniqueness of the alignment for a given pair of
799 clones, alignment orientation and the inferred insert size from the assembly.

- 800 1. Concordant best: unique alignment for clone pair, insert size within expected fosmid range,
801 expected orientation
- 802 2. Concordant tied: non-unique alignment for clone pair, insert size within expected fosmid range,
803 expected orientation
- 804 3. Discordant best: unique alignment of clone pair, insert size too small, too large or in opposite
805 expected orientation of expected fosmid clone
- 806 4. Discordant tied: non unique alignment for clone pair, insert size too small, too large or in
807 opposite expected orientation of expected fosmid clone
- 808 5. Discordant trans: clone pair has ends mapping to different contigs
809

810 Clones aligning to regions within the chromosome 8 centromeric region or 8q21.2 locus were selected
811 for FISH validation. The fosmid clones used for validation of the chromosome 8 centromeric region are:
812 174552_ABC10_2_1_000046302400_C7 for the monomeric α -satellite region,
813 171417_ABC10_2_1_000045531400_M19 for the entire D8Z2 HOR array,
814 174222_ABC10_2_1_000044375100_H13 for the central portion of the D8Z2 HOR array. The clones used
815 for validation of the 8q21.2 locus are: 174552_ABC10_2_1_000044787700_O7 for Probe 1 and
816 173650_ABC10_2_1_000044086000_F24 for Probe 2.
817

818 **CENP-A ChIP-seq analysis**

819 We mapped previously published CENP-A ChIP-seq and whole-genome sequencing (WGS) data using
820 two different approaches: 1) BWA-MEM⁸⁷, and 2) a k-mer-based mapping approach we developed
821 (described below). Both results were highly concordant, as shown in **Extended Data Fig. 9**. Diploid
822 datasets used in this analysis include MS4221 CENP-A ChIP-seq and WGS data (SRX246078,
823 SRX246081) and IMS13q CENP-A ChIP-seq and WGS data (SRX246077, SRX246080).

824

825 For BWA-MEM mapping, CENP-A ChIP-seq and WGS data were aligned to the CHM13 whole-genome
826 assembly¹¹ containing the contiguous chromosome 8 with the following parameters: `bwa mem -k 50 -c`
827 `1000000 {index} {read1.fastq.gz}` for single-end data, and `bwa mem -k 50 -c 1000000 {index}`
828 `{read1.fastq.gz} {read2.fastq.gz}` for paired-end data. The resulting SAM files were filtered using
829 SAMtools⁵⁹ with FLAG score 2308 to prevent multi-mapping of reads. With this filter, reads mapping to
830 more than one location are randomly assigned a single mapping location, thereby preventing mapping
831 biases in highly identical regions. The ChIP-seq data were normalized with deepTools⁸⁸ bamCompare
832 with the following parameters: `bamCompare -b1 {ChIP.bam} -b2 {WGS.bam} --operation ratio --binSize`
833 `1000 -o {out.bw}`. The resulting bigWig file was visualized on the UCSC Genome Browser using the
834 CHM13 chromosome 8 assembly as an assembly hub.

835

836 For the k-mer-based mapping, the initial BWA-MEM alignment was used to identify reads specific to the
837 chromosome 8 centromeric region (chr8:43600000-47200000). K-mers ($k = 50$) were identified from
838 each chromosome 8 centromere-specific dataset using Jellyfish (v2.3.0) and mapped back onto reads
839 and chromosome 8 centromere assembly allowing for no mismatches. Approximately 93-98% of all k-
840 mers identified in the reads were also found within the D8Z2 HOR array. Each k-mer from the read data
841 was then placed once at random between all sites in the HOR array that had a perfect match to that k-
842 mer. These data were then visualized using a histogram with a bin width of 500 in R (R core team,
843 2020).

844

845 **Phylogenetic analysis**

846 To assess the phylogenetic relationship between α -satellite repeats, we first masked every non- α -
847 satellite repeat in the human and NHP centromere assemblies using RepeatMasker⁸⁹ (v4.1.0). Then,
848 we subjected the masked assemblies to StringDecomposer⁵⁴ using a set of 11 α -satellite monomers
849 derived from a chromosome 8 11-mer HOR unit (described in the “Analysis of α -satellite organization
850 subsection” above). This tool identifies the location of α -satellite monomers in the assemblies, and we
851 used this to extract the α -satellite monomers from the HOR/dimeric array and monomeric regions into
852 multi-FASTA files. We ultimately extracted 12,989, 8,132, 12,224, 25,334, and 63,527 α -satellite
853 monomers from the HOR/dimeric array in human, chimpanzee (H1), chimpanzee (H2), orangutan, and
854 macaque, respectively, and 2,879, 3,781, 3,351, 1,573, and 8,127 monomers from the monomeric
855 regions in human, chimpanzee (H1), chimpanzee (H2), orangutan and macaque, respectively. We
856 randomly selected 100 and 50 α -satellite monomers from the HOR/dimeric array and monomeric
857 regions and aligned them with MAFFT^{90,91} (v7.453). We used IQ-TREE⁹² to reconstruct the maximum-
858 likelihood phylogeny with model selection and 1000 bootstraps. The resulting tree file was visualized in
859 iTOL⁹³.

860

861 To estimate sequence divergence along the pericentromeric regions, we first mapped each NHP
862 centromere assembly to the CHM13 centromere assembly using minimap2⁵⁶ (v2.17-r941) with the
863 following parameters: `-ax asm20 --eqx -Y -t 8 -r 500000`. Then, we generated a BED file of 10 kbp
864 windows located within the CHM13 centromere assembly. We used the BED file to subset the BAM file,
865 which was subsequently converted into a set of FASTA files. FASTA files contained at least 5 kbp of
866 orthologous sequences from one or more NHP centromere assemblies. Pairs of human and NHP
867 orthologous sequences were realigned using MAFFT (v7.453) and the following command: `mafft --`
868 `maxiterate 1000 --localpair`. Sequence divergence was estimated using the Tamura-Nei substitution
869 model⁹⁴, which accounts for recurrent mutations and differences between transversions and transitions
870 as well as within transitions. Mutation rate per segment was estimated using Kimura’s model of neutral

871 evolution⁹⁵. In brief, we modeled the estimated divergence (D) is a result of between-species
872 substitutions and within-species polymorphisms; i.e.,

$$D = 2\mu t + 4Ne\mu,$$

873
874
875
876 where Ne is the ancestral human effective population size, t is the divergence time for a given human–
877 NHP pair, and μ is the mutation rate. We assumed a generation time of [20, 29] years and the following
878 divergence times: human–macaque = [23e6, 25e6] years, human–orangutan = [12e6, 14e6] years,
879 human–chimpanzee = [4e6, 6e6] years. To convert the genetic unit to a physical unit, our computation
880 also assumes Ne=10,000 and uniformly drawn values for the generation and divergence times.

881

882 **DATA AVAILABILITY**

883 The complete CHM13 chromosome 8 sequence and all CHM13 ONT data, including raw signal files
884 (FAST5), base calls (FASTQ), and alignments (BAM/CRAM), are available at <https://github.com/nanopore-wgs-consortium/chm13>. In addition, the chromosome 8 sequence, CHM13 ONT FAST5
885 data, and CHM13 Iso-Seq data are accessioned under NCBI BioProject PRJNA559484. CHM13
886 PacBio HiFi data are accessioned under NCBI SRA SRX7897688, SRX7897687, SRX7897686, and
887 SRX7897685. CHM13 Strand-seq data aligned to the CHM13 chromosome 8 assembly are accessible
888 at doi:10.5281/zenodo.3998125. CHM13 BACs used in this study are listed in **Extended Data Table 8**
889 with their corresponding GenBank accession numbers. Two human PacBio Iso-Seq datasets from fetal
890 brain and testis are accessioned under NCBI BioProject PRJNA659539. The chimpanzee, orangutan,
891 and macaque ONT FAST5 and PacBio HiFi data are accessioned under NCBI BioProject
892 PRJNA659034.
893

894

895 **CODE AVAILABILITY**

896 Custom code for the SUNK-based assembly method is available at <https://github.com/glogsdon1/sunk-based-assembly>. All other code is publicly available.
897

898

899 **ACKNOWLEDGMENTS**

900 We thank S. Goodwin (CSHL) for sequence data generation; M. Jain (UCSC) and D. Miller (UW) for re-
901 base-calling sequence data; R. Tindell, H. Visse, A. Tornabene, and G. Ellis (UW) for technical
902 assistance; Z. Zhao for computational assistance; F.F. Dastvan (UW) for instrumentation; D. Gordon
903 (UW) for accessioning BACs; G. Bouffard (NHGRI) for accessioning ONT FAST5 data; J.G. Underwood
904 (FHRC/PacBio) for helpful discussions; and T. Brown (UW) for assistance in editing this manuscript.
905 We acknowledge experimental support from the W. M. Keck Microscopy Center (UW) and the
906 computational resources of the NIH HPC Biowulf cluster (<https://hpc.nih.gov>). This research was
907 supported, in part, by funding from the National Institutes of Health (NIH), HG002385 and HG010169
908 (EEE); National Institute of General Medical Sciences (NIGMS), F32 GM134558 (GAL); Intramural
909 Research Program of the National Human Genome Research Institute at NIH (SK, AMP, AR); National
910 Library of Medicine Big Data Training Grant for Genomics and Neuroscience 5T32LM012419-04
911 (MRV); NIH/NHGRI Pathway to Independence Award K99HG011041 (PH); NIH/NHGRI R21
912 1R21HG010548-01 and NIH/NHGRI U01 1U01HG010971 (KHM); and the Intramural Research
913 Program of the NIH, National Cancer Institute, Center for Cancer Research, USA (VL). EEE is an
914 investigator of the Howard Hughes Medical Institute.
915

916 **AUTHOR CONTRIBUTIONS**

917 GAL and EEE conceived the project; GAL, KH, KMM, AML, CB, MS generated long-read sequencing
918 data; GAL, MRV, PH, YM, SK, SN, PCD, AR, TD, DP, AM, AVB, MK, TAG-L, CJ, SCM, KHM, and AMP
919 analyzed sequencing data, created genome assemblies, and performed QC analyses; GAL, MRV, SK,
920 AMP and SN finalized the chromosome 8 assembly; GAL, SK, SN, AM, AVB, and KHM assessed the
921 assembly of the centromere; MAL generated pulsed-field gel Southern blots; GAL, LM, and MV
922 generated microscopy data; LGD generated and analyzed droplet digital PCR data; US provided the
923 CHM13 cell line; JGL and VL supervised experimental analyses; GAL, MRV, and EEE developed
924 figures; and GAL and EEE drafted the manuscript.

925

926 **COMPETING INTERESTS**

927 The other authors declare no competing financial interests.

928 **REFERENCES**

929

- 930 1. International Human Genome Project Consortium. Initial sequencing and analysis of the human
931 genome. *Nature* **409**, 860–921 (2001).
- 932 2. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
- 933 3. Alkan, C. *et al.* Genome-wide characterization of centromeric satellites from multiple mammalian
934 genomes. *Genome Res.* **21**, 137–145 (2011).
- 935 4. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the
936 human genome. *Nature* **431**, 931–945 (2004).
- 937 5. Willard, H. F. Chromosome-specific organization of human alpha satellite DNA. *American Journal*
938 *of Human Genetics* **37**, 524 (1985).
- 939 6. Wayne, J. S. & Willard, H. F. Nucleotide sequence heterogeneity of alpha satellite repetitive DNA: a
940 survey of alphoid sequences from different human chromosomes. *Nucleic Acids Res.* **15**, 7549–
941 7569 (1987).
- 942 7. Rudd, M. K., Schueler, M. G. & Willard, H. F. Sequence organization and functional annotation of
943 human centromeres. *Cold Spring Harb. Symp. Quant. Biol.* **68**, 141–149 (2003).
- 944 8. McNulty, S. M. & Sullivan, B. A. Alpha satellite DNA biology: finding function in the recesses of the
945 genome. *Chromosome Res.* **26**, 115–138 (2018).
- 946 9. Miga, K. H. *et al.* Centromere reference models for human chromosomes X and Y satellite arrays.
947 *Genome Res.* **24**, 697–707 (2014).
- 948 10. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human
949 genomes. *Nature* **491**, 56–65 (2012).
- 950 11. Nurk, S. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants
951 from high-fidelity long reads. *Genome Res.* gr.263566.120 (2020) doi:10.1101/gr.263566.120.
- 952 12. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly
953 with phased assembly graphs. *arXiv:2008.01237 [q-bio]* (2020).
- 954 13. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its
955 applications. *Nature Reviews Genetics* 1–18 (2020) doi:10.1038/s41576-020-0236-x.

- 956 14. Steinberg, K. M. *et al.* Single haplotype assembly of the human genome from a hydatidiform mole.
957 *Genome Res* **24**, 2066–2076 (2014).
- 958 15. Chaisson, M. J. P. *et al.* Resolving the complexity of the human genome using single-molecule
959 sequencing. *Nature* **517**, 608–611 (2015).
- 960 16. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nature*
961 **585**, 79–84 (2020).
- 962 17. Ge, Y., Wagner, M. J., Siciliano, M. & Wells, D. E. Sequence, higher order repeat structure, and
963 long-range organization of alpha satellite DNA specific to human chromosome 8. *Genomics* **13**,
964 585–593 (1992).
- 965 18. Hollox, E. J., Armour, J. a. L. & Barber, J. C. K. Extensive normal copy number variation of a beta-
966 defensin antimicrobial-gene cluster. *Am. J. Hum. Genet.* **73**, 591–600 (2003).
- 967 19. Hollox, E. J. *et al.* Psoriasis is associated with increased beta-defensin genomic copy number. *Nat.*
968 *Genet.* **40**, 23–25 (2008).
- 969 20. Mohajeri, K. *et al.* Interchromosomal core duplicons drive both evolutionary instability and disease
970 susceptibility of the Chromosome 8p23.1 region. *Genome Res* **26**, 1453–1467 (2016).
- 971 21. Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**,
972 641–646 (2010).
- 973 22. Falconer, E. & Lansdorp, P. M. Strand-seq: a unifying tool for studies of chromosome segregation.
974 *Semin. Cell Dev. Biol.* **24**, 643–652 (2013).
- 975 23. Sanders, A. D., Falconer, E., Hills, M., Spierings, D. C. J. & Lansdorp, P. M. Single-cell template
976 strand sequencing by Strand-seq enables the characterization of individual homologs. *Nat Protoc*
977 **12**, 1151–1176 (2017).
- 978 24. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness,
979 and phasing assessment for genome assemblies. *bioRxiv* 2020.03.15.992941 (2020)
980 doi:10.1101/2020.03.15.992941.
- 981 25. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat.*
982 *Methods* **14**, 407–410 (2017).

- 983 26. Devriendt, K. *et al.* Delineation of the critical deletion region for congenital heart defects, on
984 chromosome 8p23.1. *Am J Hum Genet* **64**, 1119–1126 (1999).
- 985 27. Giglio, S. *et al.* Heterozygous submicroscopic inversions involving olfactory receptor-gene clusters
986 mediate the recurrent t(4;8)(p16;p23) translocation. *Am. J. Hum. Genet.* **71**, 276–285 (2002).
- 987 28. Cantsilieris, S. & White, S. J. Correlating multiallelic copy number polymorphisms with disease
988 susceptibility. *Hum. Mutat.* **34**, 1–13 (2013).
- 989 29. Tyson, C. *et al.* Expansion of a 12-kb VNTR containing the REXO1L1 gene cluster underlies the
990 microscopically visible euchromatic variant of 8q21.2. *European Journal of Human Genetics* **22**,
991 458–463 (2014).
- 992 30. Warburton, P. E. *et al.* Analysis of the largest tandemly repeated DNA families in the human
993 genome. *BMC Genomics* **9**, 533 (2008).
- 994 31. Hasson, D. *et al.* Formation of novel CENP-A domains on tandem repetitive DNA and across
995 chromosome breakpoints on human chromosome 8q21 neocentromeres. *Chromosoma* **120**, 621–
996 632 (2011).
- 997 32. Hasson, D. *et al.* The octamer is the major form of CENP-A nucleosomes at human centromeres.
998 *Nat. Struct. Mol. Biol.* **20**, 687–695 (2013).
- 999 33. Alkan, C. *et al.* Organization and evolution of primate centromeric DNA from whole-genome
1000 shotgun sequence data. *PLoS Comput. Biol.* **3**, 1807–1818 (2007).
- 1001 34. Cacheux, L., Ponger, L., Gerbault-Seureau, M., Richard, F. A. & Escudé, C. Diversity and
1002 distribution of alpha satellite DNA in the genome of an Old World monkey: *Cercopithecus solatus*.
1003 *BMC Genomics* **17**, 916 (2016).
- 1004 35. Jain, M. *et al.* Linear assembly of a human centromere on the Y chromosome. *Nat. Biotechnol.* **36**,
1005 321–323 (2018).
- 1006 36. Warburton, P. E. *et al.* Immunolocalization of CENP-A suggests a distinct nucleosome structure at
1007 the inner kinetochore plate of active centromeres. *Current Biology* **7**, 901–904 (1997).
- 1008 37. Vafa, O. & Sullivan, K. F. Chromatin containing CENP-A and α -satellite DNA is a major component
1009 of the inner kinetochore plate. *Current Biology* **7**, 897–900 (1997).

- 1010 38. Smith, G. P. Evolution of repeated DNA sequences by unequal crossover. *Science* **191**, 528–535
1011 (1976).
- 1012 39. Shepelev, V. A., Alexandrov, A. A., Yurov, Y. B. & Alexandrov, I. A. The evolutionary origin of man
1013 can be traced in the layers of defunct ancestral alpha satellites flanking the active centromeres of
1014 human chromosomes. *PLOS Genetics* **5**, e1000641 (2009).
- 1015 40. Alexandrov, I., Kazakov, A., Tumeneva, I., Shepelev, V. & Yurov, Y. Alpha-satellite DNA of
1016 primates: old and new families. *Chromosoma* **110**, 253–266 (2001).
- 1017 41. Salser, W. *et al.* Investigation of the organization of mammalian chromosomes at the DNA
1018 sequence level. *Fed. Proc.* **35**, 23–35 (1976).
- 1019 42. Marshall, O. J., Chueh, A. C., Wong, L. H. & Choo, K. H. A. Neocentromeres: new insights into
1020 centromere structure, disease development, and karyotype evolution. *American Journal of Human*
1021 *Genetics* **82**, 261 (2008).
- 1022 43. Warburton, P. E. *et al.* Molecular cytogenetic analysis of eight inversion duplications of human
1023 chromosome 13q that each contain a neocentromere. *Am J Hum Genet* **66**, 1794–1806 (2000).
- 1024 44. Koga, A. *et al.* Evolutionary origin of higher-order repeat structure in alpha-satellite DNA of primate
1025 centromeres. *DNA Res.* **21**, 407–415 (2014).
- 1026 45. Alexandrov, I. A., Mitkevich, S. P. & Yurov, Y. B. The phylogeny of human chromosome specific
1027 alpha satellites. *Chromosoma* **96**, 443–453 (1988).
- 1028 46. Singer, D. & Donehower, L. Highly repeated DNA of the baboon: Organization of sequences
1029 homologous to highly repeated DNA of the African green monkey. *Journal of Molecular Biology*
1030 **134**, 835–842 (1979).
- 1031 47. Plohl, M., Meštrović, N. & Mravinac, B. Centromere identity from the DNA point of view.
1032 *Chromosoma* **123**, 313–325 (2014).
- 1033 48. Henikoff, S., Ahmad, K. & Malik, H. S. The centromere paradox: stable inheritance with rapidly
1034 evolving DNA. *Science* **293**, 1098–1102 (2001).
- 1035 49. Malik, H. S. & Henikoff, S. Adaptive evolution of Cid, a centromere-specific histone in *Drosophila*.
1036 *Genetics* **157**, 1293–1298 (2001).

- 1037 50. Fishman, L. & Saunders, A. Centromere-associated female meiotic drive entails male fitness costs
1038 in monkeyflowers. *Science* **322**, 1559–1562 (2008).
- 1039 51. Vollger, M. R. *et al.* Improved assembly and variant detection of a haploid human genome using
1040 single-molecule, high-fidelity long reads. *Ann. Hum. Genet.* (2019) doi:10.1111/ahg.12364.
- 1041 52. Huddleston, J. *et al.* Reconstructing complex regions of genomes using long-read sequencing
1042 technology. *Genome Res* **24**, 688–696 (2014).
- 1043 53. Logsdon, G. A. HMW gDNA purification and ONT ultra-long-read data generation. *protocols.io*
1044 (2020) doi:dx.doi.org/10.17504/protocols.io.bchhit36.
- 1045 54. Dvorkina, T., Bzikadze, A. V. & Pevzner, P. A. The string decomposition problem and its
1046 applications to centromere analysis and assembly. *Bioinformatics* **36**, i93–i101 (2020).
- 1047 55. Jain, C. *et al.* Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36**, i111–
1048 i118 (2020).
- 1049 56. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100
1050 (2018).
- 1051 57. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform.
1052 *Bioinformatics* **26**, 589–595 (2010).
- 1053 58. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS
1054 alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
- 1055 59. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079
1056 (2009).
- 1057 60. Porubsky, D. *et al.* breakpointR: an R/Bioconductor package to localize strand state changes in
1058 Strand-seq data. *Bioinformatics* **36**, 1260–1261 (2020).
- 1059 61. Porubsky, D. *et al.* A fully phased accurate assembly of an individual human genome. *bioRxiv*
1060 855049 (2019) doi:10.1101/855049.
- 1061 62. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human
1062 genomes. *Nat Commun* **10**, 1784 (2019).

- 1063 63. Sanders, A. D. *et al.* Characterizing polymorphic inversions in human genomes by single-cell
1064 sequencing. *Genome Res.* (2016) doi:10.1101/gr.201160.115.
- 1065 64. Ghareghani, M. *et al.* Strand-seq enables reliable separation of long reads by chromosome via
1066 expectation maximization. *Bioinformatics* **34**, i115–i123 (2018).
- 1067 65. Mikheenko, A., Bzikadze, A. V., Gurevich, A., Miga, K. H. & Pevzner, P. A. TandemTools: mapping
1068 long reads and assessing/improving assembly quality in extra-long tandem repeats. *Bioinformatics*
1069 **36**, i75–i83 (2020).
- 1070 66. Lee, I. *et al.* Simultaneous profiling of chromatin accessibility and methylation on human cell lines
1071 with nanopore sequencing. *bioRxiv* 504993 (2019) doi:10.1101/504993.
- 1072 67. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnology* **29**, 24–26 (2011).
- 1073 68. Dougherty, M. L. *et al.* Transcriptional fates of human-specific segmental duplications in brain.
1074 *Genome Res.* **28**, 1566–1576 (2018).
- 1075 69. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning
1076 sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
- 1077 70. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project.
1078 *Genome Res.* **22**, 1760–1774 (2012).
- 1079 71. Pertea, M. *et al.* CHES: a new human gene catalog curated from thousands of large-scale RNA
1080 sequencing experiments reveals extensive transcriptional noise. *Genome Biology* **19**, 208 (2018).
- 1081 72. Shumate, A. & Salzberg, S. L. Liftoff: an accurate gene annotation mapping tool. *bioRxiv*
1082 2020.06.24.169680 (2020) doi:10.1101/2020.06.24.169680.
- 1083 73. R Core Team. *R: A language and environment for statistical computing.* (R Foundation for
1084 Statistical Computing, 2020).
- 1085 74. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse
1086 genomes. *Science* **367**, (2020).
- 1087 75. Mafessoni, F. *et al.* A high-coverage Neandertal genome from Chagyrskaya Cave. *PNAS* **117**,
1088 15132–15136 (2020).

- 1089 76. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse
1090 populations. *Nature* **538**, 201–206 (2016).
- 1091 77. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science*
1092 **338**, 222–226 (2012).
- 1093 78. Prado-Martinez, J. *et al.* Great ape genetic diversity and population history. *Nature* **499**, 471–475
1094 (2013).
- 1095 79. Prüfer, K. *et al.* A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**,
1096 655–658 (2017).
- 1097 80. Hach, F. *et al.* mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat. Methods* **7**, 576–
1098 577 (2010).
- 1099 81. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: A short tandem repeat profiler for personal
1100 genomes. *Genome Res* **22**, 1154–1162 (2012).
- 1101 82. Iwata-Otsubo, A. *et al.* Expanded satellite repeats amplify a discrete CENP-A nucleosome
1102 assembly site on chromosomes that drive in female meiosis. *Current Biology* **27**, 2365–2373
1103 (2017).
- 1104 83. Logsdon, G. A. *et al.* Human artificial chromosomes that bypass centromeric DNA. *Cell* **178**, 624-
1105 639.e19 (2019).
- 1106 84. Ventura, M. *et al.* Gorilla genome structural variation reveals evolutionary parallelisms with
1107 chimpanzee. *Genome Res.* gr.124461.111 (2011) doi:10.1101/gr.124461.111.
- 1108 85. Darby, I. A. *In Situ Hybridization Protocols.* (Humana Press, 2000).
- 1109 86. Haaf, T. & Willard, H. F. Chromosome-specific alpha-satellite DNA from the centromere of
1110 chimpanzee chromosome 4. *Chromosoma* **106**, 226–232 (1997).
- 1111 87. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
1112 *arXiv:1303.3997 [q-bio]* (2013).
- 1113 88. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for
1114 exploring deep-sequencing data. *Nucleic Acids Res* **42**, W187–W191 (2014).
- 1115 89. Smit, A. F. A., Hubley, R. & Green, P. *RepeatMasker Open-4.0.* (2013).

- 1116 90. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements
1117 in performance and usability. *Mol Biol Evol* **30**, 772–780 (2013).
- 1118 91. Nakamura, T., Yamada, K. D., Tomii, K. & Katoh, K. Parallelization of MAFFT for large-scale
1119 multiple sequence alignments. *Bioinformatics* **34**, 2490–2492 (2018).
- 1120 92. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective
1121 stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268–274
1122 (2015).
- 1123 93. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and
1124 annotation. *Bioinformatics* **23**, 127–128 (2007).
- 1125 94. Tamura, K. & Nei, M. Estimation of the number of nucleotide substitutions in the control region of
1126 mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512–526 (1993).
- 1127 95. Kimura, M. *The neutral theory of molecular evolution*. (Cambridge University Press, 1983).
1128 doi:10.1017/CBO9780511623486.
1129

FIGURES

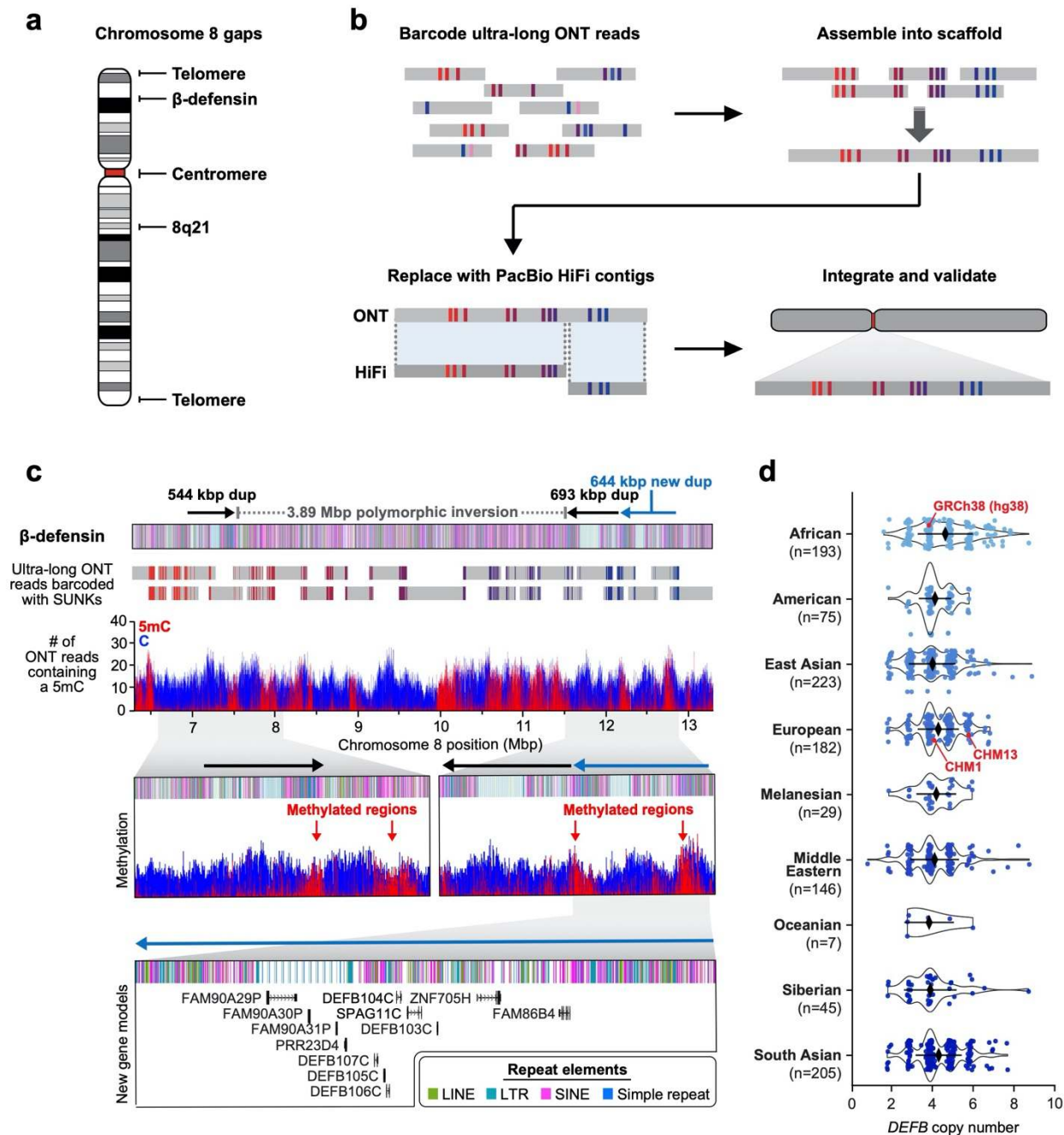


Figure 1. Telomere-to-telomere assembly of human chromosome 8 via a targeted assembly method. **a)** Gaps in the GRCh38 chromosome 8 reference sequence. **b)** Targeted assembly method to resolve complex repeat regions in the human genome. Ultra-long ONT reads (gray) are barcoded with singly unique nucleotide k-mers (SUNKs; colored bars) and assembled into a sequence scaffold. Regions within the scaffold sharing high sequence identity with PacBio HiFi contigs (dark gray) are replaced, thereby improving the base accuracy to >99.99%. The PacBio HiFi assembly is integrated into an assembly of chromosome 8 from the CHM13 genome¹¹ and subsequently validated with orthogonal technologies. **c)** Sequence, structure, methylation status, and genetic composition of the CHM13 β -defensin locus. The CHM13 locus contains three segmental duplications (SDs) (dups) located at chr8:7098892-7643091, chr8:11528114-12220905, and chr8:12233870-12878079 in the

assembly. A 3,885,023 bp inversion (located at chr8:7643092-11528113) separates the first and second duplication. Although two SDs had been previously reported²⁰, the other duplication (light blue) is newly resolved. Resolution of the entire locus was achieved via assembly of 26 ultra-long ONT reads (gray) barcoded with SUNKs (colored bars). The methylation status of the region was determined from mapped ultra-long ONT reads using Nanopolish²⁵. 5-methylcytosine (5mC) is indicated in red and unmethylated cytosine is indicated in blue. Iso-Seq data reveal that the new duplication contains twelve new protein-coding genes, five of which are *DEFB* genes (**Extended Data Fig. 15** shows a schematic of all *DEFA*- and *DEFB*-related genes across the β -defensin locus). **d)** Copy number of the *DEFB* genes [chr8:7783837–7929198 in GRCh38 (hg38)] throughout the human population. CHM13 has six copies of *DEFB* genes, one set per SD per haplotype, while CHM1 and GRCh38 only have four copies (red data points).

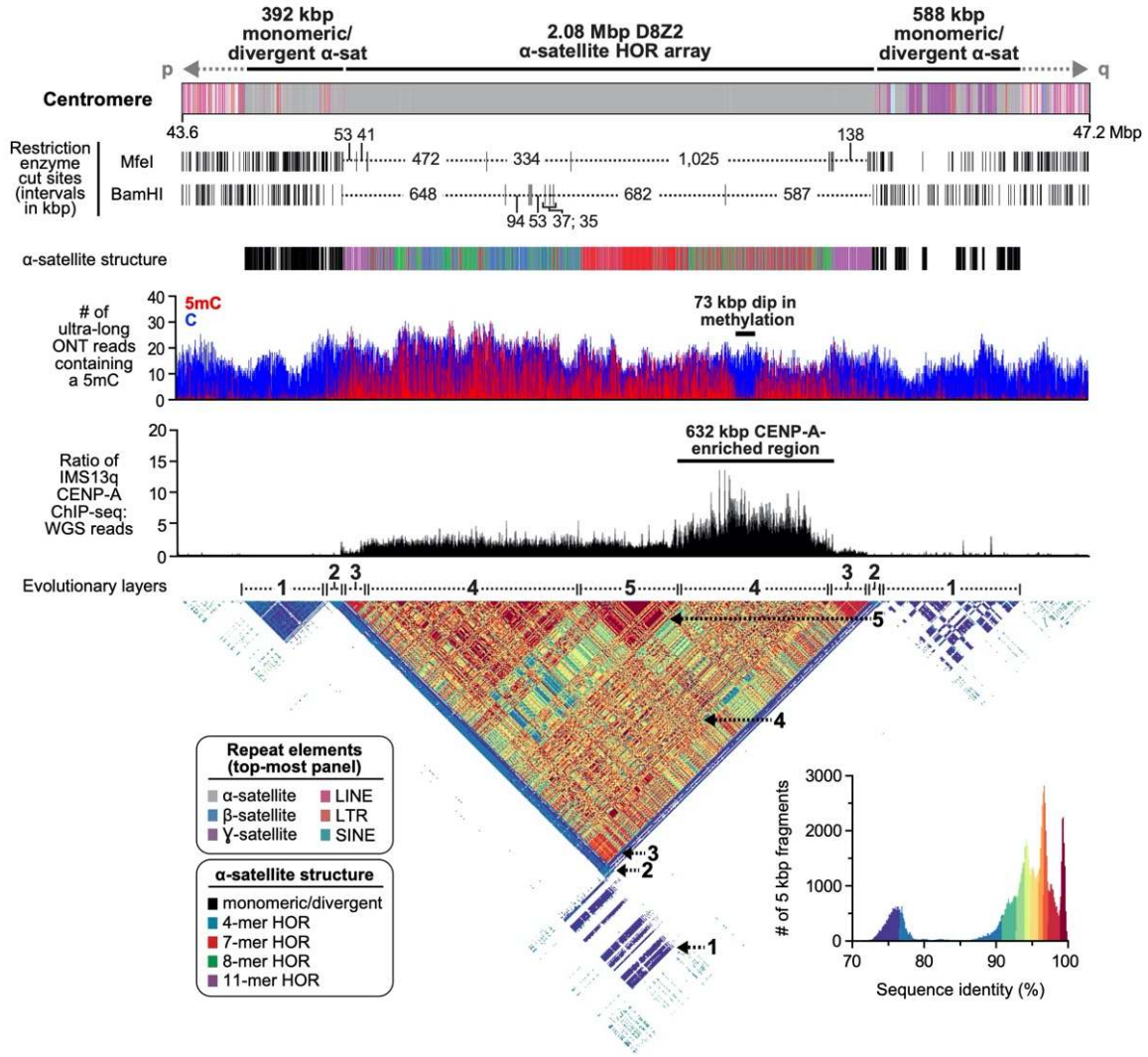
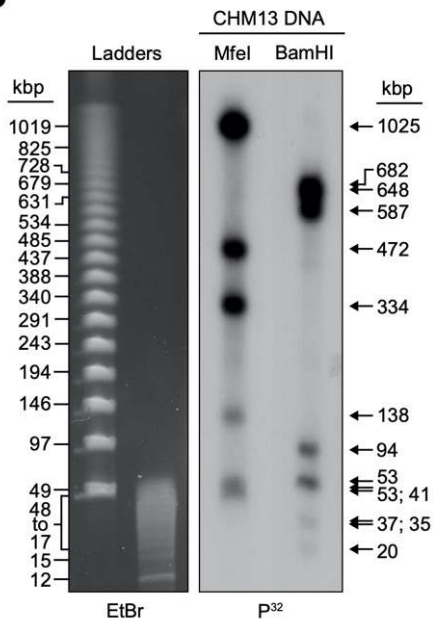
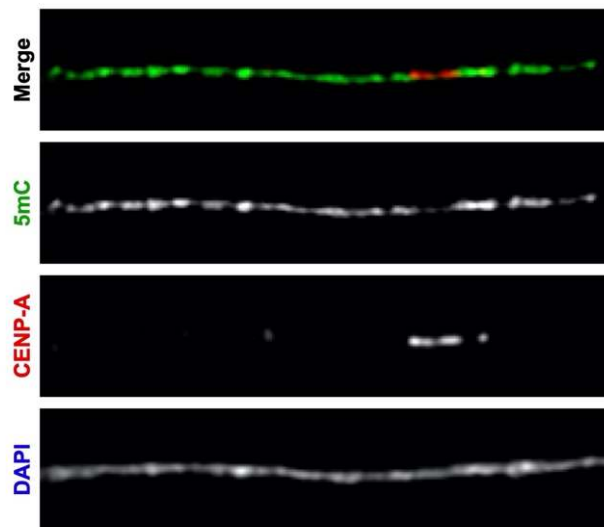
a**b****c**

Figure 2. Sequence, structure, and epigenetic map of the chromosome 8 centromeric region.

a) Schematic showing the composition of the CHM13 chromosome 8 centromere. The centromeric region is comprised of a 2.08 Mbp D8Z2 α -satellite HOR array flanked by regions of monomeric and/or divergent α -satellite interspersed with retrotransposons, β -satellite, and γ -satellite. The predicted restriction digest pattern is shown and supported by the pulsed-field gel (PFG) Southern blot in **Panel b**. The D8Z2 α -satellite HOR array is primarily composed of four types of higher-order repeats (HORs; see **Extended Data Fig. 8, Methods** for details) and is heavily methylated except for a 73 kbp hypomethylation region. Mapping of normalized CENP-A ChIP-Seq data from a diploid human genome known as IMS13q³¹ reveals that centromeric chromatin is primarily located within a 632 kbp region encompassing the hypomethylated region (**Extended Data Fig. 9** includes another CENP-A ChIP-seq dataset and details). A pairwise sequence identity map across the centromeric region indicates that the centromere is composed of five distinct evolutionary layers (indicated with dashed arrows). **b)** PFG Southern blot of CHM13 DNA confirms the structure and organization of the chromosome 8 centromeric HOR array indicated in **Panel a**. Left: EtBr staining; Right: P³²-labeled chromosome 8 α -satellite-specific probe. **c)** Representative images of a CHM13 chromatin fiber showing that CENP-A is enriched in an unmethylated region. Bar = 1 micron.

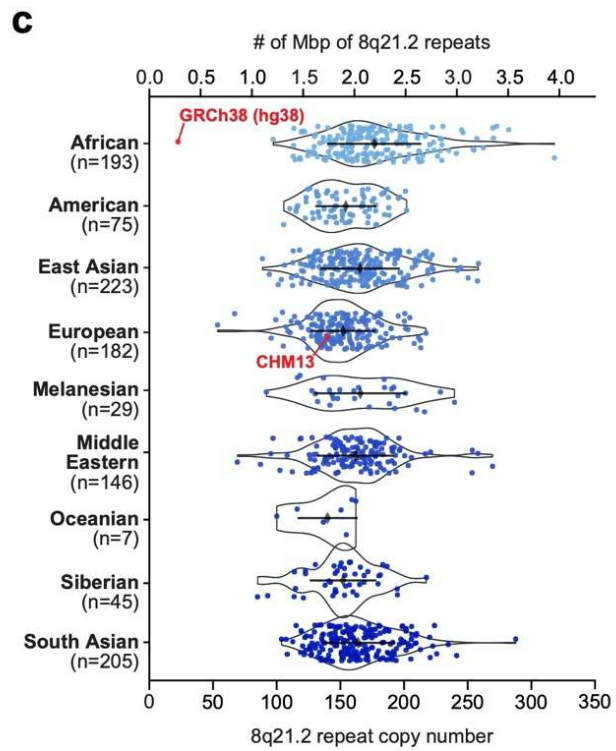
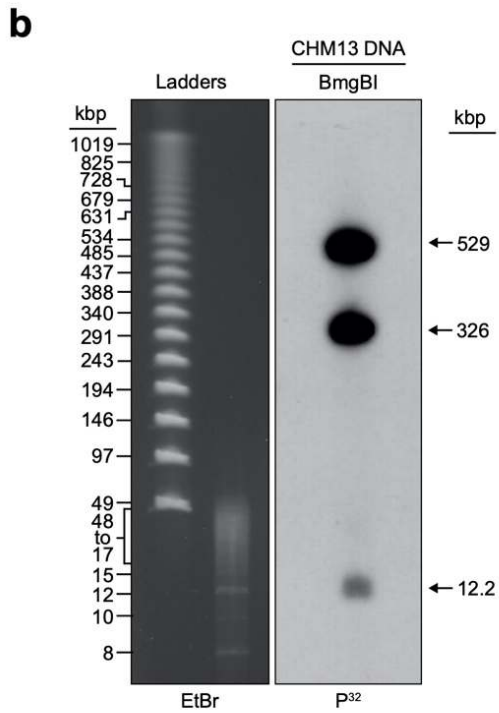
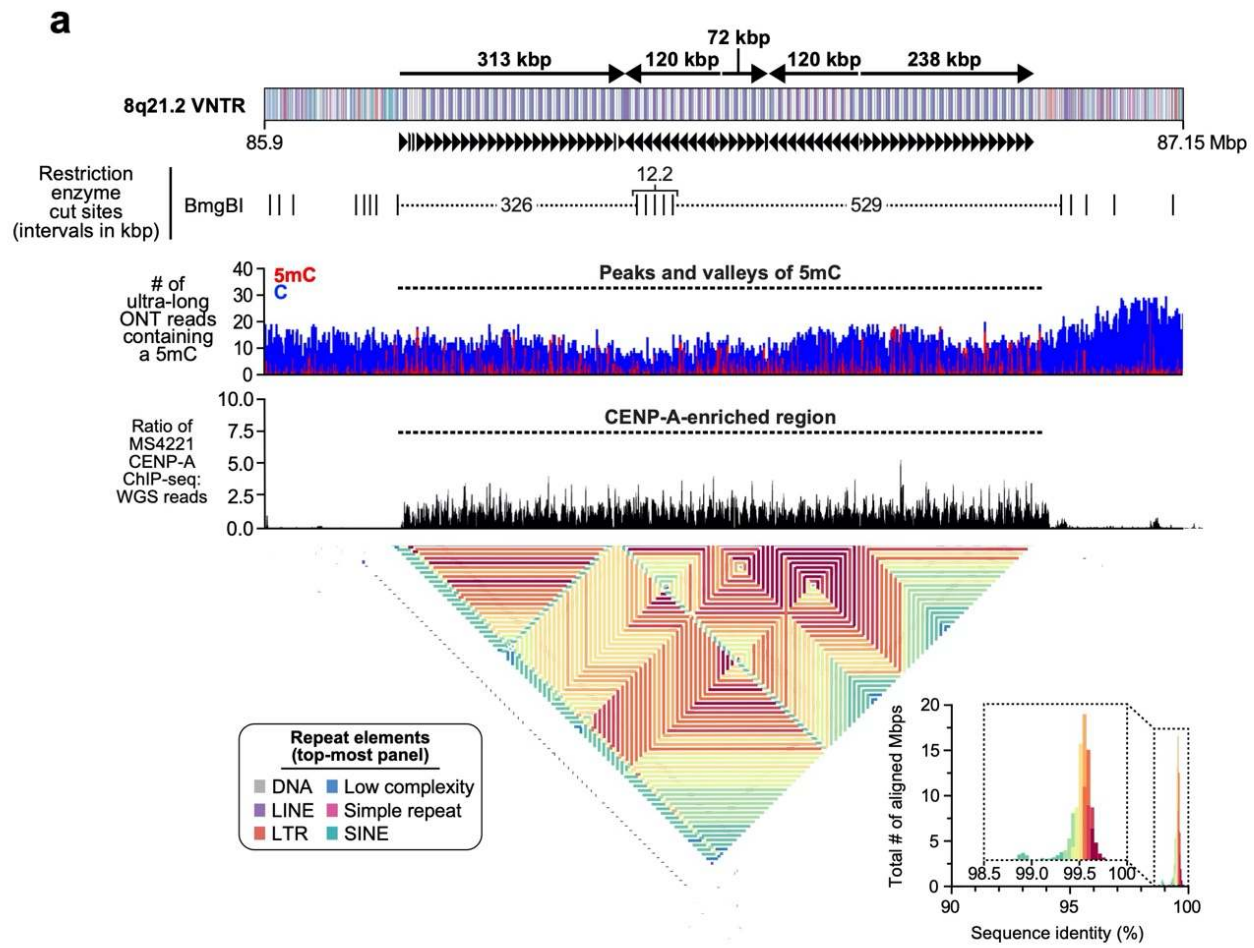


Figure 3. Sequence, structure, and epigenetic map of the neocentromeric chromosome 8q21.2 VNTR. **a)** Schematic showing the composition of the CHM13 8q21.2 VNTR. This VNTR is comprised of 67 full and 7 partial 12.192 kbp repeats that span 863 kbp in total. The predicted restriction digest pattern is indicated. Each repeat is methylated within a 3 kbp region and hypomethylated within the rest of the sequence. Mapping of CENP-A CHIP-seq data from the chromosome 8 neocentric cell line known as MS4221^{31,32} (**Methods**) reveals that centromeric chromatin is primarily located on the hypomethylated portion of the repeat. A pairwise sequence identity map across the region indicates a mirrored symmetry within a single layer, consistent with the evolutionarily young status of the tandem repeat. **b)** PFG Southern blot of CHM13 DNA digested with BmgBI confirms the size and organization of the chromosome 8q21.2 VNTR. Left: EtBr staining; Right: P³²-labeled chromosome 8q21.2-specific probe. **c)** Copy number of the 8q21 repeat [chr8:85792897–85805090 in GRCh38] throughout the human population. CHM13 is estimated to have 144 total copies of the 8q21 repeat, or 72 copies per haplotype, while GRCh38 only has 26 copies (red data points).

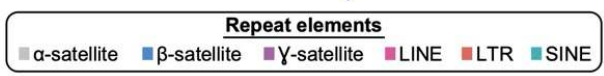
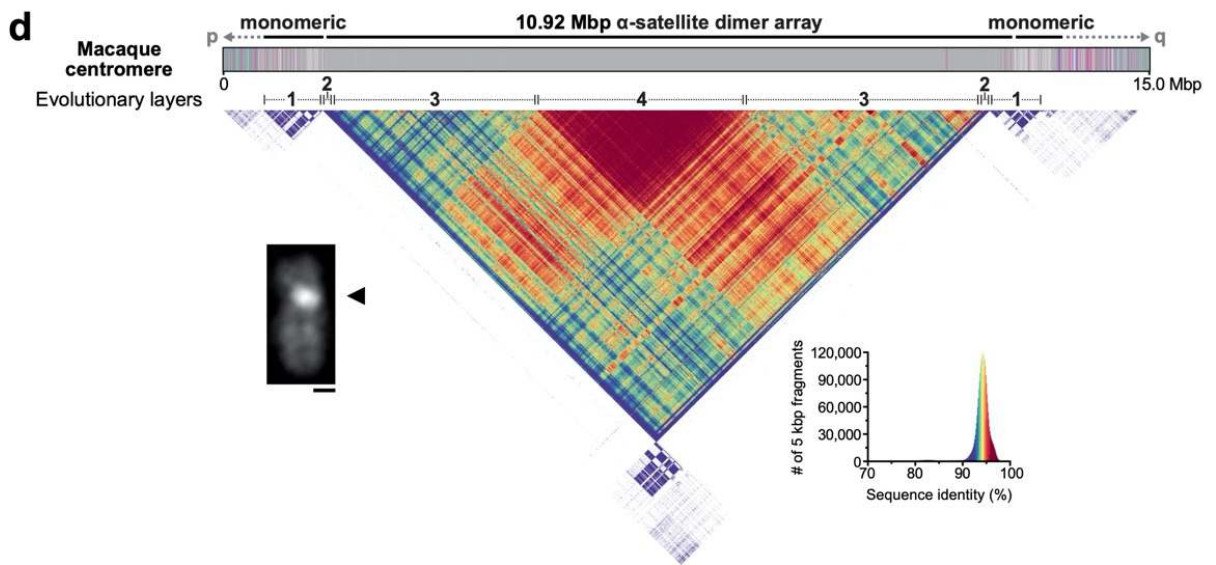
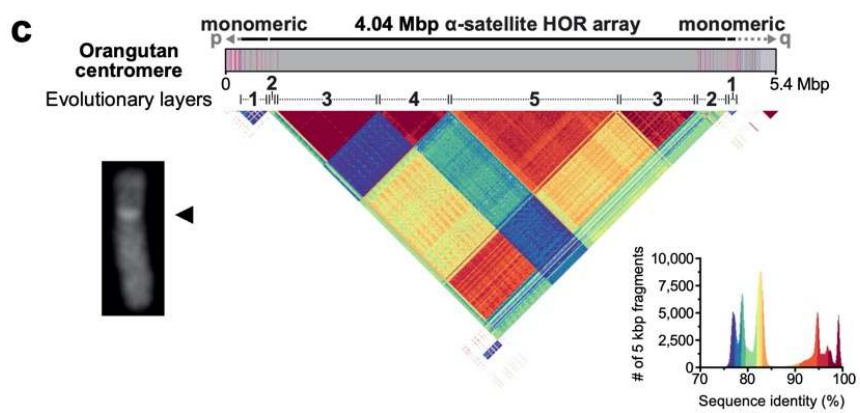
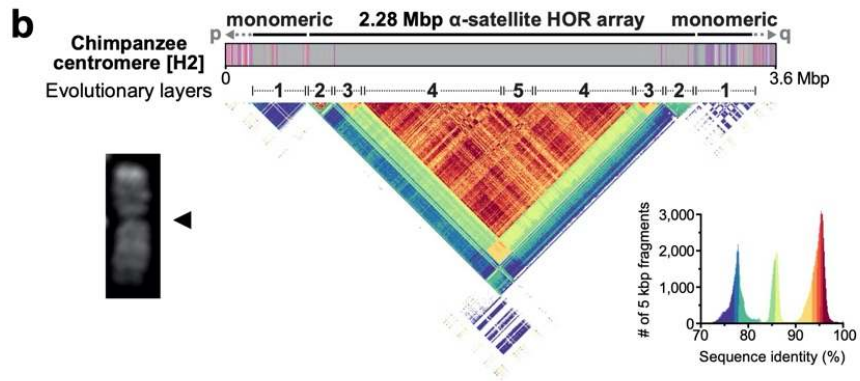
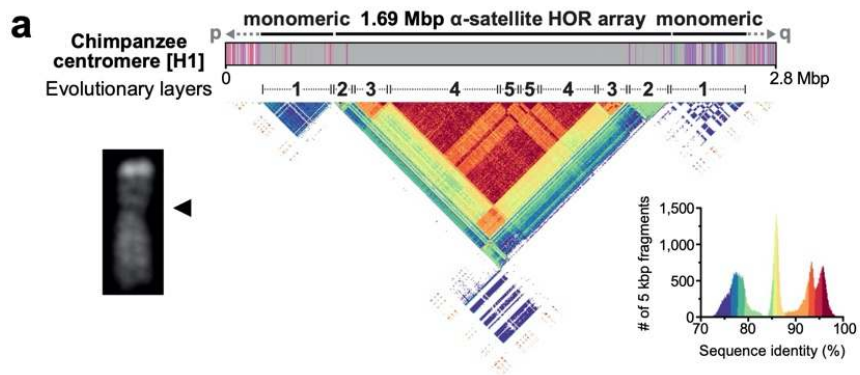


Figure 4. Sequence and structure of the chimpanzee, orangutan, and macaque chromosome 8 centromeres. a-d) Structure and sequence identity of the chimpanzee H1 (**Panel a**), chimpanzee H2 (**Panel b**), orangutan (**Panel c**), and macaque (**Panel d**) chromosome 8 centromeres. Each centromere has a mirrored organization consisting of either four or five distinct evolutionary layers. The size of each centromeric region is consistent with microscopic analyses, showing increasingly bright DAPI staining with increasing centromere size. Bar = 1 micron.

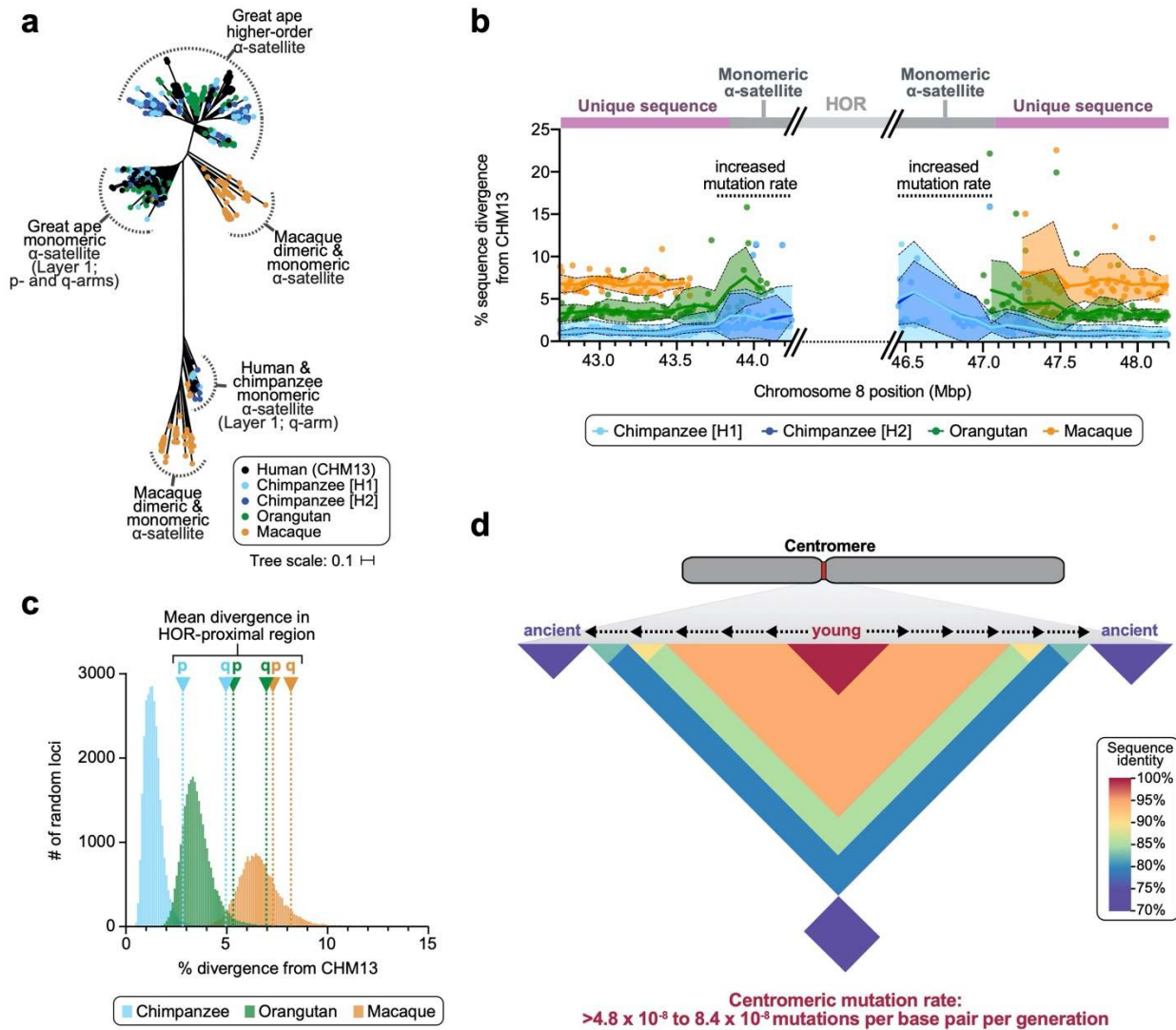


Figure 5. Evolution of the chromosome 8 centromere. **a)** Phylogenetic tree of human, chimpanzee, orangutan, and macaque α -satellites from the HOR and monomeric regions of the chromosome 8 centromere. A portion of the human and chimpanzee monomeric α -satellite is evolutionarily closer to the macaque α -satellite (bottom of the tree; see **Extended Data Fig. 13** for bootstrapping annotations). **b)** Plot showing the sequence divergence between the CHM13 and nonhuman primates (NHPs) in the regions flanking the chromosome 8 α -satellite HOR array. The mean and standard deviation (bold line and shaded region) are calculated over a sliding window of 200 kbp with a 100 kbp overlap. Individual data points from 10 kbp pairwise sequence alignments are shown. **c)** Histogram of the sequence divergence between CHM13 and chimpanzee, orangutan, or macaque at thousands of random 10 kbp loci. **d)** Model of centromere evolution. Centromeric α -satellite HORs evolve in the center of the array via unequal crossing over and homogenization, pushing older, more ancient HORs to the edges, consistent with hypotheses previously put forth^{38,39,41}. The centromeric mutation rate is estimated to be at least 4.8 to 8.4×10^{-8} mutations per base pair per generation, which is 2.2 to 3.8 higher than the mean mutation rate measured from nearly 20,000 random loci.

Figures

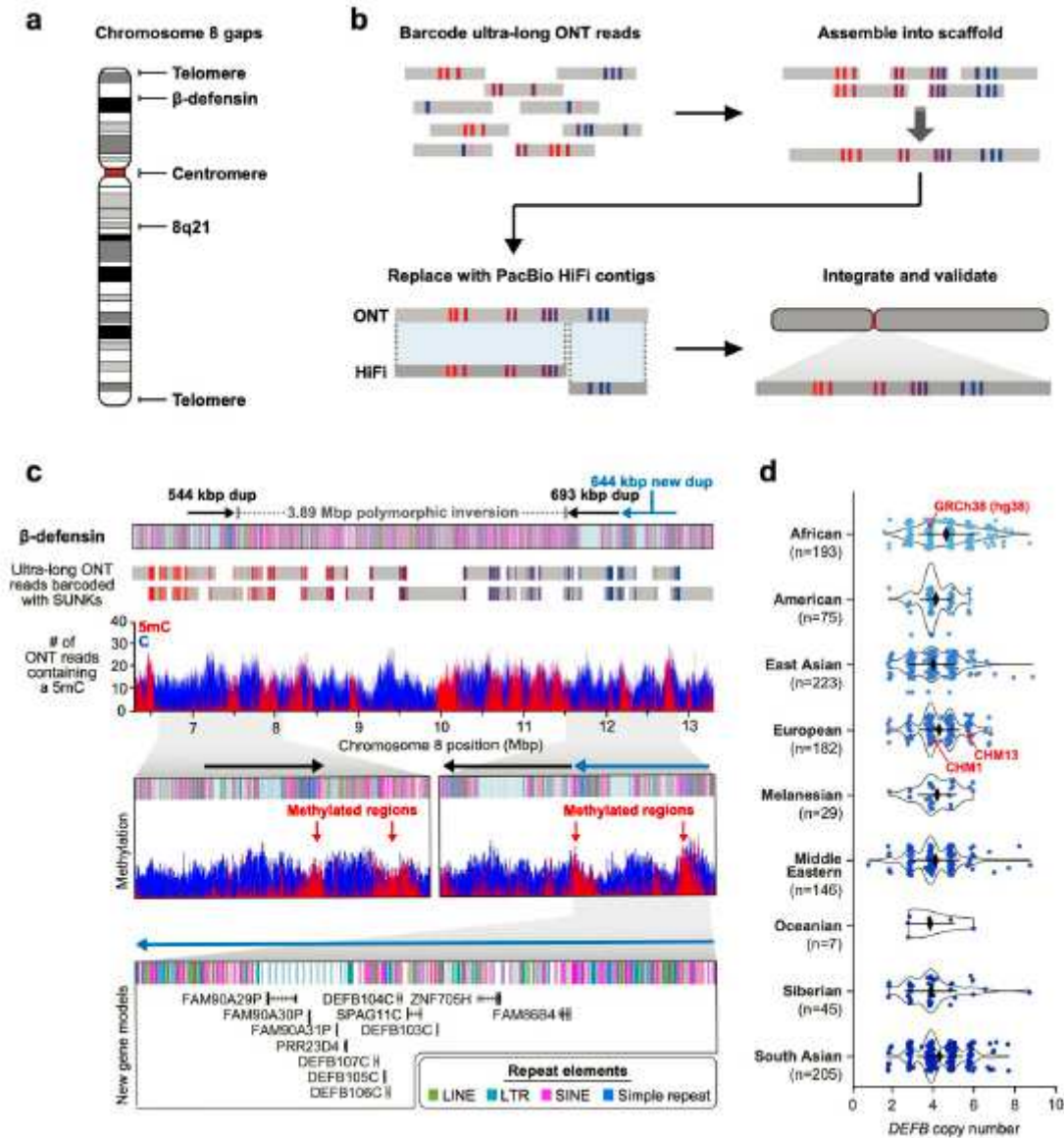


Figure 1

Telomere-to-telomere assembly of human chromosome 8 via a targeted assembly method. a) Gaps in the GRCh38 chromosome 8 reference sequence. b) Targeted assembly method to resolve complex repeat regions in the human genome. Ultra-long ONT reads (gray) are barcoded with singly unique nucleotide k-mers (SUNKs; colored bars) and assembled into a sequence scaffold. Regions within the scaffold sharing high sequence identity with PacBio HiFi contigs (dark gray) are replaced, thereby improving the base accuracy to >99.99%. The PacBio HiFi assembly is integrated into an assembly of chromosome 8 from the CHM13 genome11 and subsequently validated with orthogonal technologies. c) Sequence, structure, methylation status, and genetic composition of the CHM13 β -defensin locus. The CHM13 locus contains three segmental duplications (SDs) (dups) located at chr8:7098892-7643091, chr8:11528114-12220905, and chr8:12233870-12878079 in the assembly. A 3,885,023 bp inversion (located at chr8:7643092-

11528113) separates the first and second duplication. Although two SDs had been previously reported²⁰, the other duplication (light blue) is newly resolved. Resolution of the entire locus was achieved via assembly of 26 ultra-long ONT reads (gray) barcoded with SUNKs (colored bars). The methylation status of the region was determined from mapped ultra-long ONT reads using Nanopolish²⁵. 5-methylcytosine (5mC) is indicated in red and unmethylated cytosine is indicated in blue. Iso-Seq data reveal that the new duplication contains twelve new protein-coding genes, five of which are DEFB genes (Extended Data Fig. 15 shows a schematic of all DEFA- and DEFB-related genes across the β -defensin locus). d) Copy number of the DEFB genes [chr8:7783837–7929198 in GRCh38 (hg38)] throughout the human population. CHM13 has six copies of DEFB genes, one set per SD per haplotype, while CHM1 and GRCh38 only have four copies (red data points).

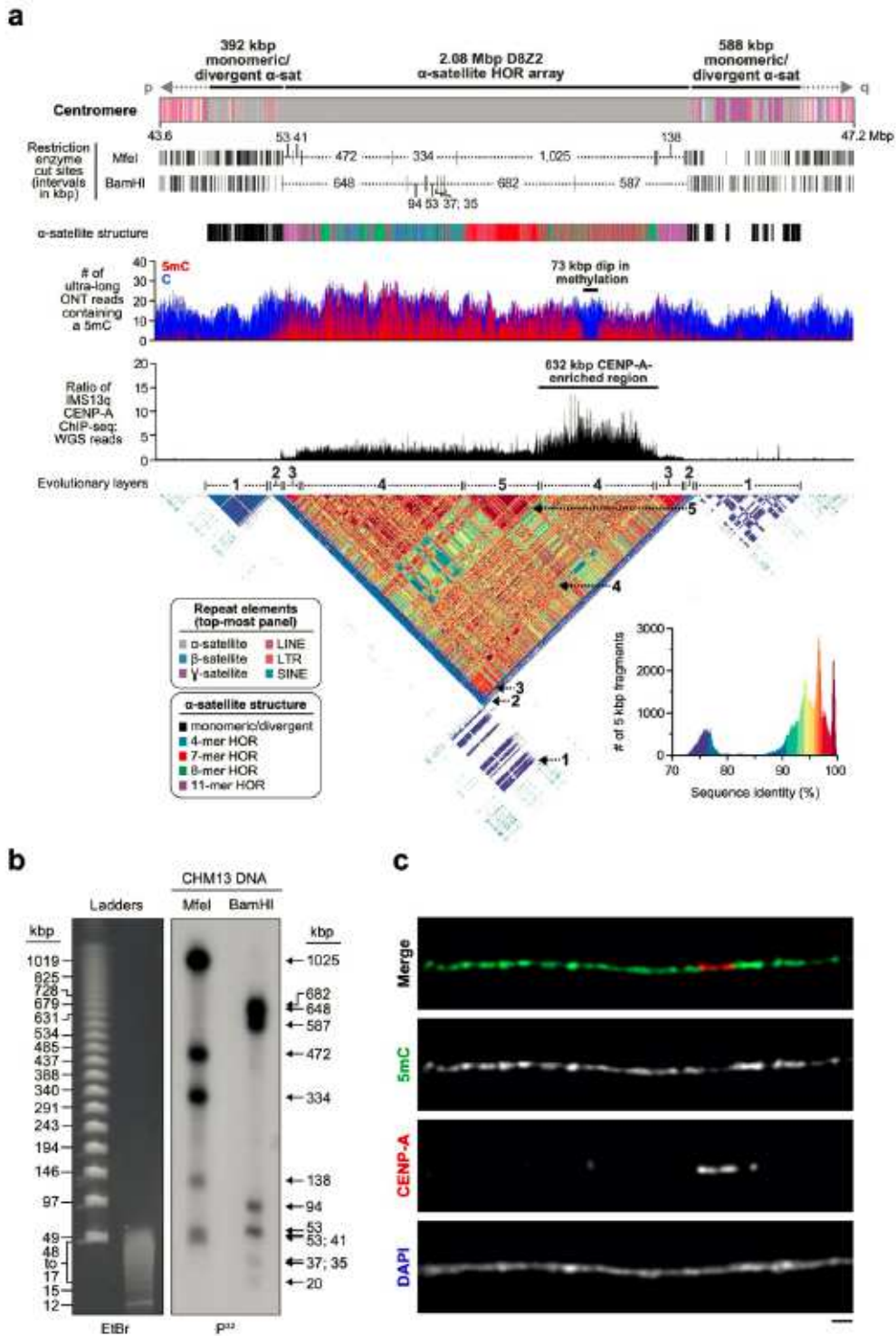


Figure 2

Sequence, structure, and epigenetic map of the chromosome 8 centromeric region. a) Schematic showing the composition of the CHM13 chromosome 8 centromere. The centromeric region is comprised of a 2.08 Mbp D8Z2 α -satellite HOR array flanked by regions of monomeric and/or divergent α -satellite interspersed with retrotransposons, β -satellite, and γ -satellite. The predicted restriction digest pattern is shown and supported by the pulsed-field gel (PFG) Southern blot in Panel b. The D8Z2 α -satellite HOR array is

primarily composed of four types of higher-order repeats (HORs; see Extended Data Fig. 8, Methods for details) and is heavily methylated except for a 73 kbp hypomethylation region. Mapping of normalized CENP-A ChIP-Seq data from a diploid human genome known as IMS13q31 reveals that centromeric chromatin is primarily located within a 632 kbp region encompassing the hypomethylated region (Extended Data Fig. 9 includes another CENP-A ChIP-seq dataset and details). A pairwise sequence identity map across the centromeric region indicates that the centromere is composed of five distinct evolutionary layers (indicated with dashed arrows). b) PFG Southern blot of CHM13 DNA confirms the structure and organization of the chromosome 8 centromeric HOR array indicated in Panel a. Left: EtBr staining; Right: P32-labeled chromosome 8 α -satellite-specific probe. c) Representative images of a CHM13 chromatin fiber showing that CENP-A is enriched in an unmethylated region. Bar = 1 micron.

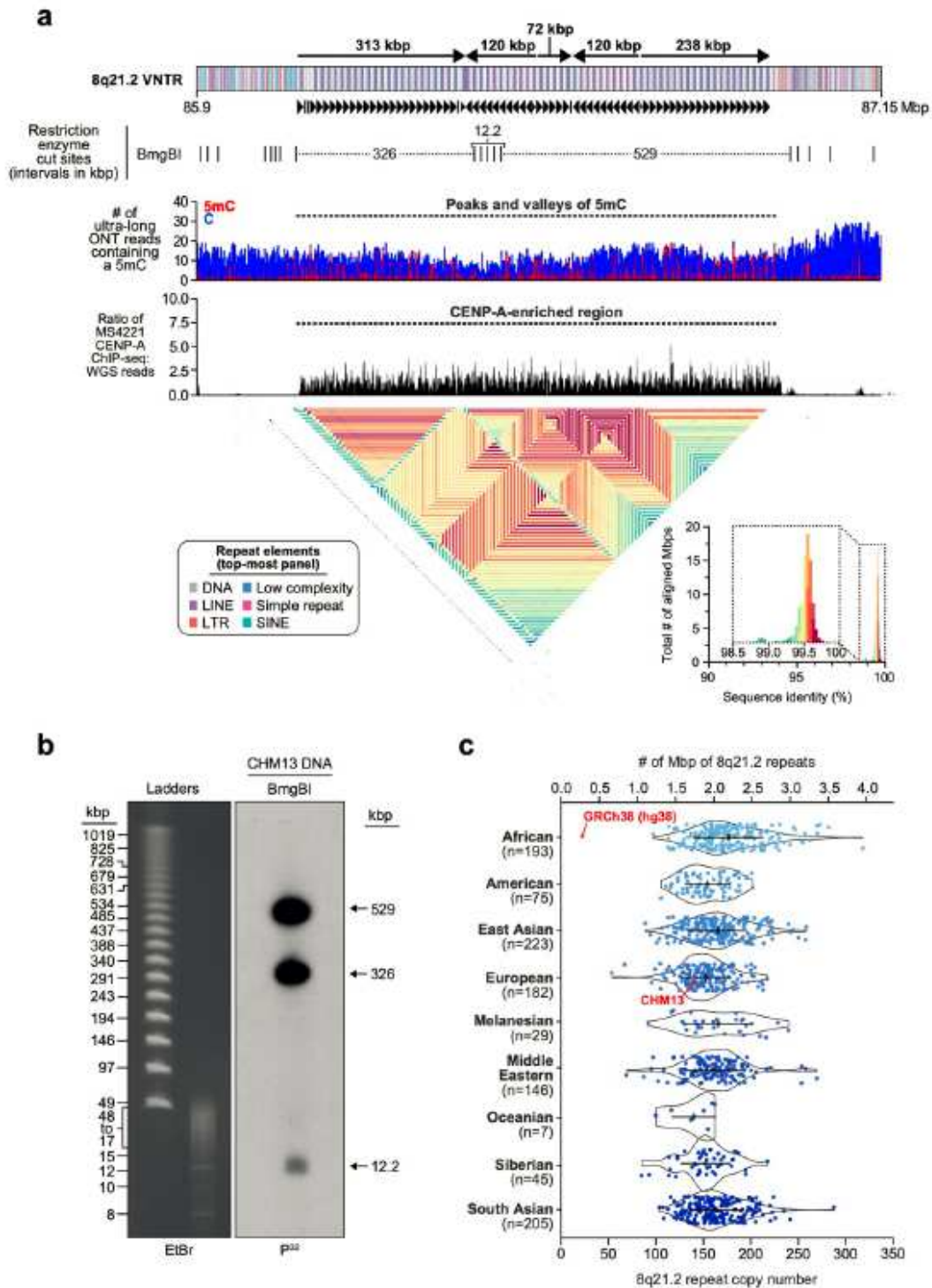


Figure 3

Sequence, structure, and epigenetic map of the neocentromeric chromosome 8q21.2 VNTR. a) Schematic showing the composition of the CHM13 8q21.2 VNTR. This VNTR is comprised of 67 full and 7 partial 12.192 kbp repeats that span 863 kbp in total. The predicted restriction digest pattern is indicated. Each repeat is methylated within a 3 kbp region and hypomethylated within the rest of the sequence. Mapping of CENP-A ChIP-seq data from the chromosome 8 neodicentric cell line known as MS422131,32

(Methods) reveals that centromeric chromatin is primarily located on the hypomethylated portion of the repeat. A pairwise sequence identity map across the region indicates a mirrored symmetry within a single layer, consistent with the evolutionarily young status of the tandem repeat. b) PFG Southern blot of CHM13 DNA digested with BmgBI confirms the size and organization of the chromosome 8q21.2 VNTR. Left: EtBr staining; Right: P32-labeled chromosome 8q21.2-specific probe. c) Copy number of the 8q21 repeat [chr8:85792897–85805090 in GRCh38] throughout the human population. CHM13 is estimated to have 144 total copies of the 8q21 repeat, or 72 copies per haplotype, while GRCh38 only has 26 copies (red data points).

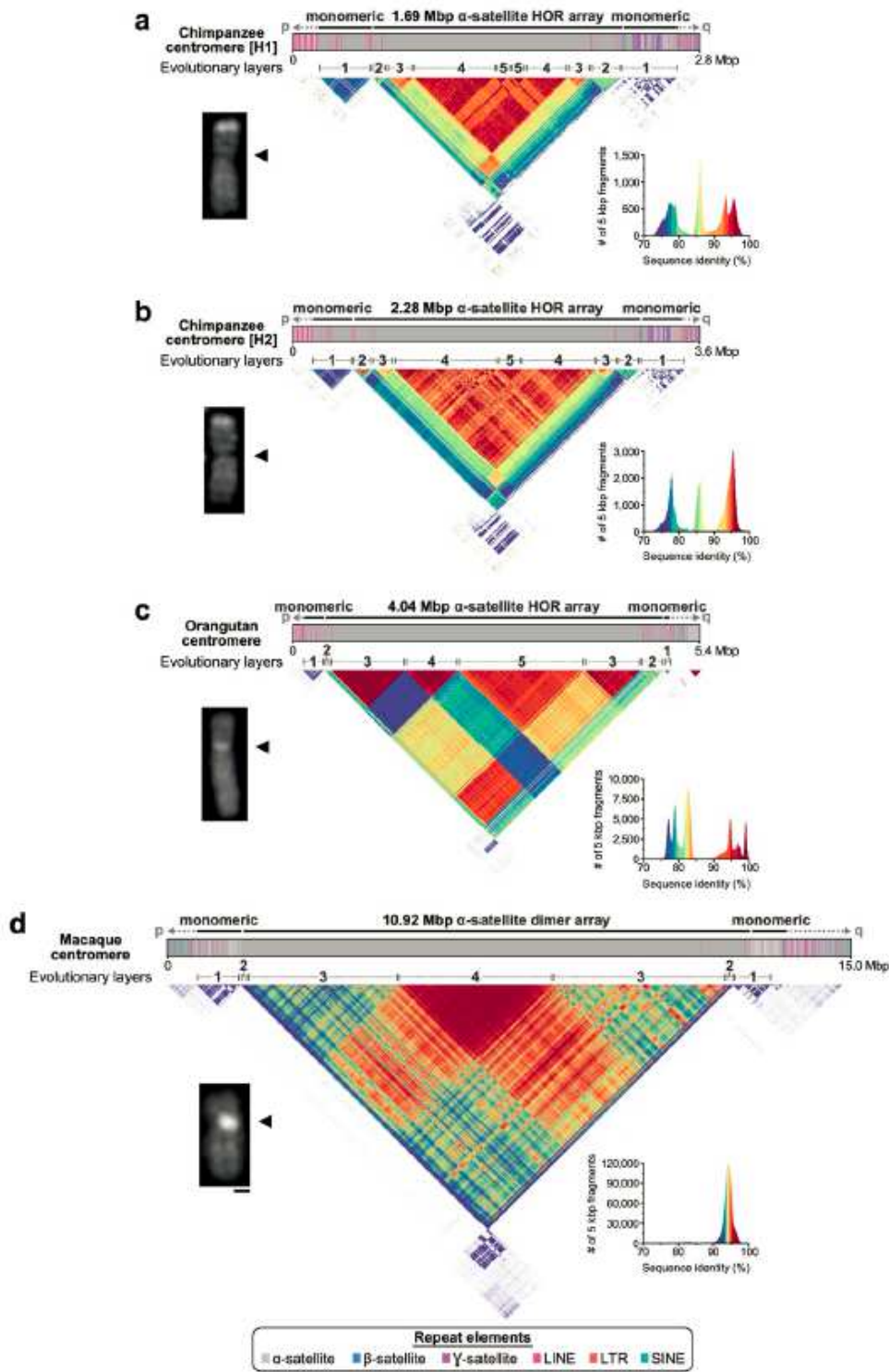


Figure 4

Sequence and structure of the chimpanzee, orangutan, and macaque chromosome 8 centromeres. a-d) Structure and sequence identity of the chimpanzee H1 (Panel a), chimpanzee H2 (Panel b), orangutan (Panel c), and macaque (Panel d) chromosome 8 centromeres. Each centromere has a mirrored organization consisting of either four or five distinct evolutionary layers. The size of each centromeric

region is consistent with microscopic analyses, showing increasingly bright DAPI staining with increasing centromere size. Bar = 1 micron.

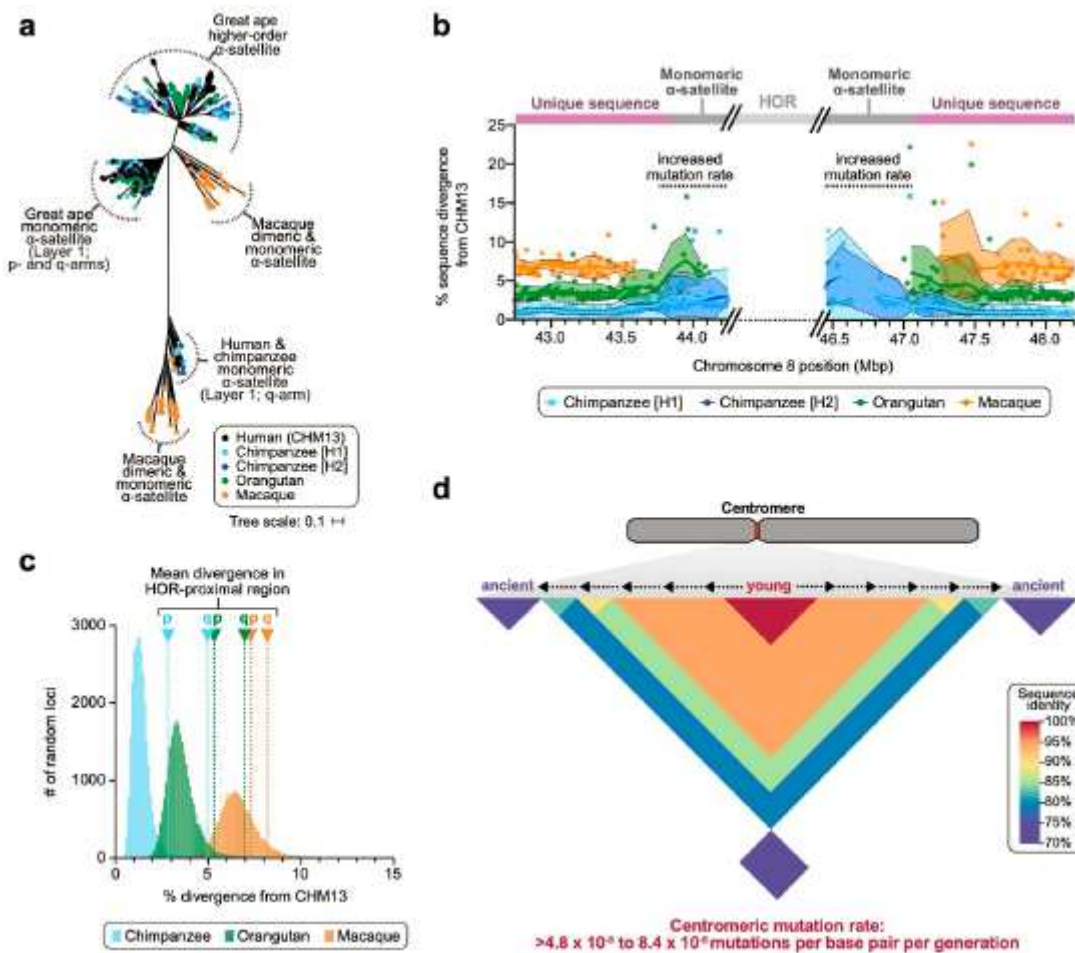


Figure 5

Evolution of the chromosome 8 centromere. a) Phylogenetic tree of human, chimpanzee, orangutan, and macaque α -satellites from the HOR and monomeric regions of the chromosome 8 centromere. A portion of the human and chimpanzee monomeric α -satellite is evolutionarily closer to the macaque α -satellite (bottom of the tree; see Extended Data Fig. 13 for bootstrapping annotations). b) Plot showing the sequence divergence between the CHM13 and nonhuman primates (NHPs) in the regions flanking the chromosome 8 α -satellite HOR array. The mean and standard deviation (bold line and shaded region) are calculated over a sliding window of 200 kbp with a 100 kbp overlap. Individual data points from 10 kbp pairwise sequence alignments are shown. c) Histogram of the sequence divergence between CHM13 and chimpanzee, orangutan, or macaque at thousands of random 10 kbp loci. d) Model of centromere evolution. Centromeric α -satellite HORs evolve in the center of the array via unequal crossing over and homogenization, pushing older, more ancient HORs to the edges, consistent with hypotheses previously put forth^{38,39,41}. The centromeric mutation rate is estimated to be at least 4.8 to 8.4×10^{-8} mutations per base pair per generation, which is 2.2 to 3.8 higher than the mean mutation rate measured from nearly 20,000 random loci.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ExtendedDataTable1.DifferencesinCHM13andGRCh38hg38chromosome8DEFAandDEFBgenes.xlsx](#)
- [ExtendedDataTable4.Chromosome8centromericmutationrate.xlsx](#)
- [ExtendedDataTable6.PacBioIsoSeqdatasets.xlsx](#)
- [ExtendedDataTable7.GeneswithgreatersequenceidentitytoCHM13chromosome8thanGRCh38.xlsx](#)
- [ExtendedDataTable8.CHM13BACsusedinthisstudy.xlsx](#)
- [Logsdonetaextendeddata.pdf](#)