Complete chromosome 21 centromere sequencing of families with Down syndrome reveals centromere size asymmetry

F. Kumara Mastrorosa[1], Alessia Daponte[2], Julie Wertz[1], Allison N. Rozanski[1], William T. Harvey[1], David Porubsky[1], Jordan Knuth[1], Gage H. Garcia[1], Marcelo Ayllon[1], Katherine M. Munson[1], Kendra Hoekzema[1], Weiya He[3], Stephanie L. Sherman[3], Emily G. Allen[3], Tracie C. Rosser[3], Claudia Rita Catacchio[2], Mario Ventura[2], Glennis A. Logsdon[1,2,*], Evan E. Eichler[1,3,*]

[1] Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA

[2] Department of Biosciences, Biotechnology and Environment, University of Bari Aldo Moro, Bari, 70124, Italy

[3] Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA

[4] Current address: Department of Genetics, Epigenetics Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

[5] Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA

* Correspondence: glogsdon@pennmedicine.upenn.edu, ee3@uw.edu

ABSTRACT

Down syndrome, the most common form of human intellectual disability, is caused by nondisjunction and chromosome 21 trisomy (T21). Small centromeres have been hypothesized to contribute to its aetiology and studies on mammals suggest that larger centromeres are more efficiently transmitted, yet complete sequencing of chromosome 21 (chr21) centromeres has been particularly challenging. Using long-read sequencing, we sequenced and assembled the centromeres from eight families that include a child with free T21 (1 trio, 6 child–mother duos, and 1 singleton) all resulting from maternal meiosis I errors. Two of these families carry the smallest chr21 centromeres (143 and 181 kbp) observed in female individuals to date, exhibiting a ~10.7- and ~19.4-fold centromeric α-satellite higher-order repeat array size difference between the maternally inherited homologs, respectively. In both cases, the longer centromere harbors a poorly defined centromere dip region, marked by DNA hypomethylation, in the proband but not in the mother. A comparison of all proband chr21 centromeres (*n*=24) to those of controls (*n*=261) shows that small centromeres are not enriched in families with T21 (p-value=0.73); contrarily, chr21 extreme centromere size asymmetry (>10-fold) is unique of T21 (p-value=0.003), suggesting that this feature may represent a genetic risk factor for a subset of families with free T21. Additionally, phylogenetic reconstruction reveals that human chr21 has been particularly prone to such variation with some of the biggest size differences occurring over the last ~17 thousand years of human evolution.

KEYWORDS

Down syndrome, Trisomy 21, centromere, chromosome 21, intellectual disability, aneuploidy, long-read sequencing, methylation, genome assembly

42 **LIST OF ABBREVIATIONS**

43 **CDR** = centromere dip region

44 **chr21** = chromosome 21

45 **FISH =** fluorescence *in situ* hybridization

46 **Gbp =** Gigabase pairs

47 **IF =** immunofluorescence

48 **kbp** = kilobase pairs

49 **kya** = thousand years ago

50 **H1/2/3/4** = haplotype 1/2/3/4

51 **HiFi** = high fidelity

52 **HGSVC** = Human Genome Structural Variation Consortium

53 **HPRC** = Human Pangenome Reference Consortium

54 **HOR** = higher-order repeat

55 **MMI/IIE** = maternal meiosis I/II error

56 **Mbp** = Megabase pairs

57 **PacBio** = Pacific Biosciences

58 **T21** = Trisomy 21

59 **T2T** = telomere-to-telomere

60 **UL-ONT** = ultra-long Oxford Nanopore Technologies sequencing

INTRODUCTION

Trisomy 21 (T21) is the most common autosomal chromosomal aneuploidy in humans[1] and the most common genomic cause of intellectual disability[2]. The clinical condition, known as Down syndrome, in most cases (>95%) is caused by the presence of an entire extra chromosome 21 (chr21) in the cells of the individual with the syndrome (free T21)[2]. Other cases are caused by a portion or the entire long arm of chr21 translocated to another chromosome as a result of Robertsonian translocations t(14;21) or t(21;21) or mosaic occurrences[2]. The majority of cases arise as a result of maternal meiosis I errors (MMIEs; ~66%), followed by maternal meiosis II errors (MMIIEs; ~21%), paternal meiosis I (~3%) and meiosis II (~5%) errors, or mitotic events after the zygote formation (~5%)[2].

Centromeres play an essential role in accurate chromosome segregation during meiosis by promoting equal segregation of homologous chromosomes and sister chromatids. Human centromeric regions consist of different types of satellite DNA, with the most abundant type being α-satellite: a ~171 base pair (bp) nearly identical sequence tandemly repeated up to several megabase pairs (Mbp) in length[3]. For most human chromosomes, the functional centromere is composed of α-satellite repeats organized into higher-order repeat (HOR) arrays that tend to diverge in sequence at the pericentromeric edges. Until recently, the complete sequence of all human centromeres was unknown because they could not be assembled with standard short-read sequencing data. However, in the last five years, improvements in long-read sequencing technologies and computational assembly methods have made it possible to assemble centromeres[4–7], the short arm of acrocentric chromosomes[8,9], large segmental duplications[10,11], and other complex loci[8,12,13] previously regarded as inaccessible.

85 Functionally, the kinetochore is marked by the presence of nucleosomes containing the histone
86 variant H3 CENP-A, which provides the foundation for the assembly of inner- and outer-
87 kinetochore proteins and plays a crucial role in cell division and chromosome segregation[14,15].
88 The analysis of the complete sequence of human centromeres[5–7] has made it possible to define
89 the likely site of kinetochore attachment: a pocket of hypomethylated CpG DNA called the
90 centromere dip region (CDR)[16,17], where chromatin immunoprecipitation followed by sequencing
91 (ChIP-seq) shows an enrichment of CENP-A. All completely sequenced human centromeres to
92 date derive from primary tissue (blood) or cell lines and possess CDRs up to tens of kilobase
93 pairs (kbp) in length where CENP-A accumulates, implicating this as a general epigenetic
94 signature of the kinetochore[7].
95
96 Given their role in cell division, centromeres, and in particular length differences, have been
97 postulated to play an important role in chromosome nondisjunction, including Down
98 syndrome[18,19]. Some studies on mice suggest that centromeres with longer repeat arrays are
99 preferentially retained in the egg during maternal meiosis I[20–23] and that in general, centromeric
100 features, such as sequence and length, and kinetochore components[23–26] are correlated: a
101 concept known as “centromere strength.” A recent investigation in humans also suggested that
102 individuals with T21 might uniformly have smaller chr21 centromeres (possibly as a result of
103 reduced size of the centromeric α-satellite HOR array)[27]. Centromere size was estimated,
104 however, based on quantitative, real-time PCR[27], which can be challenging given the high
105 degree of sequence identity between chromosome 13 and 21 α-satellite DNA and a limited
106 understanding of the sequence diversity within the allelic α-satellite HOR array[27]. Other studies
107 showed that cohesion loss between sister centromeres[28] and reduced recombination[29] increase
108 the risk of nondisjunction. Nevertheless, the most well-established nongenetic risk factor is
109 increased maternal age at conception[19], though there is not a well-established genetic risk
110 factor for free T21.

111

112 It is currently unknown whether centromere variation amongst humans increases the likelihood
113 of nondisjunction. For instance, centromere size, sequence, allelic differences, and epigenetic
114 features might influence nondisjunction risk and indicate predisposition to such events.
115 Recently, combining both Pacific Biosciences (PacBio) high-fidelity (HiFi) long-read sequencing
116 and ultra-long Oxford Nanopore Technologies (UL-ONT) sequencing, we demonstrated that it is
117 possible to completely sequence and distinguish human autosomal alleles for chr21[7]. Here, we
118 present the first complete linear sequence and epigenetic profile of chr21 centromeres in eight
119 individuals with MMIE free T21 and eight of their parents (7 mothers and 1 father). We also
120 compare genetic and epigenetic features of the chr21 centromeres to a control group of 261
121 completely sequenced chr21 centromeres from two hydatidiform moles and control individuals
122 generated as part of the Human Genome Structural Variation Consortium[30] (HGSVC) and
123 Human Pangenome Reference Consortium[31] (HPRC) to reveal unique features of T21
124 centromeres.

125

126 **RESULTS**

127 **Maternally inherited chr21 centromeres show size asymmetry in an individual with T21.**
128 To study the chr21 centromeres of an individual with T21, we initially sequenced the DNA of a
129 proband (maternal age at childbirth = 29) using both PacBio HiFi sequence data (59×) and UL-
130 ONT sequence data (23×). We used Verkko, a hybrid genome assembly workflow[32], to
131 assemble all proband chr21 centromere haplotypes (total genome assembly size = 6.118 Gbp,
132 contigs = 3,466, contig N50 = 7.72 Mbp). After standard validation and quality control[5,7]
133 (**Methods**, **Supplementary Figure 1**), sequence assembly of phased chr21 centromere
134 haplotypes confirmed an MMIE where haplotypes 1 (H1) and 2 (H2) arose from maternal
135 nondisjunction and haplotype 3 (H3) was inherited from the father (**Figure 1a**).

136

137 H1 and H3 centromeres are the most similar in size with an α-satellite HOR array of ~1.53 Mbp
138 and ~1.47 Mbp, respectively. By contrast, maternal H2 is 10.7-fold smaller, with an α-satellite
139 HOR array of only 143 kbp (**Figure 1a**). The α-satellite HOR array sequences from all three
140 haplotypes are predominantly composed of 11-monomer α-satellite HORs (~86-93%) with a
141 smattering of rarer α-satellite HORs 2-, 10-, 13-, 17-, and 18-monomers in length (each
142 constituting <3% of the sequence) (**Figure 1a; Supplementary Table 1**). Other than their size,
143 the three chr21 centromeres share a remarkably similar organization, with the α-satellite HOR
144 array flanked by human satellite 1a (HSat1a; annotated as "SAR'' by RepeatMasker) on the p-
145 arm and various classes of mobile elements (SINEs and LINEs) on the q-arm. Also, H3 is
146 distinguished by a 317 bp AluY retrotransposon insertion in its α-satellite HOR array (**Figure**
147 **1a**).
148
149 Given the use of long-read sequencing, which allows for CpG methylation to be detected, we
150 were able to assess the position and distribution of each chr21 centromere CDR, the pocket of
151 hypomethylated DNA indicating the location of the kinetochore[5,6,17]. Both H2 and H3 match this
152 description with a CDR of 120 kbp and 135 kbp, respectively. By contrast, the maternally
153 inherited H1 does not have a clearly defined CDR. The CDR is fragmented into five smaller
154 pockets ranging from 10 to 50 kbp and spanning across 280 kbp of centromeric α-satellite
155 HORs (**Figure 1b**). For each CDR, we calculated the percentage of methylated CpGs (defined
156 as having ≥25% methylated CpGs at the same position; **Methods**) across the region comprising
157 the CDR pockets (from the start of the first CDR window to the end of the last one). We found
158 that the percentage of methylated CpGs within H1's CDR was much higher than those of H2
159 and H3 CDRs (~71% vs. ~39% and ~38%, respectively; **Figure 1b**). This more diffuse pattern
160 of CDRs is associated with the largest HOR array within this family. We orthogonally validated
161 the centromere methylation profiles using PacBio HiFi sequencing data (**Supplementary**
162 **Figure 2**).

163

164 Because of the central role of centromeric proteins, such as CENP-A and CENP-C, in defining
165 the site of kinetochore assembly other than the CDRs, we assessed the level of CENP-A and
166 CENP-C on chr21 centromeres in proband AG16777 and their parents using an
167 immunofluorescence (IF) combined with fluorescence *in situ* hybridization (FISH) targeting the
168 α-satellite sequence, CENP-A, and CENP-C proteins (**Supplementary Figure 3**). We observed,
169 however, no statistically significant differences in CENP-A or CENP-C signal intensities in any
170 member of the family.

171

172 For each chr21 centromere, we also generated sequence identity heatmaps, based on pairwise
173 sequence alignment of 5 kbp segments using StainedGlass[33] (**Figure 1c**). All three haplotypes
174 show similar compositional features, with the most identical segments mapping to the central
175 section of the α-satellite HOR array but not corresponding to the CDR, with the exception of H2
176 due to its smaller size. This structure is also similar to the complete chr21 α-satellite HOR arrays
177 described for two complete hydatidiform moles (CHM1 and CHM13)[5,6], as well as population
178 samples[7], and generally follows the Smith organizational model with monomeric satellite at the
179 edges as previously described in 1976[34]. For both H1 and H3, the CDR and the predicted site of
180 kinetochore attachment occur peripherally in a region composed of more heterogeneous α-
181 satellite HOR cassettes (with an admixture of 5- and 11-monomer HORs).

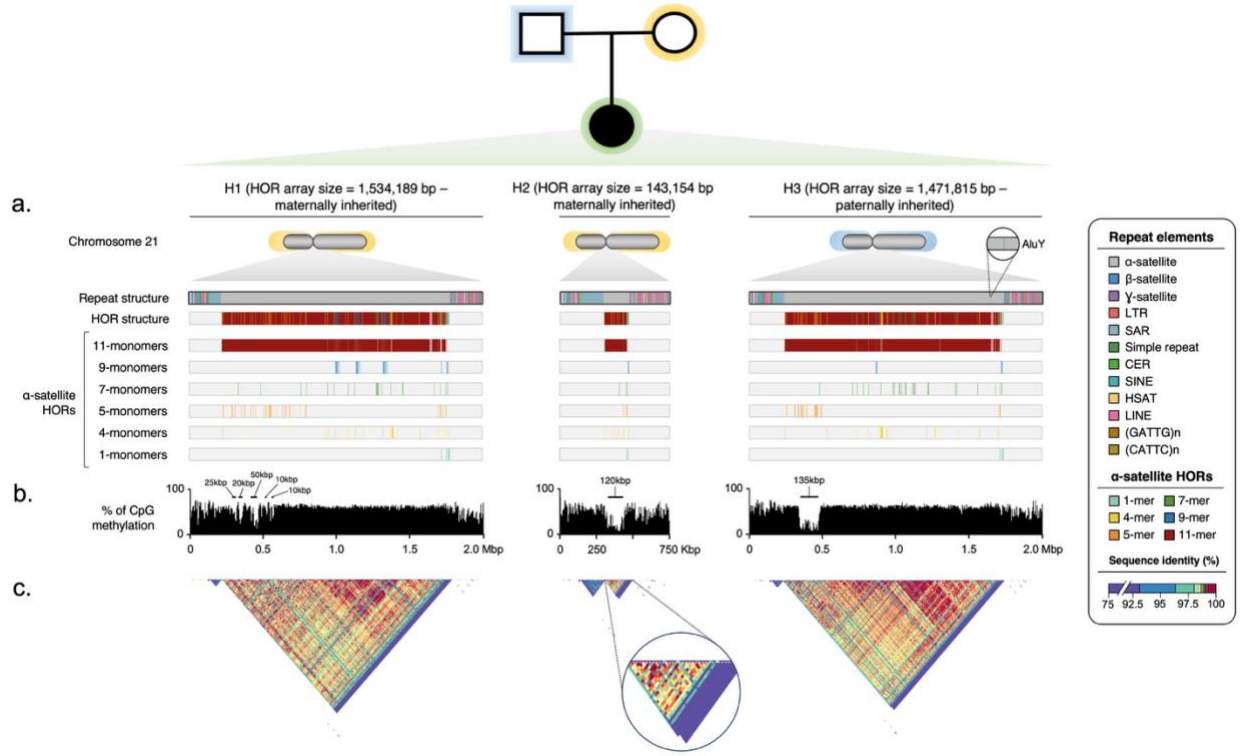
182

183 For this index family, we also studied additional proband centromeres (chromosomes 4, 12, 15)
184 for which both haplotypes were correctly assembled (**Supplementary Figure 4**). Chromosome
185 12 homologs are the only ones having the CDRs roughly in the same relative position on both
186 homologs. Chromosome 4 and 15 haplotype CDRs are positioned differently instead and one of
187 chromosome 4 is divided into three windows intervaled by small highly methylated regions

188 (**Supplementary Figure 4**) confirming the CDR heterogeneity recently observed by Logsdon et

189 al. from their analysis of cell line samples as part of the HGSVC[35].

190



191

192 **Figure 1. Genetic composition and epigenetic profile of three chromosome 21 (chr21) centromeres**
193 **from an individual with Trisomy 21 (T21). a)** Size and repeat composition of three centromeric
194 haplotypes for an individual with T21 (Coriell ID: AG16777). Different α-satellite higher-order repeat
195 (HOR) monomers mapping within the α-satellite array are colored (see legend). Only α-satellite HOR
196 monomers present more than 10 times in at least one haplotype are shown (see **Supplementary Table 1**
197 for complete monomer composition). **b)** Methylation profile of each centromere with hypomethylated
198 pockets indicating the location of the kinetochore. **c)** Pairwise sequence identity heatmaps
199 (StainedGlass[33]). Sequence identity is depicted in 5 kbp windows across the centromeres revealing
200 regions of high- and low-identity homology within each α-satellite HOR array (see legend).
201

202 **Parental chr21 centromeres show transgenerational methylation changes and suggest a**

203 **preferential CDR position.** To understand the inheritance of the chr21 centromeres, we

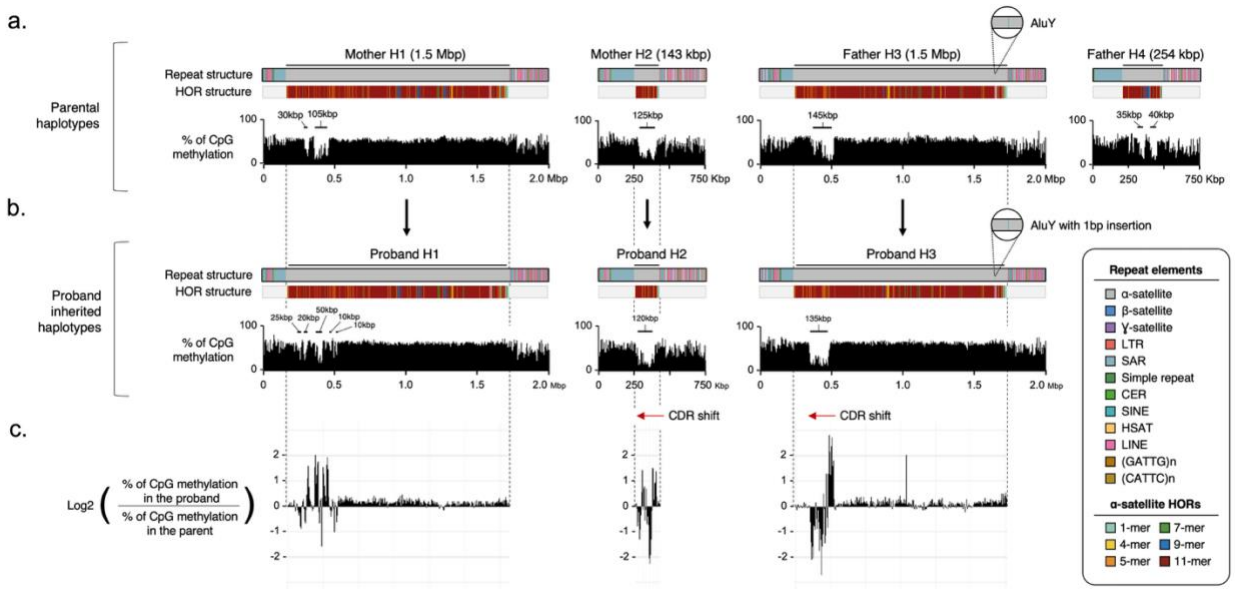
204 sequenced and assembled the genomes of both parents using DNA extracted from

205 lymphoblastoid cell lines and fully reconstructed each of the centromere haplotypes from the

206 diploid parental cell lines. We generated PacBio HiFi sequencing data at 37× coverage for the

207 mother and 40× for the father and complemented these data with UL-ONT data at 25× and 12×
208 coverage, respectively, to generate *de novo* assemblies (mother: total genome assembly size =
209 6.078 Gbp, contigs = 4,827, contig N50 = 5.41 Mbp; father: total genome assembly size = 5.947
210 Gbp, contigs = 5,032, contig N50 = 5.01 Mbp). Centromeric contigs from chr21 were extracted,
211 validated, and compared to the child with T21 (see **Methods** for assembly correction).
212
213 Sequencing confirmed that the largest and the smallest chr21 centromeres (H1 and H2,
214 respectively) were maternally transmitted, while H3 originated from the father (**Figure 2a-b**).
215 Sequence identity between parent and proband haplotypes was >99.99% identical, confirming
216 the accuracy of the assembly as well as the stability of the centromere transgenerationally and
217 in cell culture (**Supplementary Figure 5**). No difference in the α-satellite repeat composition,
218 organization, or array size was observed, except for a 1 bp deletion in the poly(T) tract
219 associated with an AluY insertion on the paternal haplotype (H3) when compared to the
220 proband (**Supplementary Figure 6**). We note that the second non-transmitted paternal
221 haplotype (father H4) has an α-satellite HOR array size of 254 kbp (**Figure 2a**). This is 112 kbp
222 larger than the small maternal haplotype (H2). Like the other haplotypes, the 11-monomer α-
223 satellite HORs constitute the majority of its array sequence (~86%), with 2- to 5-, 7-, 9-, 10-, and
224 13-monomer HORs at lower frequency (each composing <6% of the array sequence).
225
226 Comparing the centromeric CpG methylation profiles between the parents and the proband, we
227 find that the atypical methylation pattern observed in proband H1 (five hypomethylated pockets
228 10-50 kbp in size) appears to be specific to the proband (**Figure 2a-b**; see **Supplementary**
229 **Figure 2** for orthogonal validation). The mother shows a more typical CpG methylation profile,
230 with two clear CDRs of 30 and 105 kbp in length, and ~46% methylated CpGs across the region
231 comprising the CDR pockets. This is 25% less compared to the proband (~71%), and it
232 contrasts with the transmitted maternal H2 (42% vs. 39% in proband) and paternal H3 (48% vs.

233 38% in proband), which show more similar CDR patterns. The CDRs of chr21 H2 and H3
234 centromeres differ by only an estimated 5-10 kbp in size between the generations (**Figure 2c**).
235 We also observe a shift of both proband H2 and H3 CDRs towards the p-arm relative to the
236 parents despite nearly identical sequences. The maternal H2 CDR, for example, starts 37 kbp
237 from the beginning of the α-satellite array, while in the proband, the hypomethylated pocket
238 starts 11 kbp upstream. The paternal H3 has a CDR initiating 131 kbp from the beginning of the
239 α-satellite HOR array, which is shifted 25 kbp towards the p-arm in the proband (**Figure 2c**).



240

241 **Figure 2. Genetic and epigenetic features of parent–proband transmission of chr21 centromeres.**
242 Composition and organization of α-satellite HOR arrays (only α-satellite HORs present more than 10
243 times in at least one haplotype are shown; see **Supplementary Table 1** for complete HOR composition),
244 with CpG methylation profiles of **a)** parental chr21 centromeres and **b)** transmitted proband chr21
245 centromeres. **c)** Log_2 ratio of CpG methylation frequency between proband and parental haplotypes.
246

247 **Analysis of additional families with Down syndrome reveals another case of asymmetric**
248 **chr21 centromeres.** To validate the findings from the index trio, we generated PacBio HiFi and
249 UL-ONT sequence data from seven additional families: six mother-proband duos where MMIEs
250 had been previously identified and one singleton where parental DNA was not available.
251 Importantly, unlike the original trio, in six of these families both blood and cell line DNA material
252 were available for investigation. Because the maternal age of the index case was 29 years, we
253 specifically selected mothers representing a younger age range (20-34 years) (**Supplementary**
254 **Table 2**). Despite the well-known maternal age effect with T21, we note, however, that this age
255 range still represents the majority of the cases in the extended collection from the Atlanta and
256 National Down Syndrome Projects[36,37] (maternal age at birth <35 for ~54% of the cases;
257 2,171/3,985), with a mean maternal age at birth of 32.9 years old. We assembled their genomes
258 with the same methods used for the original family (probands mean genome assembly size =
259 6.072 Gbp, # of contigs = 1,237, contig N50 = 45.84 Mbp; mothers mean genome assembly
260 size = 6.064 Gbp, # of contigs = 2,217, contig N50 = 35.16) and analyzed their chr21
261 centromeric sequences. Of the proband centromeres, 95% (20/21) were completely assembled
262 while 67% (8/12) of the maternal centromeric haplotypes were assembled without any
263 detectable error.
264
265 In this extended cohort, we identified one additional family showing extreme centromere size
266 differences similar to that described in the original trio with T21. Duo 7210079 shows a 19.4-fold
267 chr21 centromere size difference between maternal homologs (original trio: 10.7-fold) with a
268 small centromere α-satellite HOR array of only 181 kbp (original trio: 143 kbp) and a longer one
269 of ~3.5 Mbp (original trio: 1.5 Mbp). In this duo as well, we observed a *de novo* epigenetic
270 change in the centromeric methylation profile, with the proband having a less-defined CDR
271 compared to the maternal for the largest haplotype (**Figure 3**). Since the 7210079.00-H3
272 centromere was not assembled completely in the proband, we leveraged the maternal locus

273 (assembled contiguously with few sequence errors) as a reference to align UL-ONT reads both
274 from mother and child and compare methylation profiles and CDR structure. Specifically, the
275 maternal CDR maps to a unique window of 150 kbp more proximally toward the p-arm, while in
276 the proband, the CDR is divided into five smaller windows (size range 75-10 kbp) despite
277 remaining in the same genomic area (**Figure 3**). Mother 7210079.20 had a similar age at
278 childbirth (age: 27) as the original trio mother (age: 29).
279
280 A third family from the cohort with T21, 7210223, showed an epigenetic change between the
281 transmitting parent and proband despite showing only a 2.6-fold centromere size difference
282 between maternal homologs (587 kbp and 1.5 Mbp). Contrary to what we observed for the other
283 two families with extreme maternal size difference, we observe a CDR shift occurring on the
284 smaller maternally inherited centromere for proband 7210223.00. The 7210223.00-H2
285 centromere (size: 587 kbp) shows four CDR windows in the mother (size range 50-10 kbp)
286 distributed across 170 kbp, while in the proband, we detect six smaller CDRs ranging in size
287 from 10 to 20 kbp distributed across 310 kbp (**Supplementary Figure 7**). For all other families,
288 only minor transgenerational changes in the CDR methylation profiles were observed for chr21
289 centromeres (**Supplementary Figure 7**). Among the remaining five probands, all have maternal
290 homologs that are at least 1 Mbp large and a size difference of less than twofold
291 (**Supplementary Table 2**).
292
293 With respect to sequence composition, most proband centromeric α-satellite HOR arrays in this
294 cohort are similar to the original, composed predominantly of 11-monomer α-satellite HORs
295 (~77-94%), with rarer HORs representing the rest of the sequence. Notably, 4710291.00-H3
296 has 10% of its sequence composed of a 4-monomer α-satellite HOR, 7210239.00-H1 has 8%,
297 and 7210008.00-H1 has 5%. The smallest haplotype of this cohort, 7210079.00-H1, has 9% of
298 its sequence organized in a 9-monomer HOR (**Supplementary Table 1**). However,

299 7210079.00-H3 is unique compared to all the other centromeres. Based on the maternal α-
300 satellite HOR array structure (**Figure 3**), it is the largest centromere sequenced amongst all
301 characterized (α-satellite HOR array of ~3.5 Mbp), with almost 1 Mbp more sequence than the
302 second largest (4710291.00-H3). Also, its predominant α-satellite HOR array unit is composed
303 of 5-monomer units (88%) with only 6.65% of its sequence made up of the more typical 11-
304 monomer units.
305
306 Because both blood and cell line DNA were available for a subset of the probands, we
307 compared the chr21 centromere methylation patterns for the two sample types for the five
308 probands with complete centromeres. PacBio HiFi data were generated from blood DNA, and
309 UL-ONT data were generated from cell line DNA. We confirm that the CDR positions remain the
310 same from primary tissue (blood) to cell line (lymphoblastoid) (**Supplementary Figure 8**). This
311 was concluded by observing ONT and HiFi data processed with the latest methylation callers
312 dorado (v0.4.2) or guppy (6.3.7 or newer) and Jasmine (v2.2.1), respectively. We also
313 confirmed the methylation callers' concordance in the original trio AG167 (**Supplementary**
314 **Figure 2**) and other samples (**Supplementary Figure 8**), for which the two long-read datasets
315 were generated from cell lines. Despite a stable CDR position, we observed an overall higher
316 level of methylation in blood and more readily defined CDR in lymphoblastoid cell lines
317 (**Supplementary Figure 8**).

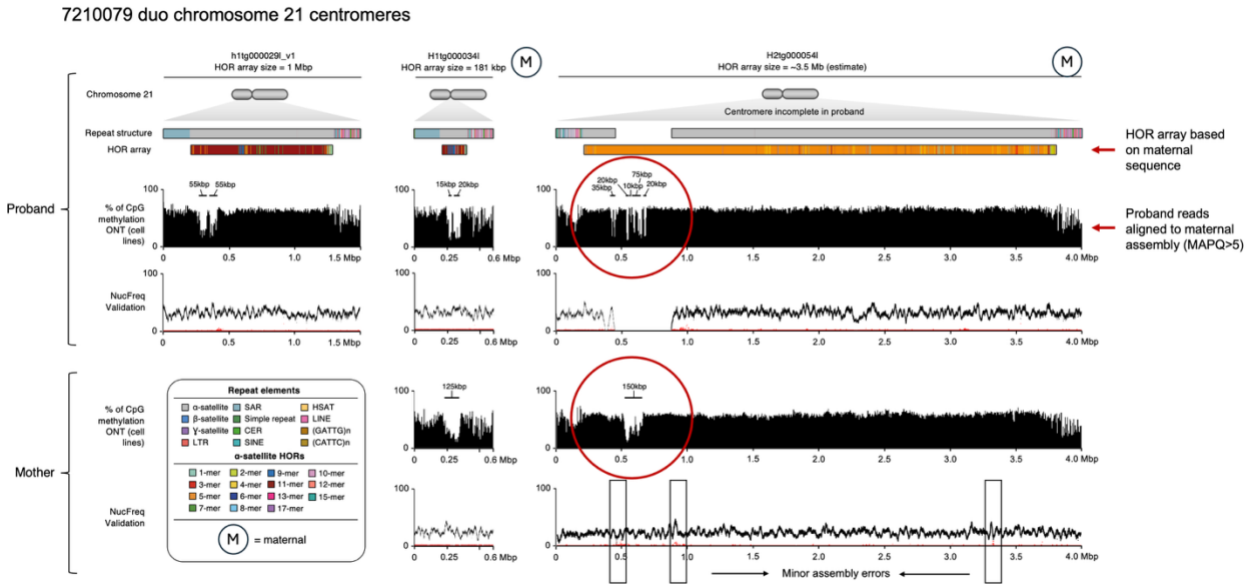
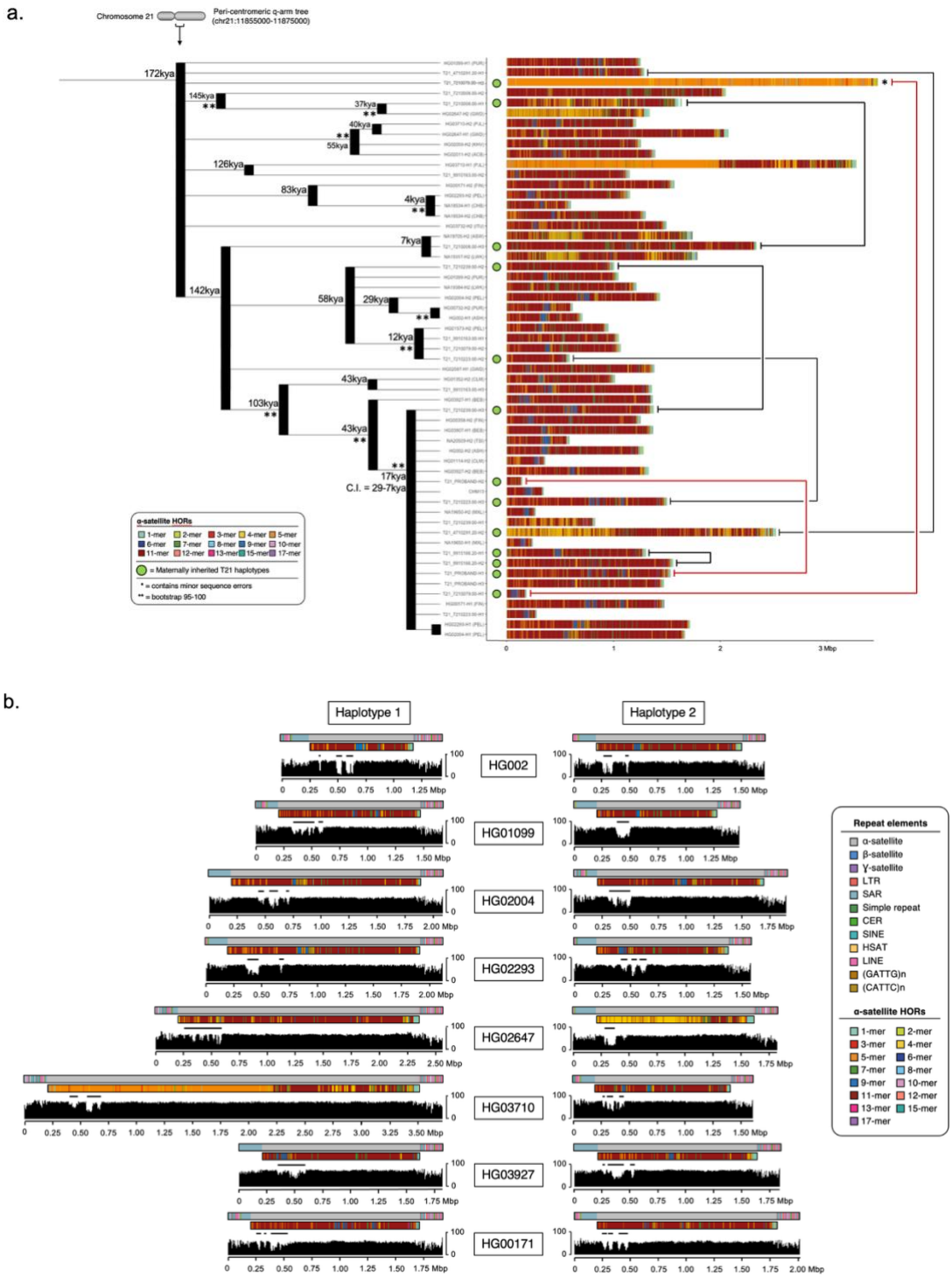


Figure 3. **Chromosome 21 (chr21) centromeres in duo 7210079 exhibit size asymmetry and transgenerational methylation changes**. Repeat structure, α-satellite HOR array organization, CpG methylation profile, and NucFreq validation of each chr21 centromere in 7210079 mother–child duo. Transgenerational DNA methylation changes are circled in red. The 7210079.00 proband H3 was not completely assembled. Since the same haplotype was assembled in the mother, we used the maternal assembly as reference to study the methylation profile in the proband.

325 **Population chr21 centromere diversity and phylogeny.** To understand genetic and
326 epigenetic diversity in the general population, we initially selected 35 fully resolved chr21
327 centromere haplotypes from 16 different ethnic groups representing all five ancestry
328 superpopulations as defined by the 1000 Genomes Project[38]. We observed considerable
329 diversity in α-satellite HOR composition and length amongst these 35 complete chr21
330 centromeres. While the most prevalent α-satellite HOR is 11-monomers in length and comprises
331 ~83% to ~95% of the α-satellite HOR array sequence, in most of the complete chr21
332 centromeres, we also detect less abundant α-satellite HORs that are 12, 14, 15, 16, and 20 α-
333 satellite monomers in length (**Supplementary Table 1, Figure 4a-b**). There are two notable
334 haplotypes whose composition differs markedly from the predominant 11-monomer α-satellite
335 HOR organization: an individual of Gambian (GWD) descent, HG02647, carries a 4-monomer α-
336 satellite HOR composing ~15% of its H2 centromere sequence; and an individual of Punjabi
337 (PJB) descent, HG03710, with 44% of its H1 centromere sequence composed of 5-monomer α-
338 satellite HORs (**Figure 4c, Supplementary Table 1**). The 7210079.00-H3 described earlier is
339 the only other chr21 centromere containing a majority of 5-monomer HORs (88%), and along
340 with HG03710-H1, represent the two largest chr21 centromeres sequenced to date. Sequence
341 identity plots (**Figure 4c**) also reveal that the 11-monomer-enriched α-satellite HOR region of
342 HG03710-H1 shows greater sequence identity when compared to the 5-monomer-enriched
343 regions. In contrast, the 4-monomer α-satellite HOR of HG02647-H2 shows greater sequence
344 identity when compared to other flanking monomers (**Figure 4c**). If sequence divergence is a
345 proxy for evolutionary age, our analyses may suggest that the 4-monomer α-satellite HOR has
346 more recently evolved, while the 5-monomer α-satellite HOR is more ancient. Notwithstanding,
347 both of these rarer centromere haplotypes map their CDRs to the noncanonical 5-monomer and
348 4-monomer α-satellite HORs.
349

350 We also compared the diversity in length and location of the CDRs for the various chr21
351 centromere haplotypes. The number of CDRs typically range from 1 to 4 and span a total of 90-
352 335 kbp in total length (**Figure 4b; Supplementary Figure 9**). None show patterns similar to
353 the one observed in the proband (H1) with multiple small CDRs mapped in close proximity.
354 However, some haplotypes, such as HG02647-H1 and HG03927-H1, show large CDR windows
355 with peaks of higher methylation within. We observe four chr21 centromere haplotypes that
356 have two hypomethylated regions separated by more than 100 kbp: HG002-H1 and H2;
357 HG02293-H1; and HG01573-H2. The dispersed CDRs tend to be smaller (10 to 25 kbp) and
358 contained within a single pocket, except for HG01573-H2, which has two. Irrespective of
359 sequence composition, 91% (20/22) of the chr21 CDRs map to the p-arm of the centromere,
360 suggesting a polarity or preference for CDR location (**Supplementary Figure 9**).
361
362 We also constructed a chr21 centromere phylogenetic tree using 20 kbp of flanking
363 pericentromeric sequence in the q-arm from the genomes assembled in this study and
364 chimpanzee as an outgroup (**Methods**). We have shown previously that these flanking
365 sequences can serve as a reasonable proxy of evolutionary history due to the large block of
366 linkage disequilibrium created by the suppressed recombination around the centromere[7]. We
367 find that 50% (11/22) of T21 haplotypes included in the analysis are part of a large monophyletic
368 clade comprising 40% (23/57) of all chr21 centromeres (**Figure 4a**). The clade harbors the
369 largest (H1) and smallest (H2) centromeres for proband AG16777, which are predicted to have
370 arisen from a common ancestor 17 thousand years ago (kya), despite the fact that these two
371 haplotypes vary by more than 10-fold in their overall length. It also contains the small
372 centromere found in the second family (7910079) showing extreme centromere size asymmetry.
373 We note that this particular clade is significantly enriched for centromeres smaller than 500 kbp
374 (Fisher's exact test p-value = 0.001). In contrast, our analysis also reveals that the two large
375 chr21 centromeres containing an abundance of 5-monomers α-satellite HORs (7210079.00-H3

and HG03710-H1) diverged much earlier (126 and 172 kya) compared to the other haplotypes. The low degree of sequence identity with other canonical 11-monomers (**Figure 4a**) confirms an ancient origin, possibly the result of an archaic introgression.



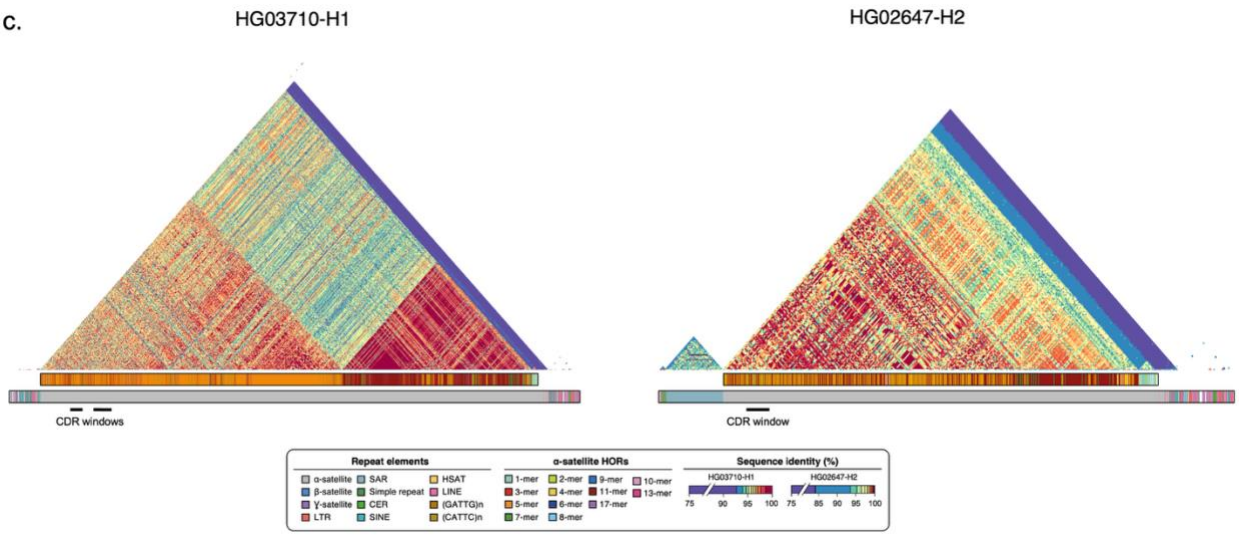
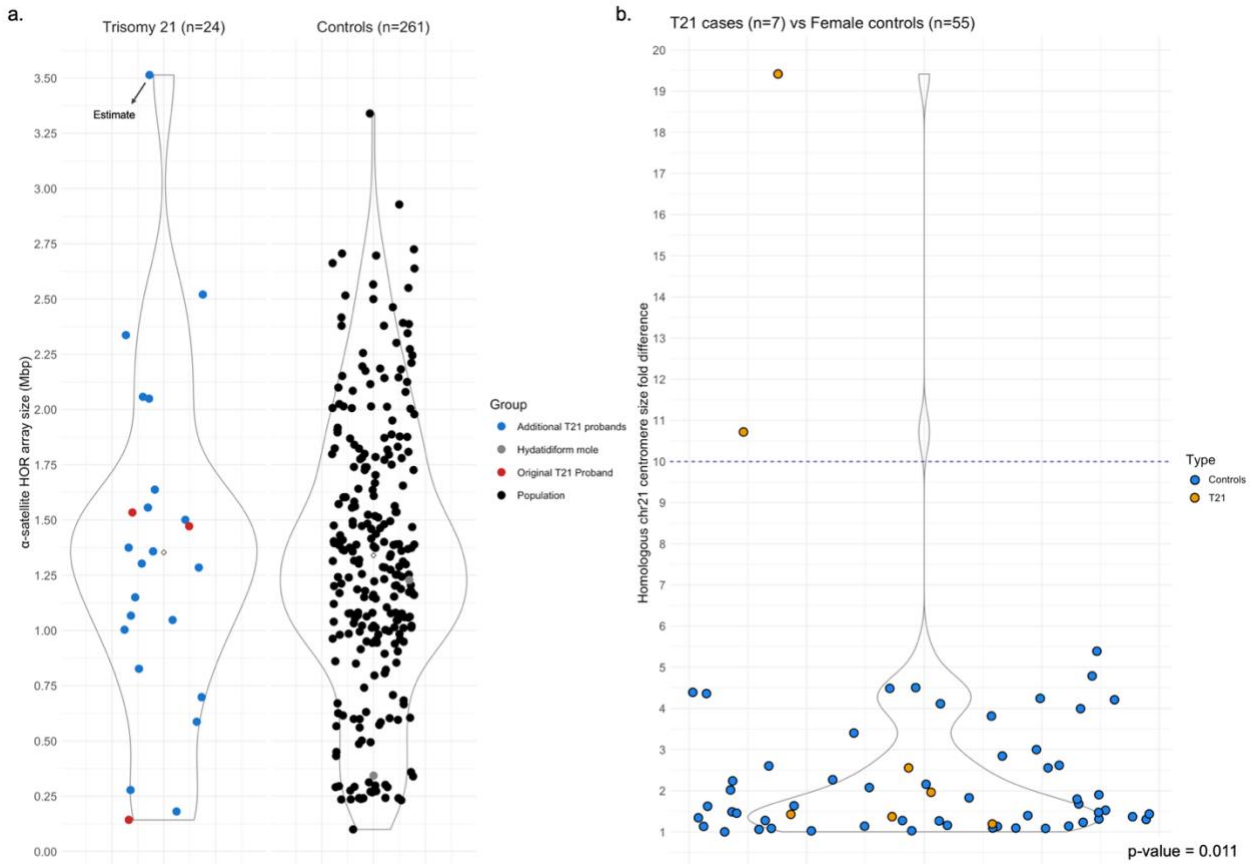


Figure 4. Diversity in α-satellite HOR array length, composition, and CpG methylation profile amongst diverse chr21 centromere haplotypes. a) Maximum-likelihood phylogenetic tree based on a multiple sequence alignment of a 20 kbp segment from the pericentromeric q-arm regions from 57 unique completely sequenced chr21 centromeres. The HOR array structure of each is depicted and estimated divergence time for each cluster is projected onto the nodes. Bootstrap values 95 to 100 (**) are marked with asterisks. Black lines connect homologous chr21 centromeres found in mothers, which were both transmitted to their child with T21. Red lines indicate the homologs exhibiting extreme centromere size asymmetry. **b)** CpG methylation profiles of 16 population haplotypes (see also **Supplementary Figure 9**). **c)** Pairwise sequence identity heatmaps (generated with StainedGlass[33] using 5 kbp windows) across chr21 centromeres enriched with 5-monomer and 4-monomer α-satellite HORs (see legend).

393 **Maternal asymmetry in T21 centromeric α-satellite HOR length**. To establish a baseline for
394 centromere length distributions in the human population, we compared the T21 centromeres to
395 an expanded set of 261 completely assembled chr21 centromeres from control individuals. The
396 control cohort consists of two hydatidiform mole (CHM13 and CHM1) chr21 centromeres, 184
397 chr21 centromeres (92 samples) from the recently released genomes part of the HPRC[31] (year
398 2 release), 40 chr21 centromeres recently published by the HGSVC[30], and an additional 35 from
399 the two consortia characterized with the same methods applied to the families with T21
400 (**Supplementary Table 3**). The centromeres from the HPRC year 2 release were characterized
401 with CenMAP (v0.3.1), a recently released centromere mapping and annotation pipeline
402 (https://github.com/logsdon-lab/CenMAP).
403
404 Comparing the size distribution of all control haplotypes (*n*=261) to those from individuals with
405 T21 (*n*=24), we find no statistically significant difference (t-test p-value = 0.92) (**Figure 5a**).
406 Similarly, the frequency of small centromeres (<500 kbp) also showed no significant difference
407 between the two groups (3/24 in T21 and 27/261 in the general population groups; Fisher's
408 exact test p-value = 0.73) despite an earlier report suggesting the contrary[27]. Notwithstanding,
409 the small centromeres identified in AG16777 and 7210079 families (mothers and children)
410 remain the smallest chr21 α-satellite HOR array sequenced to date in any female individual,
411 measuring 143 kbp and 181 kbp respectively (**Figure 5a**). For the male individuals, only one
412 centromere was found to have a smaller α-satellite HOR array: HG00739-H1, which represents
413 the smallest chr21 centromere ever sequenced with an array of only 100 kbp.
414
415 Next, we considered the diplotype of the transmitting mothers and assayed the frequency of
416 extreme centromere size asymmetry (>10-fold) in the population using 116 control samples
417 (females = 55; males = 61) and seven individuals with T21 for which both centromere
418 haplotypes were fully resolved (the singleton with T21 was excluded since we could not phase

419 the centromeres nor determine if the centromere size remained the same between the two
420 generations; **Supplementary Figure 10**). A significant difference in extreme homologous chr21
421 centromere size asymmetry cases was observed between mothers with children with T21 and
422 the general population. We identified two cases in the cohort of individuals with T21 and none in
423 116 control individuals where homologous centromere HOR length differed more than 10-fold
424 (Fisher's exact test p-value = 0.003) (**Supplementary Figure 10**). Given the fundamental
425 differences in male and female meiosis, we repeated the analysis restricting further to only
426 female individuals from the general population. The number of asymmetrical centromeres
427 remained statistically significant for families with T21 (Fisher's exact test p-value = 0.011)
428 (**Figure 5b**). We further expanded the controls by including another 23 female and 22 male
429 individuals from the HPRC year 2 release that were completely assembled but had small errors
430 in their sequences (e.g., collapses, misjoins, etc.). In these samples, the centromere size
431 estimate is not as accurate, but we do not expect a great difference between the estimate and
432 their actual sizes. Once again, none exhibited extreme centromere asymmetry. If we consider
433 these additional female samples in our comparison, the difference becomes even more
434 statistically significant (Fisher's exact test p-value = 0.006).



435
436 **Figure 5. Comparison between centromere HOR array sizes from individuals with T21 and**
437 **controls. a)** Size distribution comparison of chr21 centromeric α-satellite HOR arrays from individuals
438 with T21 and population samples. This plot includes the HOR array sizes of 24 T21, 2 hydatidiform moles,
439 and 259 population haplotypes. **b)** Homologous chr21 centromere size fold differences in individuals with
440 T21 (only maternally inherited centromeres; orange) and female population samples (blue). P-value
441 indicates a significant difference in extreme asymmetry cases between the two groups.

442 **DISCUSSION**

443 Down syndrome was first described in 1866[39] and its typical T21 karyotype description was
444 made in 1959[40]. Despite extensive research over the last 65 years, very little is still known about
445 the genetic etiology of this condition. To this day, we still do not have a method to assess
446 predisposition to this condition nor are aware of any direct genetic risk factor that leads to free
447 T21. We could only test directly for T21 during pregnancies with invasive procedures, such as
448 amniocentesis until more recent, safer noninvasive prenatal tests emerged[41]. Understanding the
449 genetic architecture of T21 and genetic factors that predispose to it may be another important
450 consideration in its prevention.

451

452 In this study, we present for the first time the complete sequence and epigenetic
453 characterization of chr21 centromeres of eight individuals with free T21, seven of their mothers,
454 and one father. In all families, T21 was caused by MMIEs and all mothers were <35 years old at
455 childbirth. We initially aimed to investigate chr21 centromere variability in individuals with T21
456 and test the hypothesis that individuals with the syndrome carry small centromeres. Previous
457 studies postulated that "centromere strength" could be associated with the amount of
458 centromeric DNA sequence[20–23]. Other independent reports revealed a statistically significant
459 higher number of small chr21 α-satellite HOR arrays in individuals with the syndrome compared
460 to controls[27], suggested an association of small chr21 α-satellite HOR arrays and MMIEs[42], and
461 proposed that individuals with T21 generally harbor small chr21 centromeres[27]. Nevertheless, a
462 2003 study explored the relationship between the size of α-satellite HOR arrays on chr21
463 centromeres and repressed recombination, one of the known phenomena that correlate with
464 nondisjunction, and concluded that the features are independent[43].

465

466 In our investigation, chr21 centromeres showed remarkable variability in length, α-satellite HOR
467 array structure, and CDR patterns within the human population (**Figure 4**), suggesting that

468 chr21 is amongst the top four most variable centromeres in the human genome[7]. We show that
469 small centromeres are found at low frequency in individuals with T21 (3/24; 3 centromeres
470 smaller than 500 kbp in 3 individuals, 2 of which of maternal origin) and that T21 centromere
471 size distribution is comparable to that of the general population (**Figure 5a**). Despite being
472 present at low frequency in the general population (10%), our data do not support the
473 observation that small centromeres alone constitute a risk factor for T21 nondisjunction. Rather,
474 our results suggest an alternative hypothesis, where centromere size asymmetry may be a
475 more important consideration and possibly one of the features that contribute to suppressed
476 recombination and nondisjunction.
477
478 Small centromeres might have been previously reported in children with T21, as they make size
479 asymmetry more likely than larger centromeres. Our results support other observations[44], which
480 noted that small centromeres are not exclusive of individuals with T21; contrarily, extreme
481 homologous chr21 centromere size asymmetry (>10-fold) appears to be absent in individuals
482 without the syndrome (**Figure 5b; Supplementary Figure 10**). We report two cases in families
483 that include an individual with T21 with similar features, but none in the general population.
484 Moreover, our phylogenetic analysis (**Figure 4a**) shows that rapid changes of chr21 centromere
485 size are possible and might facilitate the occurrence of centromere size asymmetry in the
486 population. We speculate that an extreme difference between chr21 centromere sizes makes
487 homologous pairing less efficient for example by complicating meiotic synapsis across the entire
488 centromere length. There is evidence demonstrating that larger domains of homology are
489 essential for synapsis formation and/or recombination in complex organisms such as *C.elegans*,
490 *Drosophila*, and mice but not in simpler organisms like yeast[45–49]. Nevertheless, more
491 experimental data are needed to test how asymmetry impacts meiosis, recombination, and
492 chromosome segregation.
493

494 The families (AG167 and 7210079) showing extreme centromere size asymmetry (10.7- and
495 19.4-fold) have other features in common, suggesting a common T21 aetiology. For instance,
496 they carry the smallest chr21 centromeres ever characterized in females (143 and 181 kbp),
497 both exhibit a transgenerational CDR change on the longer homologous centromere, and both
498 mothers had a similar age at time of childbirth (29 and 27 years old). Another family (7210223)
499 exhibited a transgenerational CDR change. In this case, we do not observe an extreme
500 centromere size asymmetry (2.6-fold), and the methylation change occurred on the smaller
501 haplotype (587 kbp). Mother 7210223.20 was 22 at childbirth and we did not observe any other
502 unusual centromeric features, thus suggesting an alternative mechanism underlying T21.
503
504 In this study we also investigated the amount of centromeric proteins forming the kinetochore
505 complex on each of the chr21 centromeres of family AG167. Our IF-FISH data (**Supplementary**
506 **Figure 3**) suggest that none of the chr21 centromeres have a significantly lower amount of
507 CENP-A or CENP-C proteins, arguing against the hypothesis that MMIEs are caused by
508 malfunctions in the kinetochore complex. It is unclear how the kinetochore function is impacted
509 by the altered CDRs observed in three of the children with T21 and we cannot determine
510 whether the poorly defined CDRs were already present in the maternal germ cells or developed
511 later in the children after establishment of T21.
512
513 The first family with T21 (AG167) centromere showed epigenetic transgenerational differences
514 between maternal and child centromeres in the absence of genetic changes. Specifically, we
515 observed significant differences in CDR size of child's maternal H2 and paternal H3 CDRs as
516 well as a positional shift of ~11 kbp and 25 kbp towards the p-arm in the child for the former and
517 the latter, respectively (**Figure 2c**), despite the fact that the sequence and structure between
518 parent and child were practically identical (**Supplementary Figure 5**). This reinforces the notion
519 that CDRs and kinetochore position are not defined by the primary sequence but can change

520 location potentially even within a single generation. This is also supported by two population
521 haplotypes, HG03710-H1 and HG02647-H2, which both carry their CDRs closer to the p-arm
522 yet differ radically in their HOR array monomers with 5-monomers and 4-monomers
523 corresponding to older and younger repeats, respectively.
524
525 The second cohort of families with T21 also allowed us to compare CDR differences between
526 cell lines and a primary tissue like blood. Little is known regarding the stability of CDRs
527 somatically; however, our limited data on multiple samples demonstrate that CDRs position
528 between blood and cell lines remain positionally conserved. The only detected difference is a
529 generally higher level of methylation in the blood. This could be due to multiple reasons. First,
530 comparison of methylation callers (**Supplementary Figure 8**) on cell line data indicates that
531 PacBio basecallers (Jasmine and especially the older Primrose) detect a slightly higher level of
532 methylation. Also, HiFi reads might be mapping less precisely than the much longer UL-ONT
533 reads. Lastly, it is possible that some highly specialized cells in the blood, which are generated
534 from the progenitors in the bone marrow and do not typically undergo mitosis[50], might lose
535 epigenetic signatures over time like those associated with CDRs.
536
537 There are several limitations in this study. First and foremost, the observation of asymmetry
538 rests on relatively few observations in part because the sequence and assembly of centromeres
539 is still a costly and time-consuming endeavor. Moreover, we focused here on individuals with
540 T21 conceived by younger mothers (age range: 20-34) despite the known, strong association of
541 nondisjunction with advanced maternal age. It will be important to pull apart these confounders
542 by assessing centromere asymmetry in even younger and older mothers and incorporating the
543 maternal age effect. Larger T21 population-based studies are, therefore, required to confirm this
544 finding as well as evaluating families with MMIEs and MMIIEs. Large cohorts with short-read
545 sequence data, such as the Atlanta or the National Down Syndrome Projects[36,37], have been

546 assembled and imputation of chr21 centromere length in these cohorts once a large enough
547 pangenome has been amassed may be possible instead of direct sequencing. Finally, future
548 studies should evaluate not only centromere sequence and methylation using long-read
549 sequencing data but also evaluate the composition and location of the kinetochore complex as it
550 relates to epigenetic differences of the CDR.
551

552 **Acknowledgments**

553 This work was supported, in part, by US National Institutes of Health (NIH) grants
554 R01HG010169 and U24HG007497 (to E.E.E.) and R00GM147352 (to G.A.L.). E.E.E. is an
555 investigator of the Howard Hughes Medical Institute.
556

557 This article is subject to HHMI's Open Access to Publications policy. HHMI lab heads have
558 previously granted a nonexclusive CC BY 4.0 license to the public and a sublicensable license
559 to HHMI in their research articles. Pursuant to those licenses, the author-accepted manuscript
560 of this article can be made freely available under a CC BY 4.0 license immediately upon
561 publication.
562

563 **COI (Conflicts of Interest) Statement**

564 E.E.E. is a scientific advisory board (SAB) member of Variant Bio, Inc. The other authors
565 declare no conflicts of interest.

566 **MATERIAL AND METHODS**

567 **Sample recruitment and data collection.** The following lymphoblastoid cell lines were
568 obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical
569 Research: proband (AG16777), mother (AG16778), and father (AG16782). The proband is
570 described as a mosaic with karyotype: 47,XX,+21[21]/47,XX,+21,t(21;22)(q22;q13)[29] with 10%
571 of the cells examined showing random chromosome loss and the karyotype confirming the
572 diagnosis. Parents do not have Down syndrome and have a normal karyotype. The proband
573 was 43 years old at sampling, while mother and father were 72 and 74, respectively. We
574 obtained lymphoblastoid cell lines from Emory University biobank for seven probands with
575 MMIE T21 and six mothers. Blood samples were also obtained for the same probands.
576

577 **PacBio HiFi sequencing.** PacBio HiFi data were generated per manufacturer's
578 recommendations. Briefly, high-molecular-weight (HMW) DNA was extracted from frozen blood
579 or cultured lymphoblasts using the Monarch® HMW DNA Extraction Kit for Cells & Blood (New
580 England Biolabs, T3050L). At all steps, quantification was performed with Qubit dsDNA HS
581 (Thermo Fisher, Q32854) measured on DS-11 FX (Denovix) and size distribution checked using
582 FEMTO Pulse (Agilent, M5330AA & FP-1002-0275.) HMW DNA was sheared with Megaruptor 3
583 (Diagenode, B06010003 & E07010003) using settings 28/31 or 28/30 (depending on original
584 length distribution) and used to generate PacBio HiFi libraries via the SMRTbell Prep Kit 3.0
585 (PacBio, 102-182-700). Size selection was performed with Pippin HT using a high-pass cutoff of
586 17 kbp (Sage Science, HTP0001 & HPE7510.) Samples AG16778, AG16782, and NG16777
587 were sequenced on the Sequel II platform on SMRT Cells 8M (PacBio, 101-389-001) using
588 Sequel II Sequencing Chemistry 3.2 (PacBio,102-333-300) with 2-hour pre-extension and 30-
589 hour movies, aiming for a minimum estimated coverage of 30× in PacBio HiFi reads for parents
590 (3 SMRT Cells per sample) and 60× for the proband (5 SMRT Cells, assuming a genome size
591 of 3.1 Gbp). The remaining samples were sequenced on the Revio platform on SMRT Cells

592 25M (PacBio, 102-817-900) with diffusion loading and 24-hour movies or Adaptive Loading and
593 30-hour movies, aiming for a minimum estimated coverage of 30× in PacBio HiFi reads for the
594 parents (1-1.5 SMRT Cells per sample) and 60× for the probands (2-2.5 SMRT Cells, assuming
595 a genome size of 3.1 Gbp).
596
597 **UL-ONT sequencing.** Ultra-high molecular weight gDNA was extracted from the
598 lymphoblastoid cell lines according to a previously published protocol. Briefly, 3-5 × 10^7 cells
599 were lysed in a buffer containing 10 mM Tris-Cl (pH 8.0), 0.1 M EDTA (pH 8.0), 0.5% w/v SDS,
600 and 20 mg/mL Rnase A for 1 hour at 37°C. 200 ug/mL Proteinase K was added, and the
601 solution was incubated at 50°C for 2 hours. DNA was purified via two rounds of 25:24:1 phenol-
602 chloroform-isoamyl alcohol extraction followed by ethanol precipitation. Precipitated DNA was
603 solubilized in 10 mM Tris (pH 8.0) containing 0.02% Triton X-100 at 4°C for two days. Libraries
604 were constructed using the Ultra-Long DNA Sequencing Kit (ONT, SQK-ULK001 or SQK-
605 ULK114) following the manufacturer's protocol. 75 uL of library was loaded onto a primed FLO-
606 PRO002 R9.4.1 or FLO-PRO114M R10.4.1 flow cells for sequencing on the PromethION, with
607 two nuclease washes and reloads after 24 and 48 hours of sequencing. Sequence reads >100
608 kbp in length were classified as ultra-long (UL). All ONT data were base called using guppy
609 (v6.3.7 or newer) or dorado (v0.4.2 or newer) with the SUP model.
610
611 **Genome assembly.** Proband AG16777 genome assembly was generated with the Verkko[32]
612 (v1.2) (https://github.com/marbl/verkko) workflow using HiFi sequencing data at ~59× and UL-
613 ONT (>100 kbp) data at ~23× coverage following the developer instructions. Parental
614 assemblies (AG16778 and AG16782) were generated with both Verkko[32] (v1.4)
615 (https://github.com/marbl/verkko) and hifiasm[51] (v0.19.5) (https://github.com/chhylp123/hifiasm)
616 using HiFi data at ~37× for the mother and 40× for the father and UL-ONT data at 25× and 12×,
617 respectively. HiFi developer instructions were followed. Similarly, genome assemblies for the

618 second cohort of probands with T21 and their mothers were generated using Verkko[32] (2.0 or
619 newer) and hifiasm[51] (v0.19.8 or newer).
620
621 **Cell culture, slide preparation and Immunofluorescence (IF) combined with fluorescence**
622 ***in situ* hybridization (FISH).** AG16777, AG16778 and AG16782 cells were cultured in RPMI-
623 1640 medium supplemented with 16% fetal bovine serum, 1% L-glutamine, and 1% penicillin–
624 streptomycin. Metaphase and interphase nuclei undergoing immunofluorescence with mouse
625 CENP-A monoclonal antibody (Enzo, ADI-KAM-CC006-E) were prepared by blocking the cells
626 in metaphase with colcemid, followed by washing in phosphate-buffered saline (1× PBS). The
627 cells were then counted and resuspended in a hypotonic buffer (75 mM KCl : 0.8% NaCitrate : 3
628 mM CaCl2 : 1.5 mM MgCl2, 1:1:1 ratio) for 15 minutes to reach a final concentration of 5 x 10^5
629 cells/mL. A 0.5 mL aliquot of the cell suspension was then cytocentrifuged onto Superfrost Plus
630 slides at 1,500 rpm for 5 minutes using a Shandon Cytospin 3 (Thermo Scientific). The slides
631 were immediately used for immunofluorescence. For slides undergoing immunofluorescence
632 with mouse CENP-C monoclonal antibody (Abcam, ab50974), colcemid-treated cells were
633 incubated in hypotonic solution (0,56% KCl) for 30 min and prefixed by adding methanol:acetic
634 acid (3:1) fixative solution. Prefixed cells were collected by centrifugation and then fixed in
635 methanol:acetic acid (3:1). Spreads were dropped on a glass slide and incubated at RT
636 overnight.
637
638 Immunofluorescence was performed on the metaphase spreads using mouse CENP-C and
639 CENP-A monoclonal antibodies as previously described with minor modifications[52]. In brief,
640 each slide was rehydrated by immersion in 1× PBS-azide (10 mM NaPO4, pH 7.4, 0.15 M NaCl,
641 1 mM EGTA and 0.01% NaN3) for 15 min at room temperature. Chromosomes were then
642 swollen by washing the slides (three times, 2 min each) with 1× TEEN (1 mM triethanolamine-
643 HCl, pH 8.5, 0.2 mM NaEDTA, and 25 mM NaCl), 0.5% Triton X-100 and 0.1% BSA. The

644 primary antibodies were diluted to a final concentration of 0,001 μg/μl in the same solution and
645 then added (100 μl) onto the slides. Each slide was incubated for 2 h at 37 °C. Excess of primary
646 antibody was removed by washing the slides at room temperature (three times, 2, 5 and 3 min
647 each) with 1× KB buffer (10 mM Tris-HCl, pH 7.7, 0.15 M NaCl and 0.1% BSA). A goat anti-
648 mouse IgG secondary antibody conjugated to FITC (Abcam, ab6785) was diluted 1:100 in the
649 same solution, and 100 μl was then added to the slides that were then incubated for 45 min at
650 37 °C in a dark chamber. After incubation with the secondary antibody, the slides were washed
651 once with 1× KB for 2 min, prefixed with 4% paraformaldehyde in 1× KB for 45 min at room
652 temperature, washed with dH2O by immersion for 10 min at room temperature, and fixed with
653 methanol and acetic acid (3:1) for 15 min. FISH was then performed using a plasmid containing
654 the α-satellite sequence shared between chromosomes 13 and 21 (pZ21A) directly labelled by
655 nick-translation with Cy3-dUTP (Enzo, 42501) according to a standard procedure with minor
656 modifications[53]. In brief, 100 ng of labelled probe was used for the FISH experiments; DNA
657 denaturation was performed at 70 °C for 8 min and hybridization at 37 °C in 2× SSC, 50% (v/v)
658 formamide, 10% (w/v) dextran sulphate and 3 mg sonicated salmon sperm DNA, in a volume of
659 10 μl. Post-hybridization washing was performed under high stringency conditions: at 60 °C in
660 0.1× SSC (three times, 5 min each). Nuclei and chromosome metaphases were simultaneously
661 DAPI-stained. Digital images were obtained using a Leica DMRXA2 epifluorescence
662 microscope equipped with a cooled CCD camera (Princeton Instruments). DAPI, Cy3 and
663 fluorescein fluorescence signals, detected with specific filters, were recorded separately as
664 grayscale images. Pseudocolouring and merging of images were performed using ImageJ
665 (v.1.53k). The CENP-C fluorescence intensity at chr21 centromeres was measured on 13
666 metaphases for each sample using a modified version of the ImageJ macro, CRaQ[54]. Signals
667 from chr21 were isolated, and a 7 × 7 pixel box was placed around the centroid position of each
668 signal. The pixel intensity within each box was measured and normalized dividing it by the lower

669 chr21 signal intensity for each analyzed metaphase. Plots were then obtained using the
670 ggplot2[55] package in R[56].
671
672 **Centromere characterization and validation.** Assemblies of individuals with T21 and from the
673 general population were aligned against the T2T-CHM13[8] (v2.0) reference genome with
674 minimap2[57] (v2.24) (https://github.com/lh3/minimap2) using the following options "-t {threads} -I
675 15G -a –eqx -x asm20 -s 5000". Fasta sequences from the contigs mapping against chr21
676 centromere (chr21:10700000-11650000) and flanking regions were extracted. RepeatMasker[58]
677 (v4.1.0) was used to identify the regions containing the α-satellite array (marked as
678 "ALR/Alpha'') and HumAS-HMMER (https://github.com/fedorrik/HumAS-HMMER_for_AnVIL) to
679 determine the α-satellite HOR array composition with the following command: hmmer-run.sh
680 {path_to_directory_with_fasta} AS-HORs-hmmer3.0-170921.hmm {threads}. For the contigs
681 aligning to the reference genome in the minus orientation, we converted the fasta sequences to
682 their reverse complement before running the analysis. The outputs from HumAS-HMMER
683 analyses were visualized with R[56] (v4.2.2) using the ggplot2[55] package.
684
685 Another 44 population chr21 centromere sizes were obtained from the recently published
686 HGSVC data[35]. Additionally, 115 female and 101 male samples from the HPRC year 2 release
687 were processed and validated with CenMap (v0.3.1) (https://github.com/logsdon-lab/CenMAP),
688 a newly developed pipeline for centromere mapping and annotation, to calculate their chr21
689 centromere sizes.
690
691 We validated each centromeric contig sequence aligning each sample's HiFi reads to their own
692 genome assembly using pbmm2 (v1.1.0) (https://github.com/PacificBiosciences/pbmm2) and
693 showing uniform read coverage across the centromeric region with NucFreq[59]
694 (https://github.com/mrvollger/NucFreq). To confirm that the extracted centromeric contig

695 sequences represented chr21 centromere repeats, we compared them to the α-satellite
696 monomer units extracted from the T2T-CHM13[8] reference genome using StringDecomposer[60]
697 (https://github.com/ablab/stringdecomposer) and showed they have high identity (>90%).
698 Finally, we generated pairwise sequence identity heatmaps of each centromere using
699 StainedGlass as previously described[33] (https://github.com/mrvollger/StainedGlass).
700
701 **Supplementary note**. We were not able to characterize both chr21 centromeres in 35 females
702 and 30 males from the HPRC year 2 release. In these samples, one or none of the chr21
703 centromeres was assembled, or the haplotypes could not be reliably assigned to chr21 or chr13,
704 which typically has much larger centromeres. This is common when the genomic contigs
705 containing the centromeric sequences are not sufficiently extended to encompass the
706 pericentromeric regions of both p- and q-arms and highlights the difficulty of sequencing and
707 assembling large centromeres. As an example, two female samples (HG00320 and HG04184)
708 carried a small chr21 centromere (228 and 263 kbp), but their putative homologs were either
709 incomplete or contained errors, with ambiguous assignments to chr21 or chr13. Some of these
710 contigs contained large α-satellite arrays, but despite further analysis, we were unable to
711 conclusively determine whether these putative homologs represented chr21, chr13 or a mix of
712 the two centromeres. Consequently, they were excluded. It is worth noting though that even with
713 two cases of extreme size asymmetry in the female population cohort, we would still have a
714 significant difference in the number of cases observed in the two groups (population samples =
715 80; Fisher's exact test p-value = 0.031).
716
717 **Centromere alignments and dotplots.** AG16777 proband and parental centromere sequences
718 were aligned using minimap2[57] (v2.26) and the following command: minimap2 -cx asm5
719 {proband_centromeric_fasta} {corresponding_parental_centromeric_fasta} > alignment.paf. The
720 results in .paf format were visualized with R[56] (v4.2.2) using the ggplot2[55] package. Proband H3

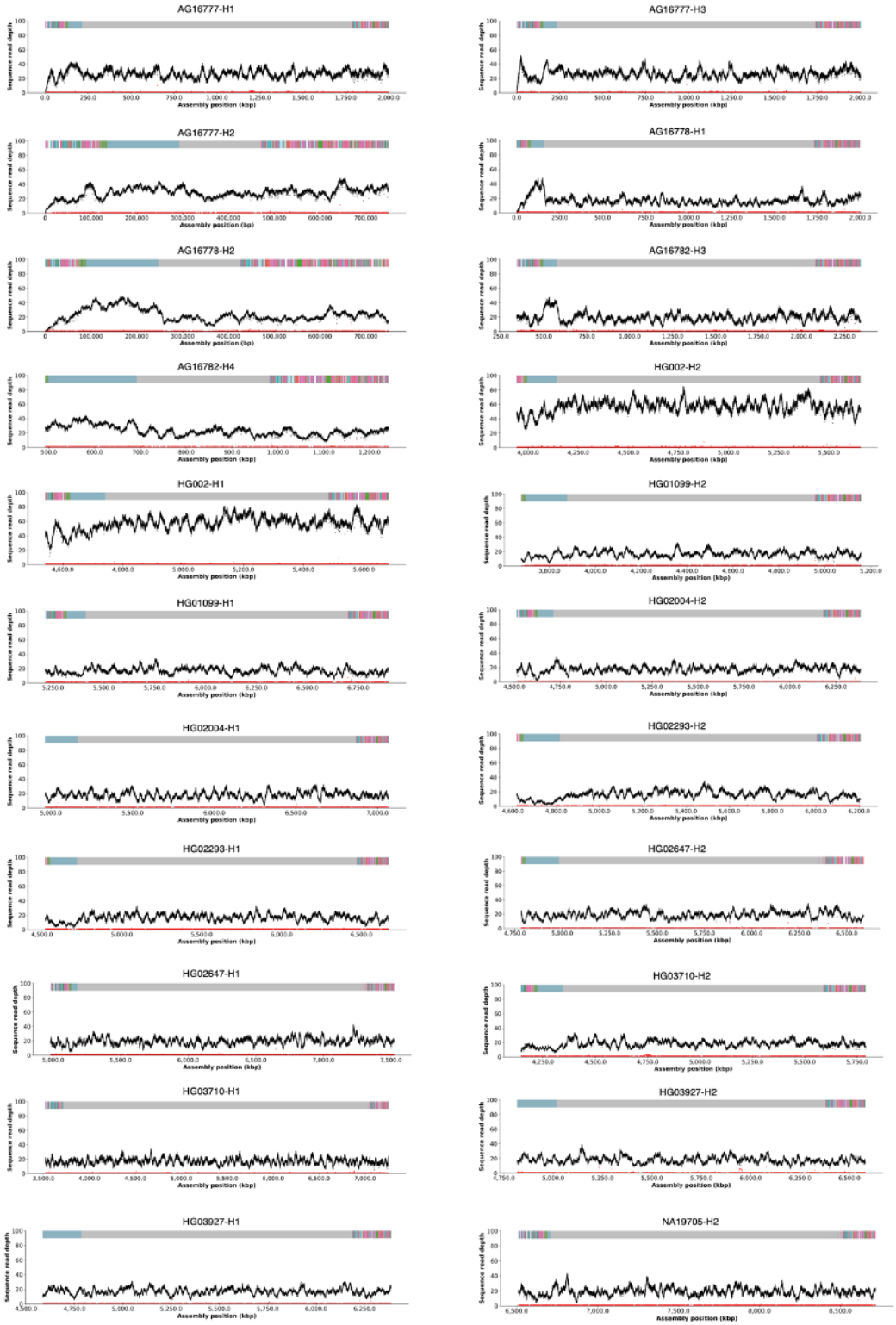
721 and father H3 AluY centromeric insertions with 10 kbp padding sequence on both sides were
722 aligned with BLAST[61].
723
724 **CpG methylation and CDR analysis.** All ONT sample data were basecalled with guppy v6.3.7
725 or newer to generate BAM files containing methylation tags linked to each read. HiFi data were
726 base called using CCS (v.6.4.0 or newer) (https://github.com/PacificBiosciences/ccs), Primrose
727 (v1.4.0 or newer) or (Jasmine (v2.0.0 or newer) (https://github.com/PacificBiosciences/jasmine),
728 which generated the equivalent BAM files containing the methylation tags linked to each read.
729 Jasmine 5mC predictions were used as a validation for the centromeric methylation profiles.
730 Briefly, we extracted the reads for each sample and aligned them to their respective genome
731 assembly using Winnowmap[62] (v2.03 or newer). We then used modbam2bed
732 (https://github.com/epi2me-labs/modbam2bed) to obtain bedMethyl files.
733
734 Using the bedMethyl files as an input and an early developer version of CDR-Finder[17], we
735 identified the CDRs for each centromere. The tool first divides the region of interest (the
736 centromeric α-satellite arrays with padding sequence on both sides, e.g., coordinates for
737 Proband H1: unassigned-0003018:0-2000000) in sequential 5 kbp bins and calculates their
738 average methylation frequency (if a bin does not have an assigned value, the bin is excluded). It
739 then computes the median CpG methylation frequency for each window containing α-satellite
740 sequences (based on RepeatMasker[58] output). It selects the bins with a lower methylation
741 frequency than the median of the region and merges consecutive bins into candidate CDRs.
742 These candidate CDRs are then evaluated on two criteria: whether or not they have flanking
743 windows on either side with greater than the maximum percent methylation (defined as within 1
744 standard deviation of the maximum) for the entire region and if they clear a user-specified size
745 threshold (25 kbp). Each call is annotated with a confidence label: “high confidence” if it clears
746 both criteria or “low confidence” if it does not. The low-confidence labels are appended with

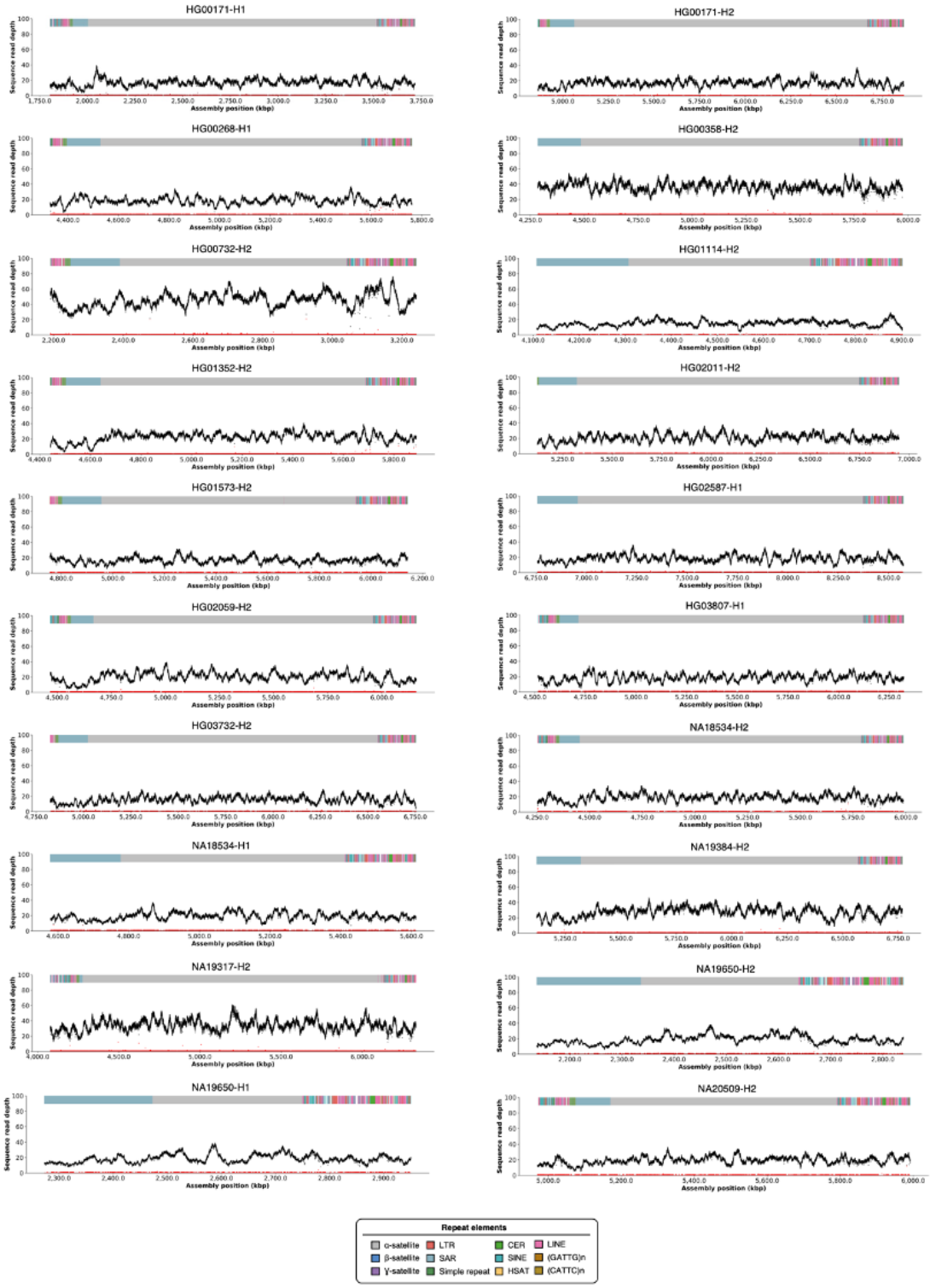
747 either “neighbor_peaks” for the first criterion and/or “size” for the second criterion. In small
748 centromeres where the median can be quite low due to the CDR taking up the majority of the
749 centromeric sequence space, the initial search algorithm of taking windows which are below the
750 median is flipped to identify the windows that are above the median and thus not part of the
751 CDR. After removing these windows from consideration, what remains are candidate CDRs,
752 and the algorithm continues as normal. All CDR calls were manually confirmed inspecting the
753 read alignments. False positives at the boundaries of the centromeres as well as calls not
754 confirmed by the reads and due to low coverage were excluded.
755
756 For each AG167 family CDR, we calculated the percentage of methylated CpGs (defined as
757 having ≥25% methylated CpGs at the same position across all UL-ONT reads) within the total
758 CDR window (from start of the first CDR window to the end of the last CDR window). Formula =
759 (number of methylated CpGs within the CDR window) * 100 / (number of CpGs within the CDR
760 window).
761
762 **Contig fixing and patching.** Centromeric contigs that showed misassemblies or collapses
763 during NucFreq[59] analysis (above) were corrected following this procedure. Sequencing reads
764 mapping to the misassembly/collapse and flanking regions were extracted and locally
765 assembled using the module Canu from the tool HiCanu[63]. The local assemblies were validated
766 using NucFreq[59] as described previously and inserted in the original assembly.
767 Verkko[32] genome assemblies containing errors were corrected with an alternative assembly
768 generated via hifiasm[51] or vice versa. Briefly, we generated assemblies of each genome with
769 hifiasm[51] (v0.19.5 or newer) and the following command: “hifiasm -o {output} -t {threads} –ul
770 {ultra-long read fastq} {hifi reads fastq}”. Then, we aligned the hifiasm assemblies to the Verkko
771 ones with minimap2[57] (v2.24) and the following command: minimap2 -t {threads} -I 15G -a –eqx
772 -x asm20 -s 5000. We identified the corresponding locations of the centromeric regions in each

773 assembly and then determined the α-satellite HOR array composition of each with HumAS-
774 HMMER (https://github.com/fedorrik/HumAS-HMMER_for_AnVIL) and the following command:
775 hmmer-run.sh {path_to_directory_with_fasta} AS-HORs-hmmer3.0-170921.hmm {threads}. We
776 compared the α-satellite HOR compositions of each array and then replaced erroneous
777 sequences in the Verkko assembly with the corresponding sequence in the hifiasm assembly
778 (or vice versa). To ensure that we had resolved the assembly error, we aligned PacBio HiFi
779 reads from the same genome to the new assembly and then ran NucFreq[59] to validate the
780 assembly.
781
782 **Phylogenetic analysis.** For the phylogenetic reconstruction of chr21 centromere haplotypes,
783 we first aligned a chimpanzee assembly to the T2T-CHM13[8] reference genome using
784 Minimap2[57] with the following options "-t {threads} -I 15G -x asm5 -a –secondary=no". We
785 identified a chr21 pericentromeric region on the q-arm where only two chimpanzee contigs
786 mapped (chr21:11855000-11875000). The chimpanzee alignment was used as an outgroup for
787 the analysis. We aligned all the human assemblies available against the T2T-CHM13[8] reference
788 genome as described above and selected only the samples with chr21 centromeric contigs
789 spanning the selected pericentromeric window. We then extracted the sequences mapping to
790 the pericentromeric region of interest of each sample. Using these sequences, we generated a
791 multi-alignment with the tool MAFFT[64,65] (v7.453) using these options: "–auto –thread {threads}".
792 The maximum-likelihood phylogeny tree was constructed with IQ-TREE[66] (v2.1.2) with the
793 following options: "-T {threads} -m MFB -B 1000". Divergence time amongst haplotypes were
794 estimated with IQ-TREE[66] (options: "–keep-ident –date-tip 0 –date-ci 100 -redo") using the
795 MAFFT multi-alignment, the consensus tree, and a time estimate file containing the known
796 divergence time between human and chimpanzee (6.4 million years) as inputs. The resulting
797 tree was visualized with iTOL[67] and FigTree (https://github.com/rambaut/figtree).
798

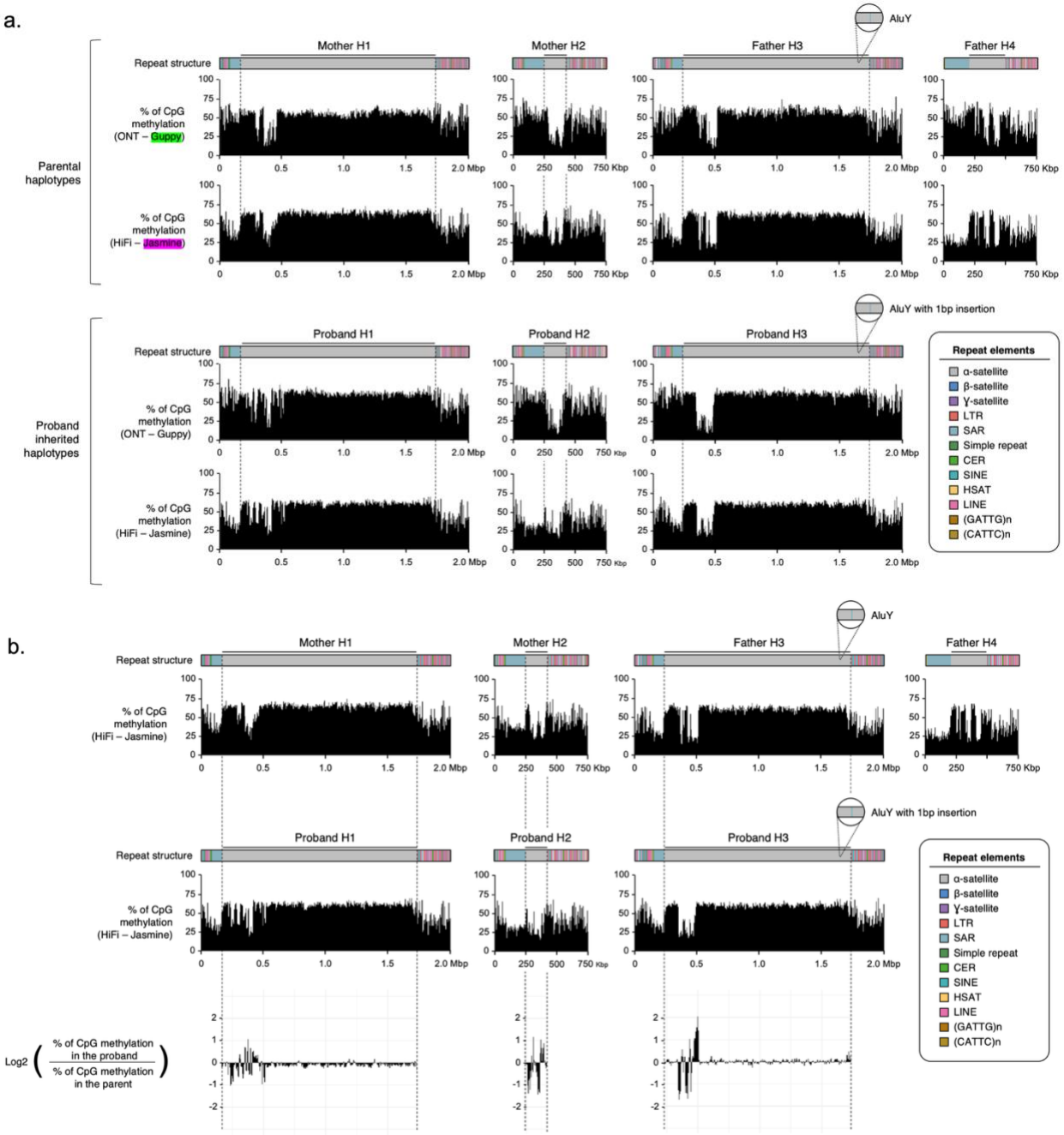
799 **Data availability.** Sequencing data were submitted to the database of Genotypes and
800 Phenotypes (dbGaP) for general research use and public access. The phs accession ID is:
801 phs003761.v1.p1.

Supplementary Figures



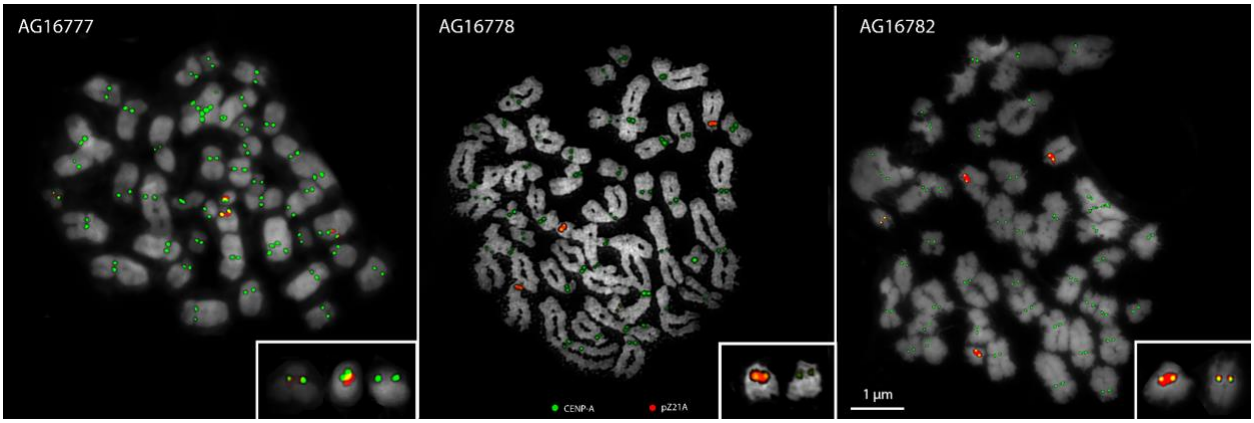


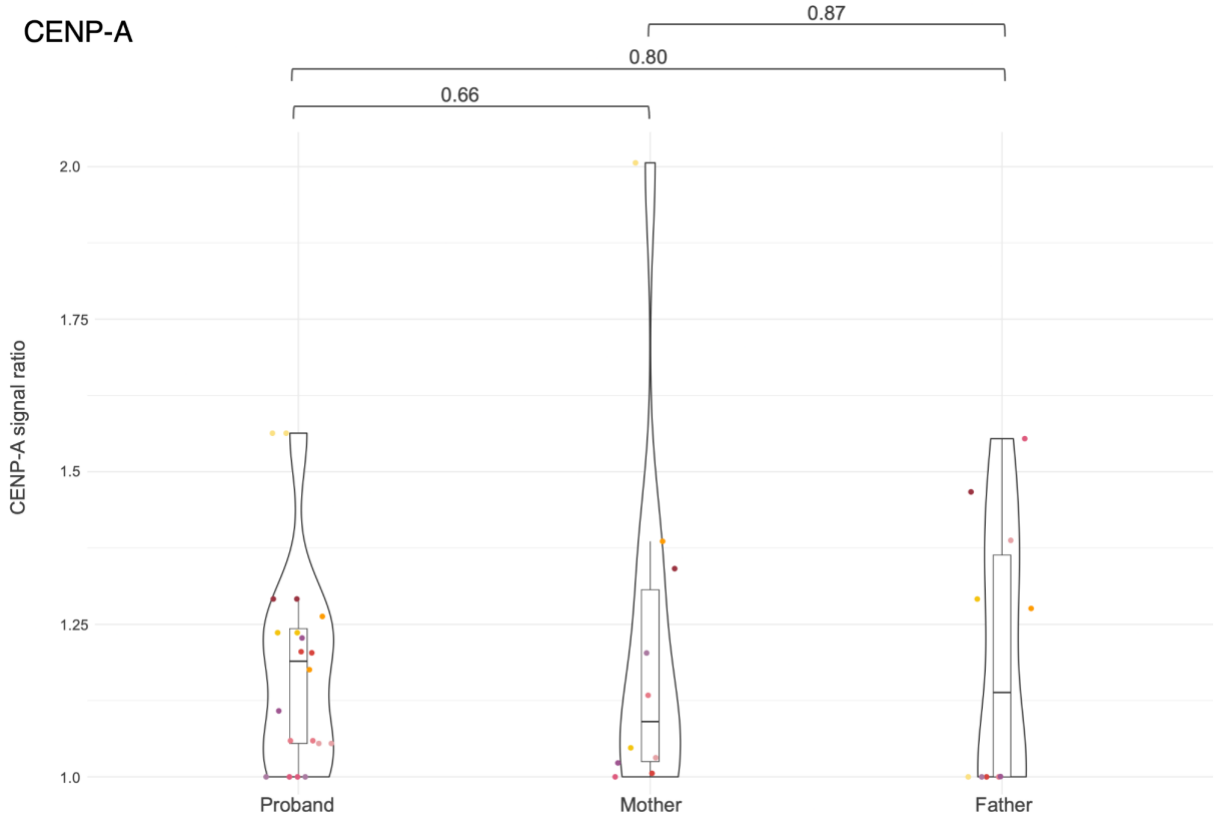
Supplementary Figure 1. NucFreq validation of all the chromosome 21 (chr21) centromeres described in the family AG167 and population samples.



Supplementary Figure 2. a) Comparison of the ONT- and HiFi-based methylation profiles in all family AG167 chr21 centromeres. **b)** Comparison between proband and parental chr21 centromere methylation profiles based on the HiFi data.

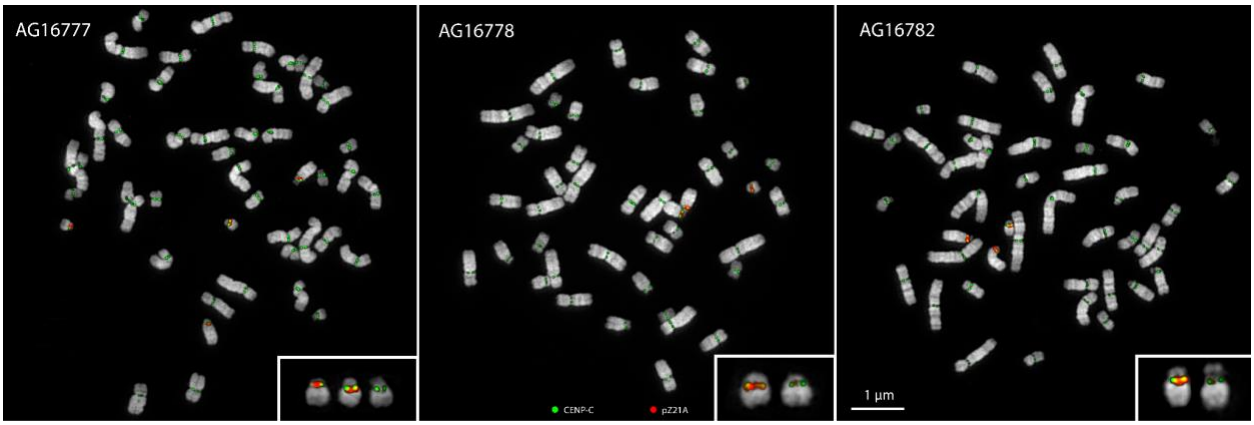
812

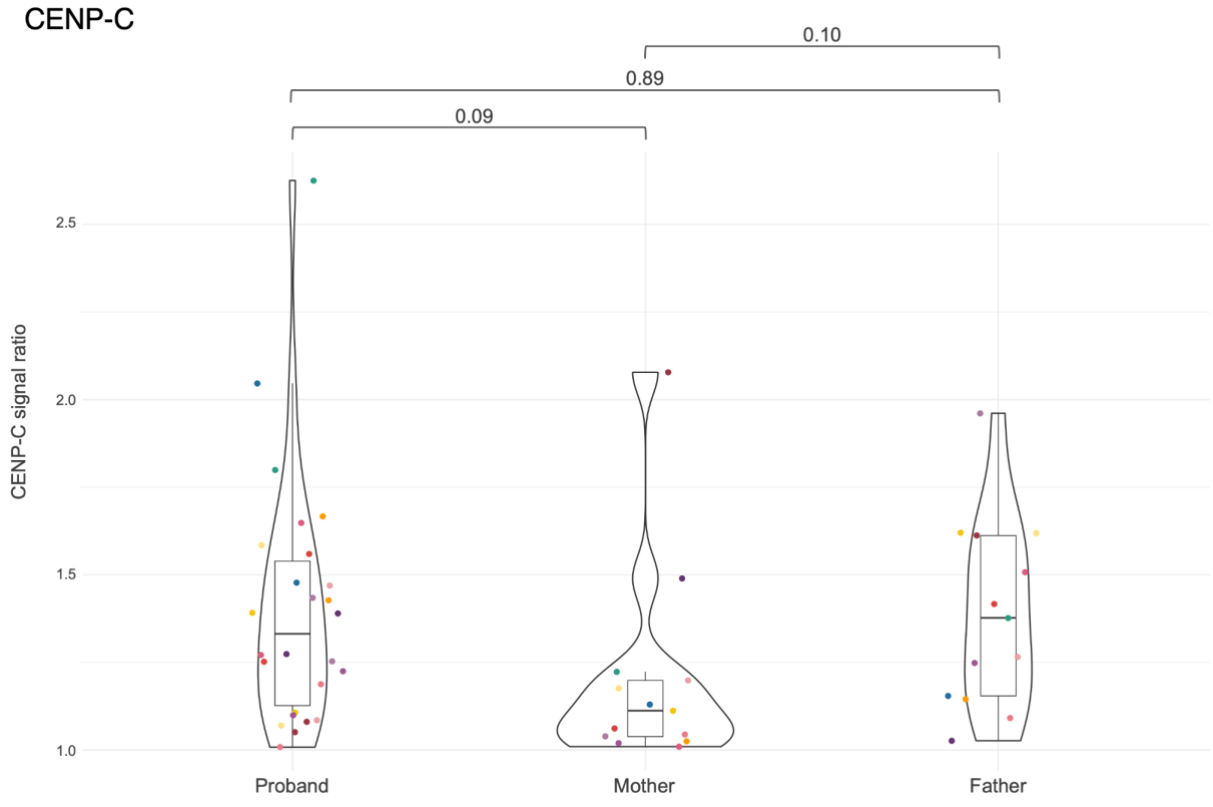

AG16777
AG16778
AG16782
CENP-A
pZ21A
1 µm


CENP-A
0.87
0.80
0.66
CENP-A signal ratio
2.0
1.75
1.5
1.25
1.0
Proband
Mother
Father

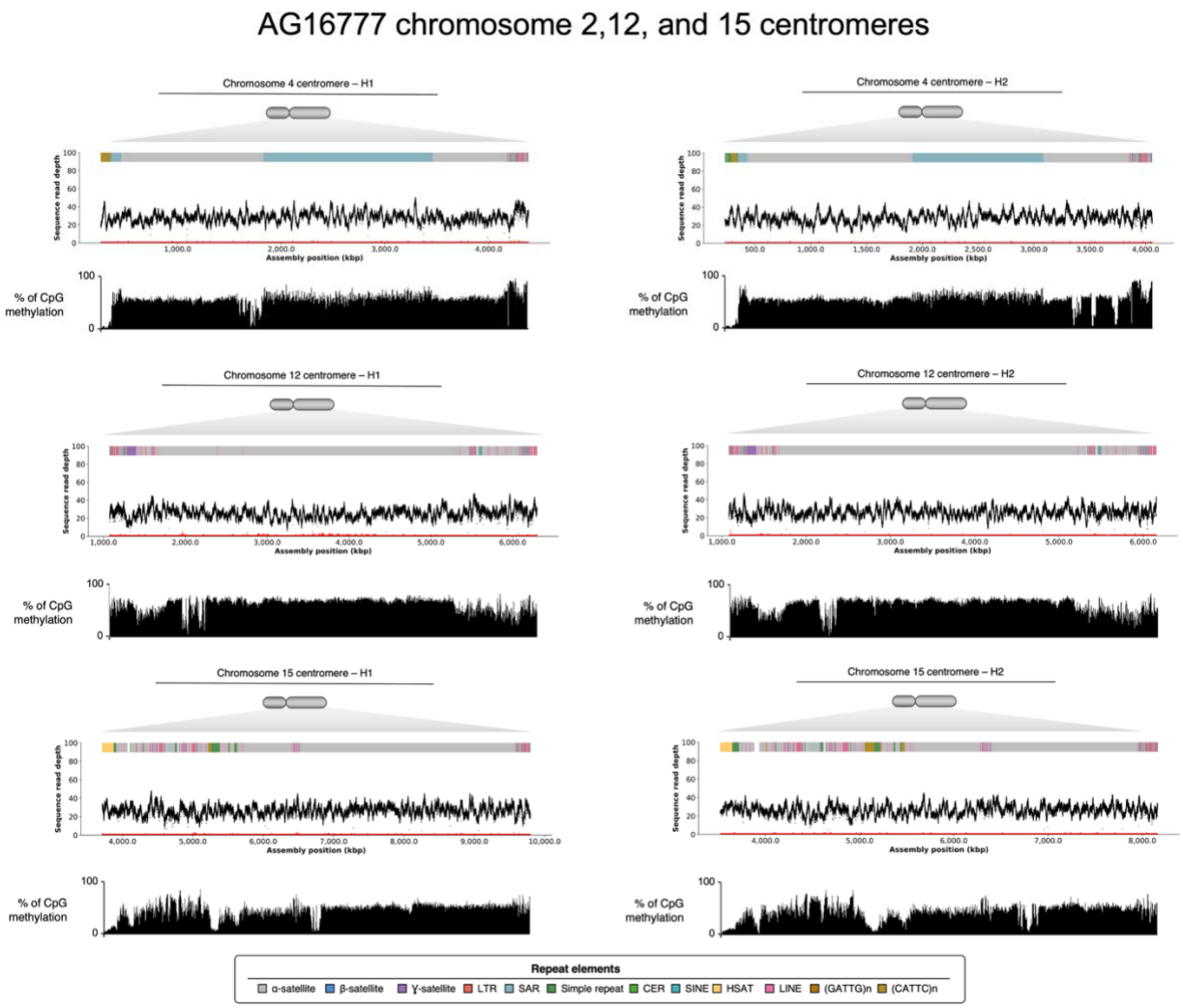
813

814

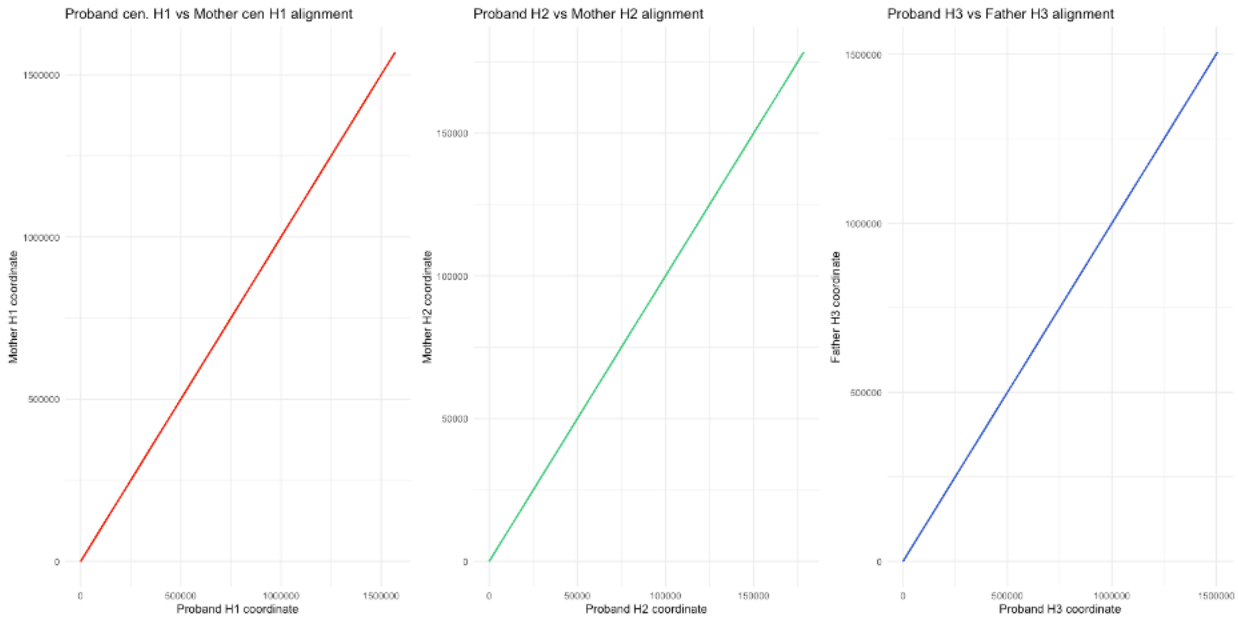




Supplementary Figure 3. CENP-A and CENP-C detection and quantification in proband AG16777, mother AG16778, and father AG16782. IF-FISH on metaphase chromosome spreads from all members of the AG167 family, with fluorescent probes targeting the chr13/21 α-satellite array (red) and CENP-A/CENP-C (green). All chr21 are highlighted in the boxes. A weaker α-satellite array probe signal on one of the chr21 centromeres of proband and mother is due to the smaller centromeric α-satellite HOR array present in each of these individuals. The violin plots represent the CENP-A and CENP-C signal ratios for each member of the family. Each data point represents a metaphase from which the values of CENP-A/CENP-C protein signals were quantified. Each signal value was divided by the homolog lowest value to obtain the ratio and self-comparisons were removed. Signal ratio of 1 indicates an identical value between homolog signals. T-test p-values are indicated at the top for each comparison.



Supplementary Figure 4. Repeat structure, NucFreq plot validation, methylation profile of chromosome 4, 12, and 15 centromeres in proband AG16777.



Supplementary Figure 5. Dotplots comparing AG167 proband and parental centromeric sequences haplotypes. The plots represent the alignments of proband and parental centromeric α-satellite array sequences.

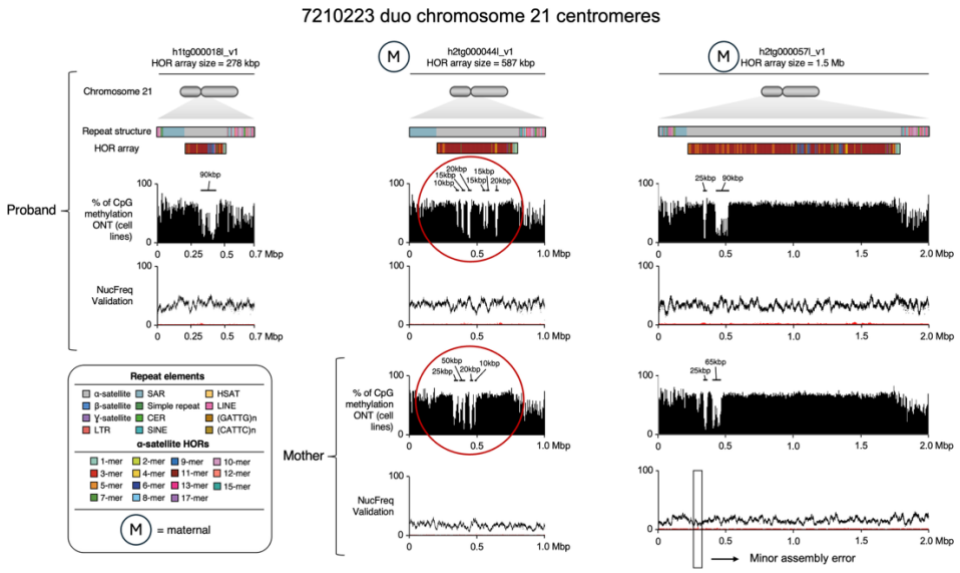
Query: Proband AluY insertion sequence + 10bp padding sequence on both sides
Subject: Father AluY insertion sequence + 10bp padding sequence on both sides

Score	Expect	Identities	Gaps	Strand
616 bits(333)	0.0	336/337(99%)	1/337(0%)	Plus/Plus

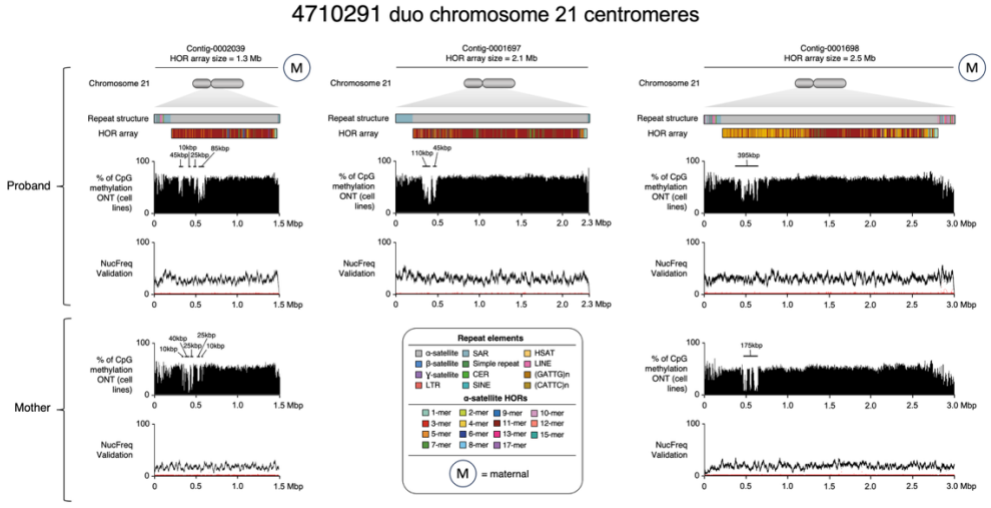
```
Query  1    AGTAGTTTCtttttttttttttttttttttttttttGAGACGGAGTCTCGCTCTGTC  60
            ||||||||| ||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    AGTAGTTTC-TTTTTTTTTTTTTTTTTTTTTTTTTTTGAGACGGAGTCTCGCTCTGTC  59

Query  61   GCCCAGGTCGGACTGCGGACTGCAGTGGCGCAATCTCGGCTCACTGCAAGCTCCGCTTCC  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  60   GCCCAGGTCGGACTGCGGACTGCAGTGGCGCAATCTCGGCTCACTGCAAGCTCCGCTTCC  119

Query  121  CGGGTTCACGCCATTCTCCTGCCTCAGCCTCCCGAGTAGCTGGGACTACAGGCGCCCGCC  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  120  CGGGTTCACGCCATTCTCCTGCCTCAGCCTCCCGAGTAGCTGGGACTACAGGCGCCCGCC  179

Query  181  ACCGCGCCCGGCTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTCACCTTGTTAGCCAG  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  180  ACCGCGCCCGGCTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTCACCTTGTTAGCCAG  239

Query  241  GATGGTCTCGATCTCCTGACCTCATGATCCACCCGCCTCGGCCTCCCAAAGTGCTGGGAT  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  240  GATGGTCTCGATCTCCTGACCTCATGATCCACCCGCCTCGGCCTCCCAAAGTGCTGGGAT  299

Query  301  TACAGGCGTGAGCCACCGCGCCCGGCCACCAAGTAGT  337
            |||||||||||||||||||||||||||||||||||||
Sbjct  300  TACAGGCGTGAGCCACCGCGCCCGGCCACCAAGTAGT  336
```

Supplementary Figure 6. BLAST-based pairwise alignment of the AluY sequences inserted in AG167 proband H3 and father H3 centromere sequences.

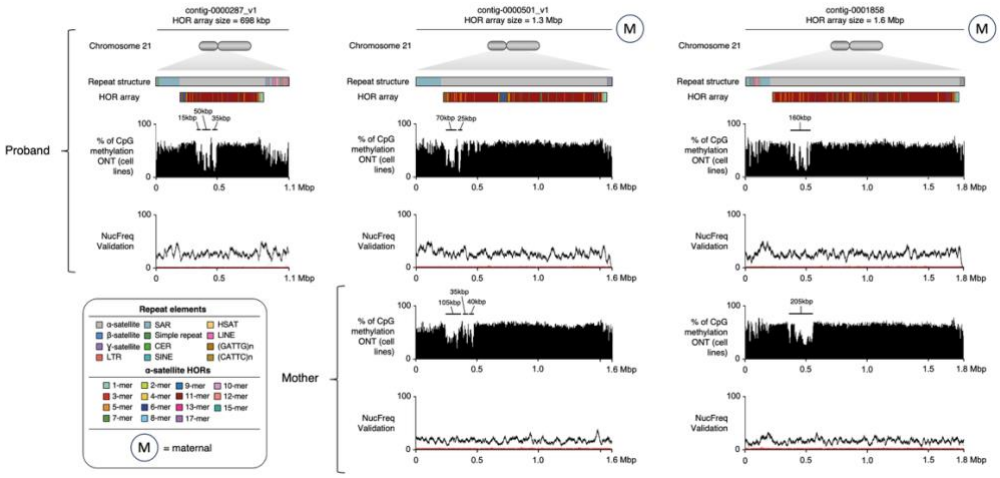
840

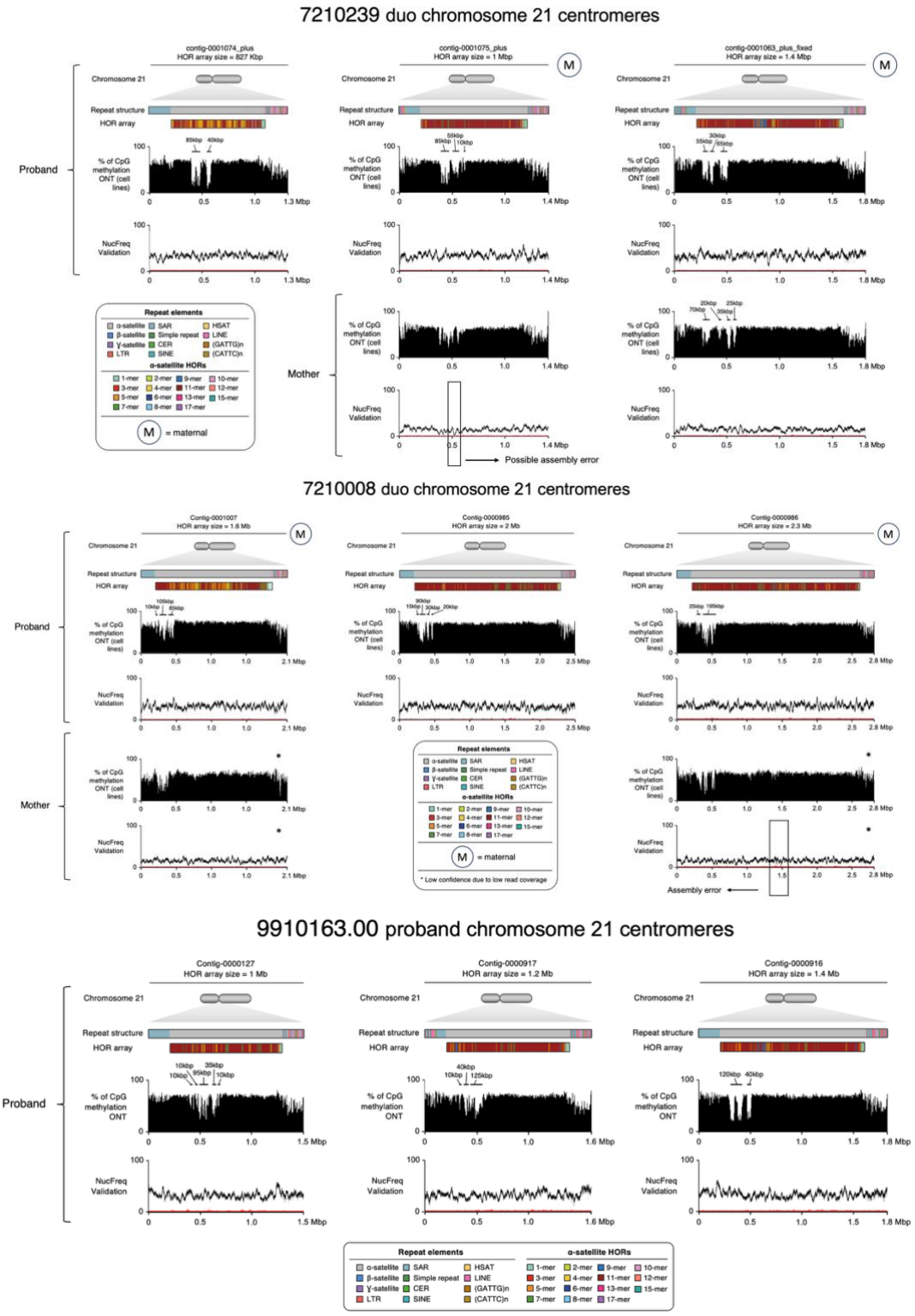

7210223 duo chromosome 21 centromeres
h1tg000016l_v1
HOR array size = 278 kbp
h2tg000044l_v1
HOR array size = 587 kbp
h2tg000057l_v1
HOR array size = 1.5 Mb
Chromosome 21
Repeat structure
HOR array
Proband
Mother
% of CpG methylation ONT (cell lines)
NucFreq Validation
Repeat elements
α-satellite
SAR
HSAT
β-satellite
Simple repeat
LINE
γ-satellite
CER
(GATTG)n
LTR
SINE
(CATTC)n
α-satellite HORs
1-mer
2-mer
9-mer
10-mer
3-mer
4-mer
11-mer
12-mer
5-mer
6-mer
13-mer
15-mer
7-mer
8-mer
17-mer
M = maternal
Minor assembly error

841


4710291 duo chromosome 21 centromeres
Contig-0002039
HOR array size = 1.3 Mb
Contig-0001697
HOR array size = 2.1 Mb
Contig-0001698
HOR array size = 2.5 Mb
Chromosome 21
Repeat structure
HOR array
Proband
Mother
% of CpG methylation ONT (cell lines)
NucFreq Validation
Repeat elements
α-satellite
SAR
HSAT
β-satellite
Simple repeat
LINE
γ-satellite
CER
(GATTG)n
LTR
SINE
(CATTC)n
α-satellite HORs
1-mer
2-mer
9-mer
10-mer
3-mer
4-mer
11-mer
12-mer
5-mer
6-mer
13-mer
15-mer
7-mer
8-mer
17-mer
M = maternal

842

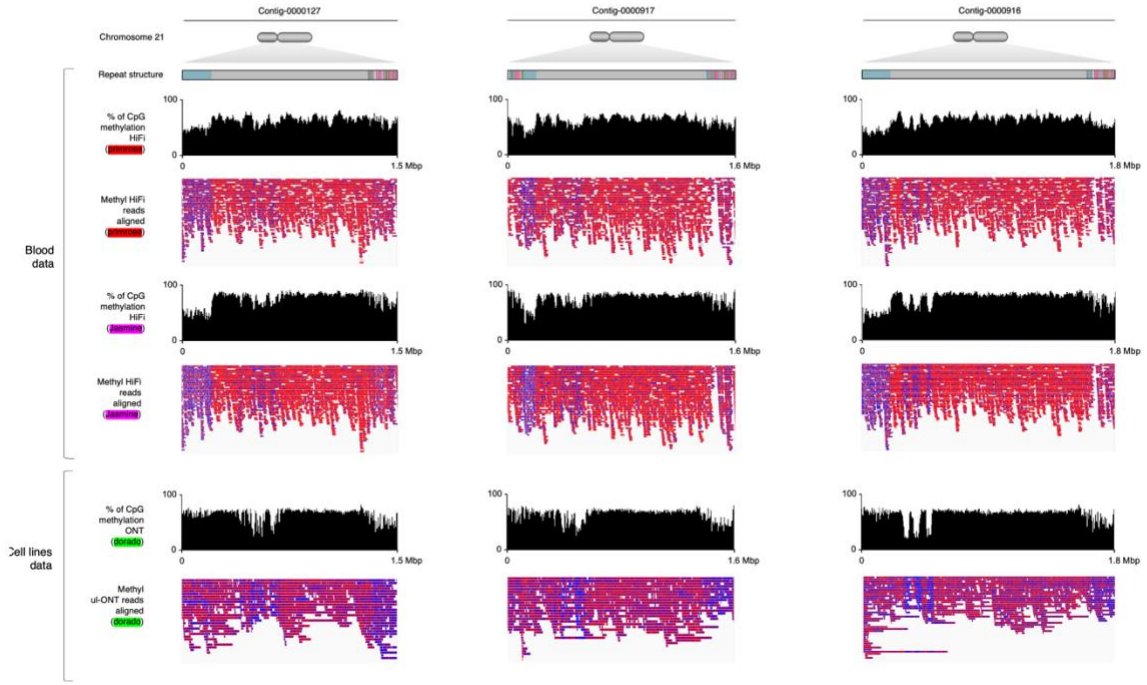

9915166 duo chromosome 21 centromeres
contig-0000287_v1
HOR array size = 698 kbp
contig-0000501_v1
HOR array size = 1.3 Mbp
contig-0001858
HOR array size = 1.6 Mbp
Chromosome 21
Repeat structure
HOR array
Proband
Mother
% of CpG methylation ONT (cell lines)
NucFreq Validation
Repeat elements
α-satellite
SAR
HSAT
β-satellite
Simple repeat
LINE
γ-satellite
CER
(GATTG)n
LTR
SINE
(CATTC)n
α-satellite HORs
1-mer
2-mer
9-mer
10-mer
3-mer
4-mer
11-mer
12-mer
5-mer
6-mer
13-mer
15-mer
7-mer
8-mer
17-mer
M = maternal



Supplementary Figure 7. Centromeres of individuals with T21 from families that do not exhibit extreme maternal homolog size asymmetry. Repeat structures, HORs and methylation profiles are shown. 7210223 duo is the only one showing transgenerational methylation changes.

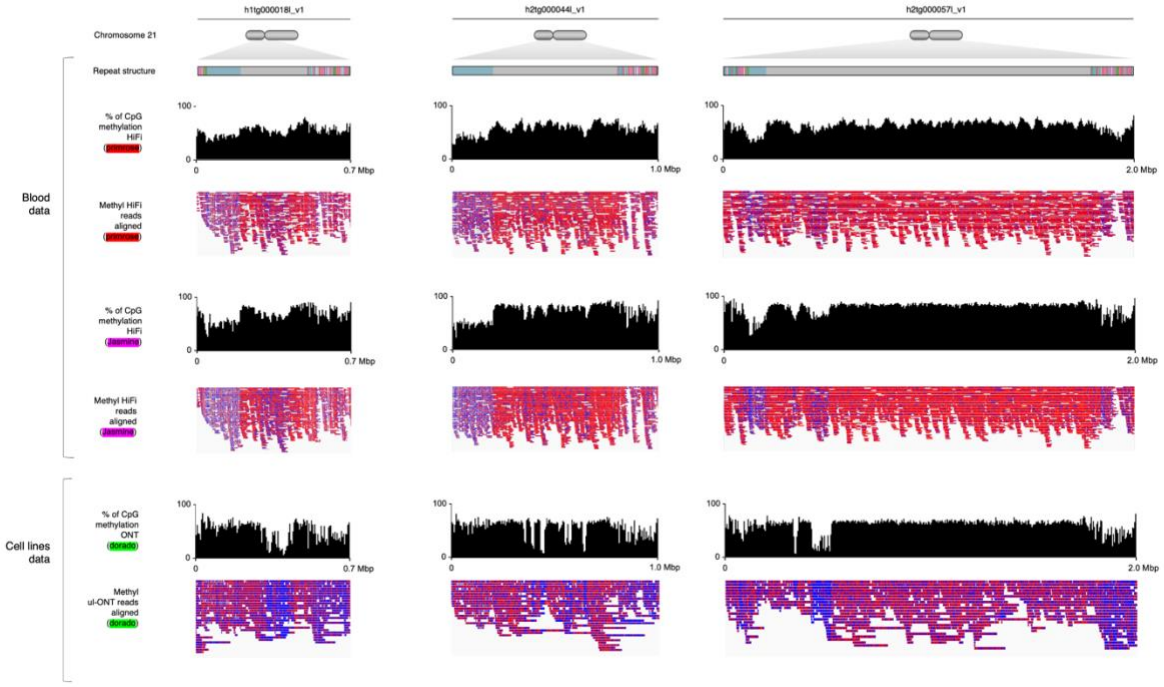
a.

Proband 9910163
chromosome 21 comparison between blood and cell lines DNA data



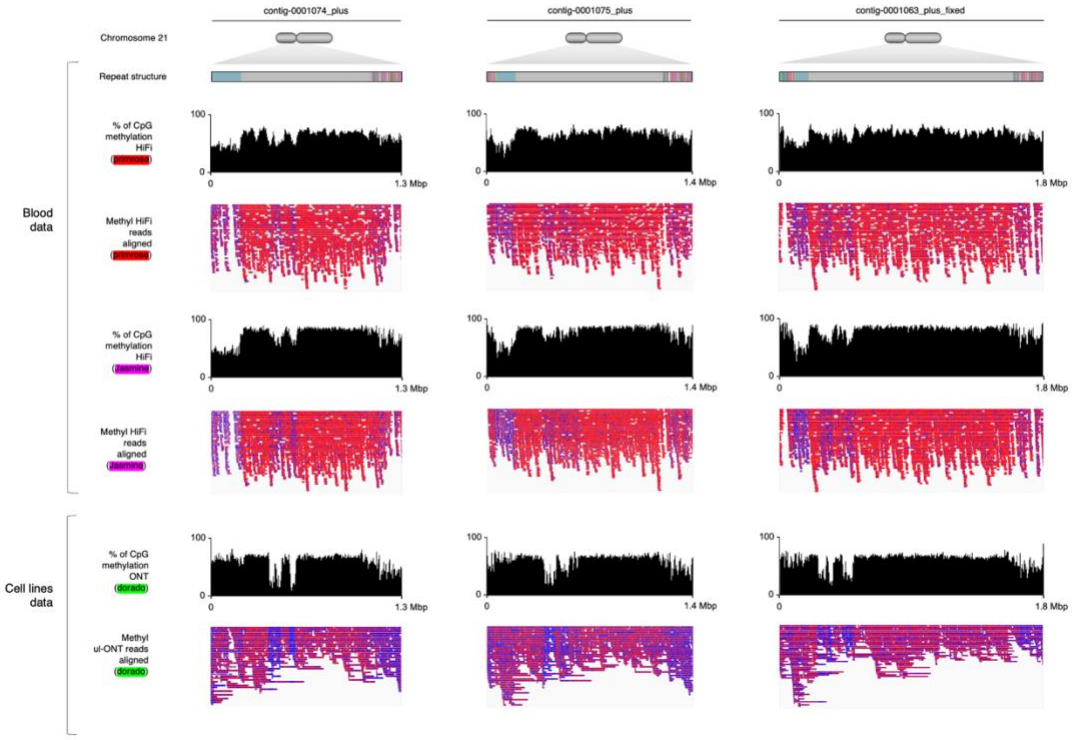
849

Proband 7210223
chromosome 21 comparison between blood and cell lines DNA data



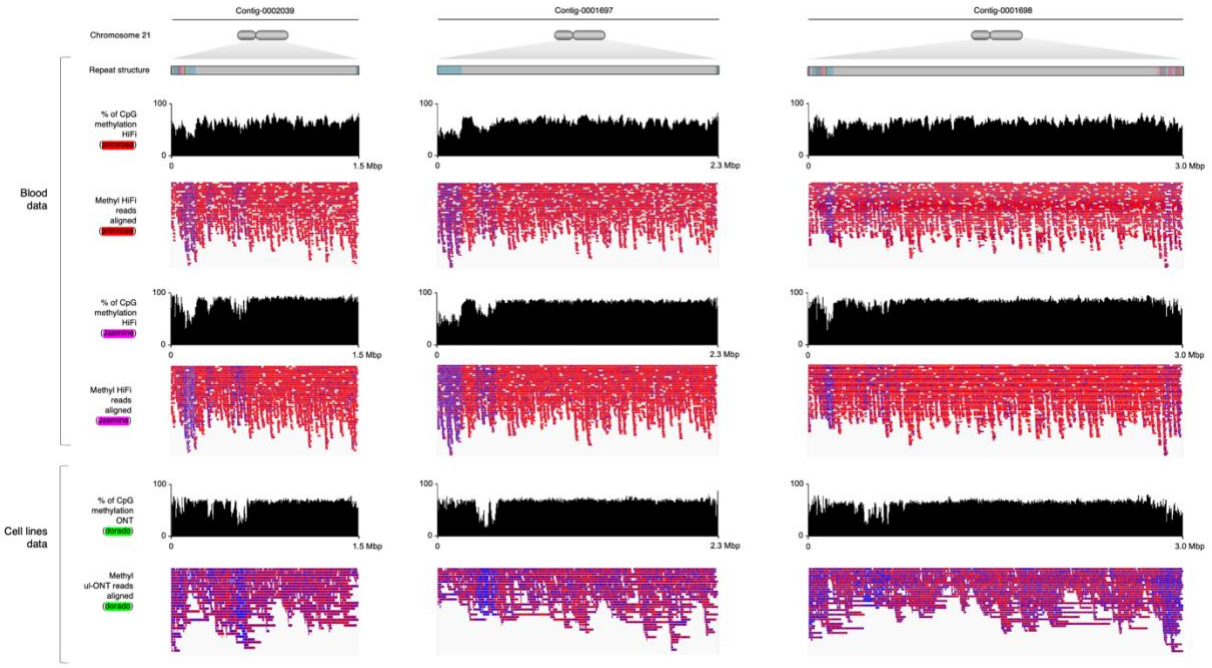
850

Proband 7210239
chromosome 21 comparison between blood and cell lines DNA data

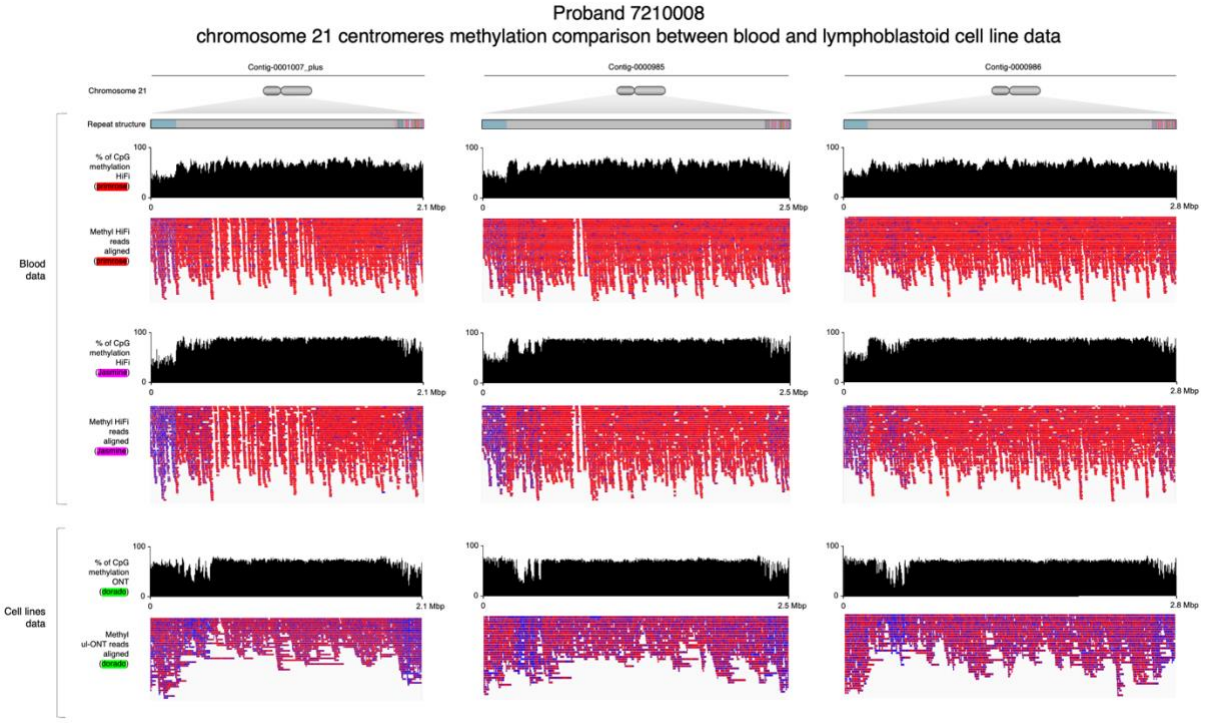


851

Proband 4710291
chromosome 21 comparison between blood and cell lines DNA data

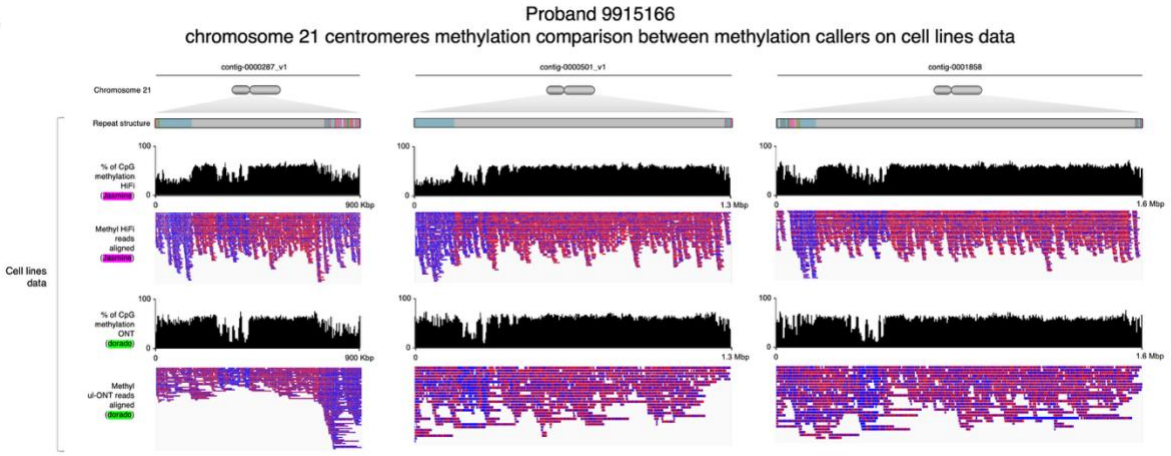


852

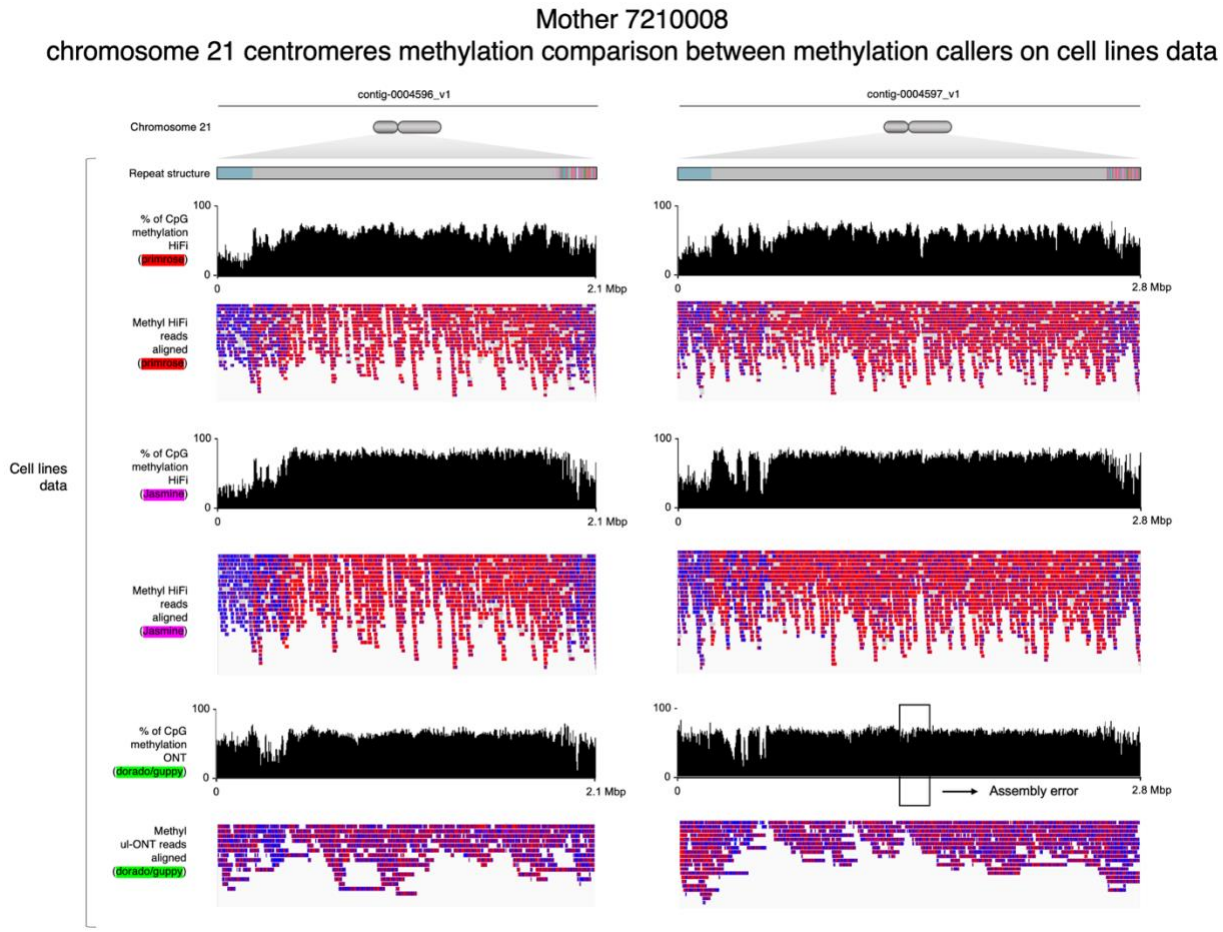

Proband 7210008
chromosome 21 centromeres methylation comparison between blood and lymphoblastoid cell line data
Contig-0001007_plus
Contig-0000985
Contig-0000986
Chromosome 21
Repeat structure
% of CpG methylation HiFi
Methyl HiFi reads aligned
Blood data
% of CpG methylation ONT
Methyl ul-ONT reads aligned
Cell lines data
2.1 Mbp
2.5 Mbp
2.8 Mbp

853

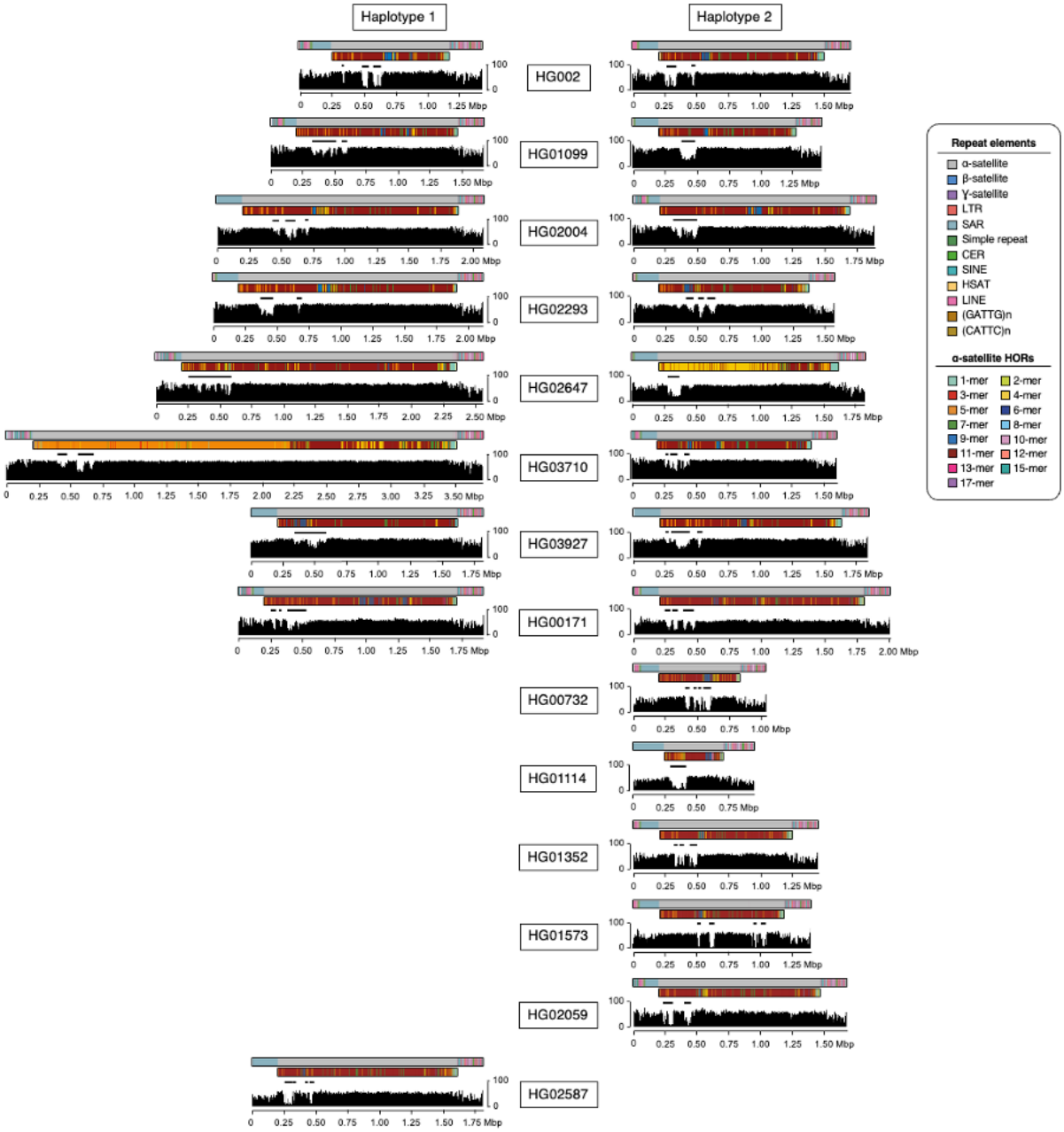
b.


Proband 9915166
chromosome 21 centromeres methylation comparison between methylation callers on cell lines data
contig-0000287_v1
contig-0000501_v1
contig-0001858
Chromosome 21
Repeat structure
% of CpG methylation HiFi
Methyl HiFi reads aligned
Cell lines data
% of CpG methylation ONT
Methyl ul-ONT reads aligned
900 Kbp
1.3 Mbp
1.6 Mbp

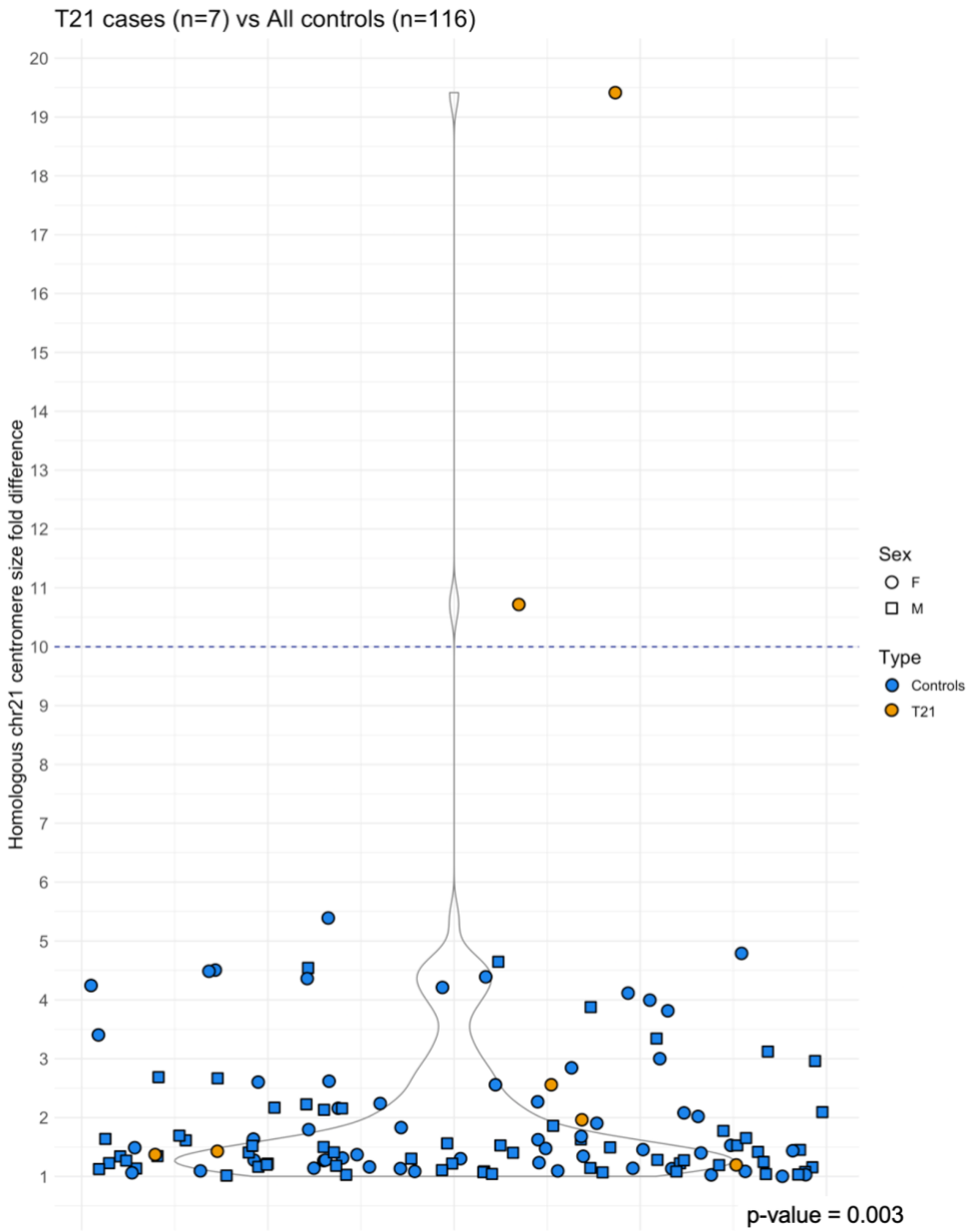
854



Supplementary Figure 8. Comparison of methylation profiles of chr21 centromeres in blood and cell line data. a) Chr21 centromere methylation profiles detected from HiFi (blood DNA) and UL-ONT (lymphoblastoid cell line DNA) in probands 9910163.00, 7210223.00, 7210239.00, 4710291.00, and 7210008.00. **b)** Performance comparison of methylation callers on cell line data in proband 9915166.00 and mother 7210008.20.



Supplementary Figure 9. Extended version of Figure 3c representing the methylation profiles and CDR positions in 22 population chr21 centromere haplotypes.



Supplementary Figure 10. Homologous chr21 centromere size fold differences in T21 centromeres (only maternally inherited) and population samples. T21 (orange) and population (blue) samples are represented with different colors; sex is represented with different shapes. P-value indicates a significant difference in asymmetry cases between the two groups.

870 **Supplementary Table 1.** All HOR monomers identified in 35 highly curated population chr21
871 centromeres.
872
873 **Supplementary Table 2.** Information on families with T21.
874
875 **Supplementary Table 3.** All chr21 centromeres described in the manuscript.
876

References

1. Sperling, K., Scherb, H. & Neitzel, H. Population monitoring of trisomy 21: problems and approaches. *Molecular Cytogenetics* **16**, 6 (2023).
2. Antonarakis, S. E. *et al.* Down syndrome. *Nat Rev Dis Primers* **6**, 1–20 (2020).
3. Logsdon, G. A. & Eichler, E. E. The Dynamic Structure and Rapid Evolution of Human Centromeric Satellite DNA. *Genes* **14**, 92 (2023).
4. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
5. Logsdon, G. A. *et al.* The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).
6. Altemose, N. *et al.* Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).
7. Logsdon, G. A. *et al.* The variation and evolution of complete human centromeres. *Nature* **629**, 136–145 (2024).
8. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
9. Guarracino, A. *et al.* Recombination between heterologous human acrocentric chromosomes. *Nature* **617**, 335–343 (2023).
10. Vollger, M. R. *et al.* Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).
11. Jeong, H. *et al.* Structural polymorphism and diversity of human segmental duplications. *Nat Genet* **57**, 390–401 (2025).
12. Rhie, A. *et al.* The complete sequence of a human Y chromosome. *Nature* **621**, 344–354 (2023).
13. Hallast, P. *et al.* Assembly of 43 human Y chromosomes reveals extensive complexity and variation. *Nature* **621**, 355–364 (2023).
14. Vafa, O. & Sullivan, K. F. Chromatin containing CENP-A and α-satellite DNA is a major component of the inner kinetochore plate. *Current Biology* **7**, 897–900 (1997).
15. Régnier, V. *et al.* CENP-A Is Required for Accurate Chromosome Segregation and Sustained Kinetochore Association of BubR1. *Mol Cell Biol* **25**, 3967–3981 (2005).
16. Gershman, A. *et al.* Epigenetic patterns in a complete human genome. *Science* **376**, eabj5089 (2022).
17. Mastrorosa, F. K. *et al.* Identification and annotation of centromeric hypomethylated regions with CDR-Finder. *Bioinformatics* **40**, btae733 (2024).
18. Hassold, T. & Hunt, P. To err (meiotically) is human: the genesis of human aneuploidy. *Nat Rev Genet* **2**, 280–291 (2001).
19. Nagaoka, S. I., Hassold, T. J. & Hunt, P. A. Human aneuploidy: mechanisms and new insights into an age-old problem. *Nat Rev Genet* **13**, 493–504 (2012).
20. Iwata-Otsubo, A. *et al.* Expanded Satellite Repeats Amplify a Discrete CENP-A Nucleosome Assembly Site on Chromosomes that Drive in Female Meiosis. *Current Biology* **27**, 2365-2373.e8 (2017).
21. Chmátal, L., Schultz, R. M., Black, B. E. & Lampson, M. A. Cell Biology of Cheating-Transmission of Centromeres and Other Selfish Elements Through Asymmetric Meiosis. *Prog Mol Subcell Biol* **56**, 377–396 (2017).
22. Lampson, M. A. & Black, B. E. Cellular and Molecular Mechanisms of Centromere Drive. *Cold Spring Harb Symp Quant Biol* **82**, 249–257 (2017).
23. Dumont, M. *et al.* Human chromosome-specific aneuploidy is influenced by DNA-dependent centromeric features. *The EMBO Journal* **39**, e102924 (2020).
24. Irvine, D. V. *et al.* Chromosome size and origin as determinants of the level of CENP-A incorporation into human centromeres. *Chromosome Res* **12**, 805–815 (2004).
25. Rudd, M. K. & Willard, H. F. Analysis of the centromeric regions of the human genome assembly. *Trends in Genetics* **20**, 529–533 (2004).
26. Dumont, M. & Fachinetti, D. DNA Sequences in Centromere Formation and Function. in

Centromeres and Kinetochores: Discovering the Molecular Mechanisms Underlying Chromosome Inheritance (ed. Black, B. E.) 305–336 (Springer International Publishing, Cham, 2017). doi:10.1007/978-3-319-58592-5_13.

27. Contreras-Galindo, R. *et al.* Rapid molecular assays to study human centromere genomics. *Genome Res* **27**, 2040–2049 (2017).

28. Angell, R. R. Predivision in human oocytes at meiosis I: a mechanism for trisomy formation in man. *Hum Genet* **86**, 383–387 (1991).

29. Smith, K. N. & Nicolas, A. Recombination at work for meiosis. *Current Opinion in Genetics & Development* **8**, 200–211 (1998).

30. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).

31. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).

32. Rautiainen, M. *et al.* Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol* **41**, 1474–1482 (2023).

33. Vollger, M. R., Kerpedjiev, P., Phillippy, A. M. & Eichler, E. E. StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps. *Bioinformatics* **38**, 2049–2051 (2022).

34. Smith, G. P. Evolution of Repeated DNA Sequences by Unequal Crossover. *Science* **191**, 528–535 (1976).

35. Logsdon, G. A. *et al.* Complex genetic variation in nearly complete human genomes. 2024.09.24.614721 Preprint at https://doi.org/10.1101/2024.09.24.614721 (2024).

36. Allen, E. G. *et al.* Maternal age and risk for trisomy 21 assessed by the origin of chromosome nondisjunction: a report from the Atlanta and National Down Syndrome Projects. *Hum Genet* **125**, 41–52 (2009).

37. Locke, A. E. *et al.* Variation in Folate Pathway Genes Contributes to Risk of Congenital Heart Defects Among Individuals With Down Syndrome. *Genet Epidemiol* **34**, 613–623 (2010).

38. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).

39. Down, J. L. Observations on an ethnic classification of idiots. 1866. *Ment Retard* **33**, 54–56 (1995).

40. Lejeune, J., Gautier, M. & Turpin, R. [Study of somatic chromosomes from 9 mongoloid children]. *C R Hebd Seances Acad Sci* **248**, 1721–1722 (1959).

41. Dondorp, W. *et al.* Non-invasive prenatal testing for aneuploidy and beyond: challenges of responsible innovation in prenatal screening. *Eur J Hum Genet* **23**, 1438–1450 (2015).

42. Maratou, K., Siddique, Y., Kessling, A. M. & Davies, G. E. Variation in alphoid DNA size and trisomy 21: a possible cause of nondisjunction. *Hum Genet* **106**, 525–530 (2000).

43. Laurent, A.-M., Li, M., Sherman, S., Roizès, G. & Buard, J. Recombination across the centromere of disjoined and non-disjoined chromosome 21. *Human Molecular Genetics* **12**, 2229–2239 (2003).

44. Lo, A. W., Liao, G. C., Rocchi, M. & Choo, K. H. Extreme reduction of chromosome-specific alpha-satellite array is unusually common in human chromosome 21. *Genome Res* **9**, 895–908 (1999).

45. Craymer, L. Techniques for manipulating chromosomal rearrangements and their application to Drosophila melanogaster. I. Pericentric inversions. *Genetics* **99**, 75–97 (1981).

46. Craymer, L. Techniques for manipulating chromosomal rearrangements and their application to Drosophila melanogaster. II. Translocations. *Genetics* **108**, 573–587 (1984).

47. McKim, K. S., Howell, A. M. & Rose, A. M. The effects of translocations on recombination frequency in Caenorhabditis elegans. *Genetics* **120**, 987–1001 (1988).

48. Moens, P. B., Heddle, J. A., Spyropoulos, B. & Heng, H. H. Identical megabase transgenes on mouse chromosomes 3 and 4 do not promote ectopic pairing or synapsis at meiosis. *Genome* **40**, 770–773 (1997).

49. Walker, M. Y. & Hawley, R. S. Hanging on to your homolog: the roles of pairing, synapsis and

recombination in the maintenance of homolog adhesion. *Chromosoma* **109**, 3–9 (2000).
50. Tak, T., Tesselaar, K., Pillay, J., Borghans, J. A. M. & Koenderman, L. What's your age again? Determination of human neutrophil half-lives revisited. *Journal of Leukocyte Biology* **94**, 595–601 (2013).
51. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021).
52. Earnshaw, W. C. & Tomkiel, J. E. Centromere and kinetochore structure. *Curr Opin Cell Biol* **4**, 86–93 (1992).
53. Lichter, P. *et al.* High-Resolution Mapping of Human Chromosome 11 by in Situ Hybridization with Cosmid Clones. *Science* **247**, 64–69 (1990).
54. Bodor, D. L., Rodríguez, M. G., Moreno, N. & Jansen, L. E. T. Analysis of protein turnover by quantitative SNAP-based pulse-chase imaging. *Curr Protoc Cell Biol* **Chapter 8**, Unit8.8 (2012).
55. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer, New York, NY, 2009). doi:10.1007/978-0-387-98141-3.
56. R Development Core Team, R. *R: A Language and Environment for Statistical Computing*. *R Foundation for Statistical Computing* vol. 1 409 (2011).
57. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
58. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015.
59. Vollger, M. R. *et al.* Long-read sequence and assembly of segmental duplications. *Nat Methods* **16**, 88–94 (2019).
60. Dvorkina, T., Bzikadze, A. V. & Pevzner, P. A. The string decomposition problem and its applications to centromere analysis and assembly. *Bioinformatics* **36**, i93–i101 (2020).
61. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).
62. Jain, C., Rhie, A., Hansen, N. F., Koren, S. & Phillippy, A. M. Long-read mapping to repetitive reference sequences using Winnowmap2. *Nat Methods* **19**, 705–710 (2022).
63. Nurk, S. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* gr.263566.120 (2020) doi:10.1101/gr.263566.120.
64. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**, 772–780 (2013).
65. Nakamura, T., Yamada, K. D., Tomii, K. & Katoh, K. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* **34**, 2490–2492 (2018).
66. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32**, 268–274 (2015).
67. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128 (2007).