Complete chromosome 21 centromere sequences from a Down syndrome family reveal size asymmetry and differences in kinetochore attachment

F. Kumara Mastrorosa[1], Allison N. Rozanski[1], William T. Harvey[1], Jordan Knuth[1], Gage Garcia[1], Katherine M. Munson[1], Kendra Hoekzema[1], 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] Current address: Department of Genetics, Epigenetics Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

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

* Correspondence: glogsdon@pennmedicine.upenn.edu, eee@gs.washington.edu

ABSTRACT

Down syndrome is the most common form of human intellectual disability caused by precocious segregation and nondisjunction of chromosome 21. Differences in centromere structure have been hypothesized to play a potential role in this process in addition to the well-established risk of advancing maternal age. Using long-read sequencing, we completely sequenced and assembled the centromeres from a parent-child trio where Trisomy 21 arose in the child as a result of a meiosis I error. The proband carries three distinct chromosome 21 centromere haplotypes that vary by 11-fold in length--both the largest (H1) and smallest (H2) originating from the mother. The longest H1 allele harbors a less clearly defined centromere dip region (CDR) as defined by CpG methylation and a significantly reduced signal by CENP-A chromatin immunoprecipitation sequencing when compared to H2 or paternal H3 centromeres. These epigenetic signatures suggest less competent kinetochore attachment for the maternally transmitted H1. Analysis of H1 in the mother indicates that the reduced CENP-A ChIP-seq signal, but not the CDR profile, pre-existed the meiotic nondisjunction event. A comparison of the three proband centromeres to a population sampling of 35 completely sequenced chromosome 21 centromeres shows that H2 is the smallest centromere sequenced to date and all three haplotypes (H1-H3) share a common origin of ~15 thousand years ago. These results suggest that recent asymmetry in size and epigenetic differences of chromosome 21 centromeres may contribute to nondisjunction risk.

KEYWORDS

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

LIST OF ABBREVIATIONS

CDR = centromere dip region

ChIP-seq = chromatin immunoprecipitation followed by sequencing

chr21 = chromosome 21

Gbp = Gigabase pairs

kbp = kilobase pairs

kya = thousand years ago

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

HiFi = high fidelity

HGSVC = Human Genome Structural Variation Consortium

HPRC = Human Pangenome Reference Consortium

HOR = higher-order repeat

MI/II = meiosis I/II

Mbp = Megabase pairs

PacBio = Pacific Biosciences

T21 = Trisomy 21

T2T = telomere-to-telomere

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 affected individual[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 (MI) errors (~66%), followed by maternal meiosis II (MII) errors (~21%), paternal MI (~3%) and MII errors (~5%), or mitotic events after the zygote formation (~5%)[2].

Centromeres play an essential role for accurate chromosome segregation during meiosis by promoting the 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: ~171-base pair (bp) nearly-identical sequences 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 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 four 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], and other complex loci[8,11,12] previously regarded as inaccessible.

Functionally, the kinetochore is marked by the presence of nucleosomes containing the histone variant H3 CENP-A, which provides the foundation for the assembly of inner- and outer-kinetochore proteins and plays a crucial role in cell division and chromosome segregation[13,14]. The analysis of the complete sequence of human centromeres[5–7] has made it possible to define the likely site of kinetochore attachment: a pocket of hypomethylated CpG DNA called the centromere dip region (CDR)[15], where chromatin immunoprecipitation followed by sequencing (ChIP-seq) shows an enrichment of CENP-A. All completely sequenced human centromeres to date derive from primary tissue (blood) or cell lines and possess CDRs up to tens of kilobase pairs (kbp) in length where CENP-A accumulates, implicating this as a general epigenetic signature of the kinetochore[7].

Given their role in cell division, centromeres have been postulated to play an important role in chromosome nondisjunction, including Down syndrome[16]. For instance, a recent report suggested that T21 individuals might have shorter chr21 centromeres (possibly as a result of reduced size of the centromeric α-satellite HOR array)[17]. Centromere size was estimated, however, based on quantitative, real-time PCR[17], which can be challenging given the high degree of sequence identity between chromosome 13 and 21 α-satellite DNA and a limited understanding of the sequence diversity within the α-satellite HOR arrays[17]. Other studies showed that cohesion loss between sister centromeres[18] and reduced recombination[19] increase the risk of nondisjunction. Nevertheless, the most well-established non-genetic risk factor is increased maternal age at conception[16].

It is currently unknown whether centromere variation amongst humans increase the likelihood of nondisjunction. For instance, centromere size, sequence, allelic differences, and epigenetic features might influence nondisjunction risk and indicate predisposition to such events.

Recently, combining both Pacific Biosciences (PacBio) high-fidelity (HiFi) long-read sequencing and ultra-long Oxford Nanopore Technologies (UL-ONT) sequencing, we demonstrated that it is possible to completely sequence and distinguish human autosomal alleles for chr21[7]. We applied this approach to a parent–child trio, where the proband has T21. Here, we present the first complete linear sequence and epigenetic profile of chr21 centromeres in an individual with maternal MI error T21 and their parents. We also compare patterns of genetic and epigenetic variation of the chr21 centromeres to a control group of 35 completely sequenced chr21 centromeres from diploid samples revealing unique features of T21 centromeres.

RESULTS

Proband chromosome 21 centromeres. To study the chr21 centromeres of a T21 proband, we selected a family where the maternal age at the time of conception was 29 years and for which cell lines were available (**Methods**). We generated both PacBio HiFi sequence data (59×) and UL-ONT sequence data (23×) using DNA extracted from the T21 lymphoblastoid cell lines. We used Verkko hybrid genome assembly workflow[20], which combines PacBio HiFi and UL-ONT sequence data to assemble all T21 proband haplotypes (total genome assembly size = 6.118 Gbp, contigs = 3,466, contig N50 = 7.72 Mbp). We aligned the proband assembly against the complete telomere-to-telomere (T2T) CHM13 (T2T-CHM13.v2) genome assembly and extracted the contigs mapping to the chr21 centromere and flanking regions. We assessed the contiguity of the chr21 centromere assemblies and performed quality control as described previously[5,7] to ensure the chr21 centromere assemblies were complete—without misassemblies or gaps (**Supplementary Figure 1**). Sequence assembly of phased chromosome 21 centromeres haplotypes confirmed a maternal MI error where haplotypes 1 (H1) and 2 (H2) arose from maternal nondisjunction and H3 was inherited from the father (**Figure 1a**).

H1 and H3 centromeres are the most similar in size with an α-satellite HOR array of ~1.53 Mbp and ~1.47 Mbp, respectively. By contrast, maternal H2 is ~11-fold smaller with a HOR array of only 142 kbp (**Figure 1a**). The α-satellite HOR array sequences from all three haplotypes are predominantly composed of 11-monomer α-satellite HORs (~87-93%) with a smattering of rarer α-satellite HORs 2- to 10-, 13-, 17-, and 18-monomers in length (each constituting <3% of the sequence) (**Figure 1a and Supplementary Table 1**). Other than their size, the three chr21 centromeres share a remarkably similar organization, with the α-satellite HOR array flanked by human satellite 1a (HSat1a; annotated as "SAR'' by RepeatMasker) on the p-arm and various classes of mobile elements (SINEs and LINEs) on the q-arm. Also, H3 is distinguished by a 317 bp AluY retrotransposon insertion in its α-satellite HOR array (**Figure 1a**).

Given the use of ONT technology, which sequences native DNA molecules and allows for CpG methylation to be detected, we compared the CpG methylation profile of each of the three chr21 centromere haplotypes based on the modification of cytosine residues. Consistent with its heterochromatic nature, the majority of CpG dinucleotides in a centromere were methylated with a small pocked of hypomethylated DNA, termed the "centromere dip region (CDR)", which is thought to define the site of kinetochore attachment[5,6]. Both H2 and H3 match this description with a CDR of 120 kbp and 135 kbp, respectively (**Methods**). By contrast, the maternally inherited H1 does not have a clearly defined CDR. The CDR is fragmented into five smaller pockets ranging from 10 to 50 kbp and spanning across 280 kbp of centromeric α-satellite HORs. For each CDR, we calculated the percentage of methylated CpGs (defined as having ≥25% methylated CpGs at the same position; **Methods**) across the region comprising the CDR pockets (from the start of the first CDR window to the end of the last one). We found that the percentage of methylated CpGs within H1's CDR was much higher than those of H2 and H3's CDRs (~71% vs. ~39% and ~38%, respectively) (**Figure 1b**). This more diffuse pattern of CDRs is associated with the largest HOR array within this family. We orthogonally validated the

centromere methylation profiles using PacBio HiFi sequencing data (**Supplementary Figure 3** and **Methods**).

Because of the central role of CENP-A in defining the site of kinetochore assembly, we assessed the centromeric CENP-A chromatin pattern on each chr21 centromeric haplotype by performing native CENP-A ChIP-seq and mapping the data back to the sequence-resolved T21 centromeres. We calculated the fold enrichment of CENP-A ChIP reads over bulk nucleosomal reads within the region that encompasses both parent-child CDRs (**Methods** and **Supplementary Table 2**). Consistent with the CDR pattern, we observed a reduction of CENP-A enrichment for H1 (~139-fold over a 280 kbp window), compared to the H2 (~351-fold over a ~136 kbp window) and H3 (~196-fold over 170 kbp window) (**Figure 1b**).

Finally, we placed the epigenetic profile for each centromere in the context of the α-satellite HOR array organization by generating sequence identity heatmaps, based on pairwise sequence alignment of 5 kbp segments using StainedGlass[21] (**Figure 1c**). All three haplotypes show similar compositional features, with the most identical segments mapping to the central section of the α-satellite HOR array but not corresponding to the CDR or site of CENP-A signal, with the exception of H2 due to its smaller size. This structure is also similar to the complete chr21 α-satellite HOR arrays described for two complete hydatidiform moles (CHM1 and CHM13)[5,6], as well as population samples[7], and follows the Smith model described in 1976[22]. Interestingly, for both H1 and H3, the CDR and the predicted site of kinetochore attachment occur peripherally in a region composed of more heterogeneous α-satellite HOR cassettes (with an admixture of 5- and 11-monomer HORs).

Parental chromosome 21 centromeres. To understand the inheritance of the chr21 centromeres, we sequenced and assembled the genomes of both parents using DNA extracted from lymphoblastoid cell lines and fully reconstructed each of the centromere haplotypes from the diploid parental cell lines. To do this, we generated PacBio HiFi sequencing data at 37× coverage for the mother and 40× for the father and complemented these data with UL-ONT data at 25× and 12× coverage, respectively, to generate *de novo* assemblies (Mother: total genome assembly size = 6.078 Gbp, contigs = 4,827, contig N50 = 5.41 Mbp; Father: total genome assembly size = 5.947 Gbp, contigs = 5,032, contig N50 = 5.01 Mbp). Centromeric contigs from chr21 were extracted, validated, and compared to the T21 child (**Supplementary Figure 1**; see **Methods** for assembly correction).

Sequencing confirmed that the largest and the smallest chr21 centromeres (H1 and H2, respectively) were maternally transmitted while H3 originated from the father (**Figure 2a-b**). Sequence identity between parent and proband haplotypes was >99.99% identical—confirming the accuracy of the assembly as well as the stability of the centromere intergenerationally and in cell culture (**Supplementary Figure 2**). No difference in the α-satellite repeat composition, organization, or array size were observed, with the exception of a 1 bp deletion in the poly(T) tract associated with an AluY insertion on the paternal haplotype (H3) when compared to the proband (**Supplementary Figure 3**). We note that the second non-transmitted paternal haplotype (Father H4) has an α-satellite HOR array size of 254 kbp (**Figure 2a**). This is 112 kbp larger than the small maternal haplotype (H2). Similar to the other haplotypes, the 11-monomer α-satellite HORs constitute the majority of its array sequence (~86%) with 2- to 5-, 7-, 9-, 10-, and 13-monomer HORs at lower frequency (each composing <6% of the array sequence).

Comparing the centromeric CpG methylation profiles between the parents and the proband, we find that the atypical methylation pattern observed in proband H1 (five hypomethylated pockets

10-50 kbp in size) appears to be specific to the proband (**Figure 2a-b**) (see **Supplementary Figure 5** for orthogonal validation). The mother shows a more typical CpG methylation profile, with two clear CDRs of 30 and 105 kbp in length, and ~46% methylated CpGs across the region comprising the CDR pockets. This is 25 percentage points less compared to the proband (~71%) and it is in contrast to the transmitted maternal H2 (42% vs. 39% in proband) and paternal H3 (48% vs. 38% in proband), which show more similar CDR patterns. The CDRs of chr21 H2 and H3 centromeres differ by only an estimated 5-10 kbp in size between the generations (**Figure 2c**). Interestingly, we observe a shift of both proband H2 and H3 CDRs towards the p-arm relative to the parents despite nearly identical sequences. The maternal H2 CDR, for example, starts 37 kbp from the beginning of the α-satellite array, while in the proband the hypomethylated pocket starts 11 kbp upstream. The paternal H3 has a CDR initiating 131 kbp from the beginning of the α-satellite HOR array, which is shifted 25 kbp towards the p-arm in the proband (**Figure 2c**).

We also compared the fold enrichment of CENP-A ChIP-seq reads over bulk nucleosomal reads within equivalent windows used for the proband haplotypes. In contrast to the CDR result, our analysis reveals that the maternal H1, despite having a clearly defined CDR, has a ~3.7-fold lower CENP-A enrichment relative to the proband (~38-fold vs ~139-fold enrichment, respectively). In contrast, the maternal H2, paternal H3, and paternal H4 haplotypes all have much more robust CENP-A fold-enrichment (~568-fold, ~504-fold, and ~371-fold enrichment, respectively).

Chromosome 21 centromere population diversity. We compared the familial T21 centromeres to 35 completely assembled centromeres from 25 population samples with diverse ancestry. These centromeres were similarly sequenced and assembled with Verkko[20] using both PacBio HiFi and ONT long-read sequencing data generated as part of the Human

Pangenome Reference Consortium (HPRC)[23] or the Human Genome Structural Variation Consortium (HGSVC)[24]. The chr21 centromere α-satellite HOR array size ranges from 242 kbp to 3.27 Mbp in these genomes, representing a >13-fold difference in α-satellite HOR array length. Based on this, the maternally inherited H2 chr21 centromere remains the smallest α-satellite HOR array sequenced to date (142 kbp). It is predicted to be almost half the size of the next largest chr21 α-satellite HOR array observed in our survey of finished chr21 centromere assemblies (**Figure 3a-b**).

There is considerable diversity in α-satellite HOR composition amongst the 35 complete chr21 centromeres. While the most prevalent α-satellite HOR is 11 monomers in length and comprises ~83% to ~95% of the α-satellite HORs sequence in most of the complete chr21 centromeres, we also detect less abundant α-satellite HORs not observed within the T21 family chr21 centromeres that are 12, 14, 15, 16, and 20 α-satellite monomers in length (**Supplementary Table 1**). There are two notable haplotypes whose composition differs markedly from the predominant 11-monomer α-satellite HOR organization: an individual of Gambian (GWD) descent, HG02647, carries a 4-monomer α-satellite HOR composing ~15% of its H2 centromere sequence; and an individual of Punjabi (PJB) descent, HG03710, with 44% of its H1 centromere sequence composed of 5-monomer α-satellite HORs (**Figure 3b**). Sequence identity plots (**Figure 3d**) reveal that the 11-monomer-enriched α-satellite HOR region of HG03710 (H1) shows greater sequence identity when compared to the 5-monomer-enriched regions. In contrast, the 4-monomer α-satellite HOR of HG02647 (H2) shows greater sequence identity when compared to other flanking monomers (**Figure 3d**). If sequence divergence is a proxy for evolutionary age, our analyses may suggest that the 4-monomer α-satellite HOR has more recently evolved while the 5-monomer α-satellite HOR is more ancient. Notwithstanding, both of these rarer centromere haplotypes map their CDRs to the non-canonical 5-monomer and 4-monomer α-satellite HORs.

We similarly compared the diversity in length and location of the CDRs for the various chr21 centromere haplotypes. The number of CDRs typically range from 1 to 4 and span a total of 90 to 335 kbp in total length (**Figure 3c; Supplementary Figure 4**). None showed a pattern similar to the one observed in the proband (H1) with multiple small CDRs mapped in close proximity. We observe four chr21 centromere haplotypes that have two hypomethylated regions separated by more than 100 kbp: HG002 (H1 and H2); HG02293 (H1); HG01573 (H2). The dispersed CDRs tend to be smaller (10 to 25 kbp) and contained within a single pocket, except for HG01573 (H2), which has two. Irrespective of sequence composition, 91% (20/22) of the chr21 CDRs map to the p-arm of the centromere, suggesting a polarity or preference for CDR location (**Supplementary Figure 4**).

Finally, we constructed a chr21 centromere phylogenetic tree using 20 kbp of flanking pericentromeric sequence in the q-arm and chimpanzee as an outgroup (**Methods**). We have shown previously that these flanking sequences can serve as a reasonable proxy of evolutionary history due to the large block of linkage disequilibrium created by the suppressed recombination around the centromere[7]. We find that all T21 family haplotypes are part of a large monophyletic clade representing 35% (12/34) of all population chr21 centromeres. The clade harbors the largest (H1) and shortest (H2) T21 proband centromeres, which are predicted to have arisen from a common ancestor 15 thousand years ago (kya), despite the fact that these two haplotypes vary by more than 10-fold in their overall length. It is noteworthy that all short T21 centromeres (<500 kbp) belong to the same monophyletic clade, and the shortest T21 maternal H2 is predicted to be even more recently derived. As a comparison, the most divergent population centromeres harboring 5-monomer and 4-monomer α-satellite HOR cassettes are estimated to have diverged 338 kya and 356 kya, respectively.

DISCUSSION

In this study, we present for the first time the complete sequence and epigenetic profiles of all chr21 centromeres of a T21 family. We find that the T21 proband centromeres show an 11-fold differential in size of the two α-satellite HOR arrays transmitted from the mother, and one of them is the smallest identified to date. Small centromeric α-satellite HOR arrays have been postulated to be involved in the syndrome, and a recent study suggested that many T21 individuals harbor a small chr21 centromere[17]. This analysis, however, was based on quantitative, real-time PCR assays targeting the α-satellite HORs of chr21 to estimate the size and was incapable of determining the complete structure and architecture of the complete centromeres given its genetic diversity. Moreover, a previous report suggested that this feature could also be found in the general population[25]. It is unclear whether a small centromere alone constitutes a risk factor for T21 nondisjunction.

CENP-A ChIP-seq signal and CDRs are thought to predict the site of kinetochore attachment[5,7], and we find a robust and consistent single signal for the smallest centromere (H2) in both proband (~351-fold enrichment) and mother (~567-fold enrichment), indicating competence in forming the kinetochore. Despite the fact that the mother and proband are nearly identical (>99.99%) (**Supplementary Figure 2**) at the sequence level for the longer centromere haplotype (H1), we find that the CDRs differ markedly in their methylation profile. In the individual with T21, we find a notable reduction in CDR window length and an increase in the number of potential hypomethylated pockets. Surprisingly, the CDR differences do not, by themselves, predict the intensity of CENP-A signal since both the maternal and proband haplotypes show quantitatively weaker signals (39- and 139-fold enrichment, respectively). This suggests that kinetochore attachment dysfunction may have already appeared on the larger haplotype in the mother and potentially within her germ cells. We hypothesize that the asymmetry in the size of the maternal H1 and H2 may have catalyzed preferential kinetochore

formation on the shorter haplotype as evidenced by CENP-A ChIP-seq signal imbalance, promoting nondisjunction via precocious centromere segregation. This imbalance subsequently manifests itself epigenetically as both a disorganized CDR and a weaker CENP-A signal in the child.

The T21 family centromere characterization allowed us to discover intergenerational differences between centromeres that were previously unknown. In addition to the epigenetic changes observed in mother and proband for H1, we observed significant differences in the size of proband maternal H2 and paternal H3 CDRs as well as a positional shift of ~11 kbp and 25 kbp towards the p-arm in the proband for the former and the latter, respectively (**Figure 2c**), despite the fact that the sequence and structure between parent and proband were practically identical (**Supplementary Figure 2**). This reinforces the hypothesis that CDRs and kinetochore position are not bounded to a specific sequence but can change location potentially even within a single generation.

Chr21 centromeres show remarkable variability in length, α-satellite HOR array structure, and CDR patterns within the human population. While a complete census of human centromere length variation is underway as part of efforts to generate T2T human genome assemblies[23,24], our analysis to date suggests that T21 is amongst the top four most variable centromeres in the human genome[7]. In our survey of 40 complete chr21 centromeres (including the T21 family and CHM13), we identify six distinct centromeres shorter than 500 kbp. Remarkably, phylogenetic analyses suggest that all short chr21 centromeres share a monophyletic origin estimated to have emerged ~15 kya. This clade also stands out as showing the greatest variation in length. This implies some of the most rapid rates of expansion and contraction (~7-fold and ~10-fold difference between short proband H2 centromere and the average length inside and outside the 15 kya clade, respectively) observed to date for any centromere. It is interesting that both the

largest H1 and shortest H2 centromeres associated with T21 in this study arise from this very young clade. This could suggest that genetic background, including a high degree of sequence identity in conjunction with length asymmetry, may also contribute to T21 risk in mothers.

There are several obvious limitations to this study. First, we only analyzed one T21 family, and a much larger sample size is required to establish significance as an etiologic factor and how these features may relate to advanced maternal age—the major risk factor associated with T21[16]. It will be interesting, for example, to determine whether this asymmetry is more pronounced in younger mothers when compared to older ones. In the family studied here, the maternal age at the time of conception was ~29 years. Large T21 collections, including cell lines and blood DNA, have been established for mothers and probands that resulted from MMI and MMII errors as part of the National Down Syndrome Project[26], and such resources will be invaluable for follow-up testing.

While this is still an n-of-one study, it is noteworthy that three previous independent reports investigated the frequency of small chr21 HOR arrays on trisomic individuals. They revealed a higher frequency of short chr21 α-satellite HOR arrays in T21 individuals compared to the general population[17,25] and suggested an association of small chr21 α-satellite HOR arrays and MMI errors[27]. Notwithstanding, a 2003 study explored the relationship between the size of α-satellite HOR arrays on chr21 centromeres and repressed recombination, one of the known phenomena that correlate with nondisjunction, and concluded that the features are independent[28]. Our results suggest an alternative hypothesis, where centromere size asymmetry may be a more important consideration and possibly one of the several features that contribute to suppressed recombination and nondisjunction. Testing this will require accurate resolution of T21 centromeres, their sequence composition, and epigenetic features.

A second limitation of this study is that we examined transformed lymphoblastoid cell lines as opposed to primary material (e.g., blood). Relatively little is known regarding the stability of the CDRs during cell culture and how this might change depending on the tissue source. Also, our ability to identify functional centromeres and competent kinetochores via epigenetic traits, such as CpG methylation or CENP-A enrichment, should be further explored. Our results challenge the hypothesis that CENP-A enrichment always correlates with CpG hypomethylation within the CDR. This might be true for most centromeres but might differ for haplotypes predisposed to nondisjunction as seen in the T21 mother. Lastly, we surveyed a relatively small group of human diversity. More population controls are needed to understand the variation in α-satellite HOR array length, composition, and epigenetic characteristics.

Acknowledgments

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

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

COI (Conflicts of Interest) Statement

E.E.E. is a scientific advisory board (SAB) member of Variant Bio, Inc.

MATERIAL AND METHODS

Samples recruitment and data collection. The following lymphoblastoid cell lines were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research: proband (AG16777), mother (AG16778), and father (AG16782). The proband is described as a mosaic with karyotype: 47,XX,+21[21]/47,XX,+21,t(21;22)(q22;q13)[29] with 10% of the cells examined showing random chromosome loss and the karyotype confirming the diagnosis. Parents are unaffected and have a normal karyotype. The proband was 43 years old at sampling, while mother and father were 72 and 74, respectively. We identified 35 complete chr21 centromeres from genome assemblies generated as part of the HPRC[23] (n=7) and the HGSVC[24] (n=18).

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

UL-ONT sequencing. Ultra-high molecular weight gDNA was extracted from the lymphoblastoid cell lines according to a previously published protocol[29]. Briefly, 3-5 × 10^7 cells 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, and 20 mg/mL Rnase A for 1 hour at 37°C. 200 ug/mL Proteinase K was added, and the solution was incubated at 50°C for 2 hours. DNA was purified via two rounds of 25:24:1 phenol-chloroform-isoamyl alcohol extraction followed by ethanol precipitation. Precipitated DNA was solubilized in 10 mM Tris (pH 8.0) containing 0.02% Triton X-100 at 4°C for two days. Libraries were constructed using the Ultra-Long DNA Sequencing Kit (ONT, SQK-ULK001) with modifications to the manufacturer's protocol: ~40 ug of DNA was mixed with FRA enzyme and FDB buffer as described in the protocol and incubated for 5 minutes at RT, followed by a 5-minute heat-inactivation at 75°C. RAP enzyme was mixed with the DNA solution and incubated at RT for 1 hour before the clean-up step. Clean-up was performed using the Nanobind UL Library Prep Kit (Circulomics, NB-900-601-01) and eluted in 450 uL EB. 75 uL of library was loaded onto a primed FLO-PRO002 R9.4.1 flow cell for sequencing on the PromethION, with two nuclease washes and reloads after 24 and 48 hours of sequencing. Sequence reads >100 kbp in length were classified as ultra-long (UL). All ONT data were basecalled using guppy (v6.3.7) with the SUP model.

Genome assembly. T21 proband genome assembly was generated with the Verkko[20] (v1.2) (https://github.com/marbl/verkko) workflow using HiFi sequencing data at ~59× and UL-ONT (>100 kbp) data at ~23× coverage following the developer instructions. Parental assemblies were generated with both Verkko[20] (v1.4) (https://github.com/marbl/verkko) and hifiasm[30] (v0.19.5) (https://github.com/chhylp123/hifiasm) using HiFi data at ~37× for the mother and at

40× for the father and UL-ONT data at 25× and 12×, respectively. HiFi developer instructions were followed.

Centromere characterization and validation. T21 family and population assemblies were aligned against the T2T-CHM13[8] (v2.0) reference genome with minimap2[31] (v2.24) (https://github.com/lh3/minimap2) using the following options "-t {threads} -I 15G -a –eqx -x asm20 -s 5000". Fasta sequences from the contigs mapping against chr21 centromere (chr21:10700000-11650000) and flanking regions were extracted. RepeatMasker[32] (v4.1.0) was used to identify the regions containing the α-satellite array (marked as "ALR/Alpha'') and HumAS-HMMER (https://github.com/fedorrik/HumAS-HMMER_for_AnVIL) to determine the α-satellite HOR array composition with the following command: hmmer-run.sh {path_to_directory_with_fasta} AS-HORs-hmmer3.0-170921.hmm {threads}. For the contigs aligning to the reference genome in the minus orientation, we converted the fasta sequences to their reverse complement before running the analysis. The outputs from HumAS-HMMER analyses were visualized with R[33] (v4.2.2) using the ggplot2 package[34].

We validated the centromeric contigs sequence aligning each sample's HiFi reads to their own genome assembly using pbmm2 (v1.1.0) (https://github.com/PacificBiosciences/pbmm2) and showing uniform read coverage across the centromeric region with NucFreq[35] (https://github.com/mrvollger/NucFreq). To confirm that the extracted centromeric contig sequences represented chr21 centromere repeats, we compared them to the α-satellite monomer units extracted from the T2T-CHM13[8] reference genome using StringDecomposer[36] (https://github.com/ablab/stringdecomposer) and showed that have high identity (>90%). Finally, we generated pairwise sequence identity heatmaps of each centromere using StainedGlass as previously described[21] (https://github.com/mrvollger/StainedGlass).

Centromere alignments and dotplots. Proband and parental centromere sequences were aligned using minimap2[31] (v2.26) and the following command: minimap2 -cx asm5 {proband_centromeric_fasta} {corresponding_parental_centromeric_fasta} > alignment.paf. The results in .paf format were visualized with R[33] (v4.2.2) using the ggplot2 package[34].
Proband H3 and father H3 AluY centromeric insertions with 10 kbp padding sequence on both sides were aligned with BLAST[37].

CpG methylation and centromeric dip region (CDR) analysis. All ONT sample data were basecalled with guppy v6.3.7 or newer to generate BAM files containing methylation tags linked to each read. HiFi data were basecalled using ccs (v.6.4.0) (https://github.com/PacificBiosciences/ccs) and Jasmine (2.0.0) (https://github.com/PacificBiosciences/jasmine), which generated the equivalent BAM files containing the methylation tags linked to each read. Jasmine 5mC predictions were used as a validation for the T21 family centromeric methylation profile. Briefly, we extracted the reads for each sample and aligned them to their respective genome assembly using Winnowmap[38] (v2.03). We then used modbam2bed (https://github.com/epi2me-labs/modbam2bed) to obtain bedMethyl files.

Using the bedMethyl files as an input and a modified version of CDR-Finder (https://github.com/arozanski97/CDR-Finder), we identified the CDRs for each centromere. The tool first divides the region of interest (the centromeric α-satellite arrays with padding sequence on both sides, e.g., coordinates for Proband H1: unassigned-0003018:0-2000000) in sequential 5 kbp bins and calculates their average methylation frequency (if a bin does not have an assigned value, the bin is excluded). It then computes the median CpG methylation frequency for each window containing α-satellite sequences (based on RepeatMasker[32] output). It selects

the bins with a lower methylation frequency than the median of the region and merges consecutive bins into candidate CDRs. These candidate CDRs are then evaluated on two criteria: whether or not they have flanking windows on either side with greater than the maximum percent methylation (defined as within 1 standard deviation of the maximum) for the entire region and if they clear a user-specified size threshold (25 kbp). Each call is annotated with a confidence label: “high confidence” if it clears both criteria or “low confidence” if it does not. The low-confidence labels are appended with either “neighbor_peaks” for the first criterion and/or “size” for the second criterion. In small centromeres where the median can be quite low due to the CDR taking up the majority of the centromeric sequence space, the initial search algorithm of taking windows which are below the median is flipped to identify the windows that are above the median and thus not part of the CDR. After removing these windows from consideration, what remains are candidate CDRs, and the algorithm continues as normal. All CDR calls were manually confirmed inspecting the read alignments. False positives at the boundaries of the centromeres (n = 3) as well as calls not confirmed by the reads and due to low coverage were excluded (n = 1).

For each T21 family CDR, we calculated the percentage of methylated CpGs (defined as having ≥25% methylated CpGs at the same position across all UL-ONT reads) within the total CDR window (from start of the first CDR window to the end of the last CDR window). Formula = (number of methylated CpGs within the CDR window) * 100 / (number of CpGs within the CDR window).

Native CENP-A ChIP-seq and analysis. To determine the location of centromeric chromatin within the mother’s, father’s, and proband’s genomes, we performed two independent replicates of native CENP-A ChIP-seq for each cell line as described previously[5], with some

modifications. Results from Replicate 2 are shown in Figures 1 and 2. Briefly, 3-4 × 10^7 cells were collected and resuspended in 2 mL of ice-cold buffer I (0.32 M sucrose, 15 mM Tris, pH 7.5, 15 mM NaCl, 5 mM $MgCl_2$, 0.1 mM EGTA, and 2x Halt Protease Inhibitor Cocktail (Thermo Fisher, 78429)). 2 mL of ice-cold buffer II (0.32 M sucrose, 15 mM Tris, pH 7.5, 15 mM NaCl, 5 mM $MgCl_2$, 0.1 mM EGTA, 0.1% IGEPAL, and 2x Halt Protease Inhibitor Cocktail) was added, and samples were placed on ice for 10 min. The resulting 4 mL of nuclei were gently layered on top of 8 mL of ice-cold buffer III (1.2 M sucrose, 60 mM KCl, 15 mM, Tris pH 7.5, 15 mM NaCl, 5 mM $MgCl_2$, 0.1 mM EGTA, and 2x Halt Protease Inhibitor Cocktail (Thermo Fisher, 78429)) and centrifuged at 10,000g for 20 min at 4°C. Pelleted nuclei were resuspended in buffer A (0.34 M sucrose, 15 mM HEPES, pH 7.4, 15 mM NaCl, 60 mM KCl, 4 mM MgCl2, and 2x Halt Protease Inhibitor Cocktail) to 400 ng/mL. Nuclei were frozen on dry ice and stored at 80°C. Mnase digestion reactions were carried out on 100-150 µg chromatin, using 0.1-0.2 U/µg Mnase (Thermo Fisher, 88216) in buffer A supplemented with 3 mM $CaCl_2$ for 10 min at 37°C. The reaction was quenched with 10 mM EGTA on ice and centrifuged at 500g for 7 min at 4°C. The chromatin was resuspended in 10 mM EDTA and rotated at 4°C for 2 h. The mixture was adjusted to 500 mM NaCl, rotated for another 45 min at 4°C and then centrifuged at max speed (21,100g) for 5 min at 4°C, yielding digested chromatin in the supernatant. Chromatin was diluted to 100 ng/ml with buffer B (20 mM Tris, pH 8.0, 5 mM EDTA, 500 mM NaCl and 0.2% Tween 20) and precleared with 100 µL 50% protein G Sepharose bead (Abcam, ab193259) slurry for 20 min at 4°C, rotating. Precleared supernatant (10-20 µg bulk nucleosomes) was saved for further processing. To the remaining supernatant, 10 µg mouse monoclonal anti-human CENP-A antibody (Enzo, ADI-KAM-CC006-E) was added and rotated overnight at 4°C. Immunocomplexes were recovered by the addition of 200 mL 50% protein G Sepharose bead slurry followed by rotation at 4°C for 3 h. The beads were washed 3× with buffer B and once with buffer B without Tween 20. For the input fraction, an equal volume of input recovery buffer (0.6 M NaCl, 20 mM EDTA, 20 mM Tris, pH 7.5, and 1% SDS) and 1 mL of Rnase A (10

mg/mL) was added, followed by incubation for 1h at 37°C. Proteinase K (100 mg/ml, Roche) was then added, and samples were incubated for another 3 h at 37°C. For the ChIP fraction, 300 µL of ChIP recovery buffer (20 mM Tris, pH 7.5, 20 mM EDTA, 0.5% SDS and 500 mg/mL Proteinase K) was added directly to the beads and incubated for 3-4 h at 56°C. The resulting Proteinase K-treated samples were subjected to a phenol-chloroform extraction followed by purification with a Qiagen MinElute PCR purification column. Unamplified bulk nucleosomal and ChIP DNA were analyzed using an Agilent Bioanalyzer instrument and a 2100 High Sensitivity Kit.

Sequencing libraries were generated using the TruSeq ChIP Library Preparation Kit – Set A (Illumina, IP-202-1012) according to the manufacturer's instructions, with some modifications. Briefly, 5–10 ng bulk nucleosomal or ChIP DNA was end-repaired and A-tailed. Illumina TruSeq adaptors were ligated, libraries were size-selected to exclude polynucleosomes using an E-Gel SizeSelect II agarose gel, and the libraries were PCR-amplified using the PCR polymerase and primer cocktail provided in the kit. The resulting libraries were submitted for 150 bp, paired-end Illumina sequencing using a NextSeq 500/550 High Output Kit v2.5 (300 cycles). The resulting reads were assessed for quality using FastQC[39] (https://github.com/s-andrews/FastQC), trimmed with Sickle (v1.33) (https://github.com/najoshi/sickle) to remove low-quality 5' and 3' end bases, and trimmed with Cutadapt6[40] (v1.18) to remove adapters.

Processed CENP-A ChIP and bulk nucleosomal reads were aligned to the whole-genome assembly (WGA) from the same sample using BWA-MEM[41] (v0.7.17) and the following parameters: bwa mem -k 50 -c 1000000 {index} {read1.fastq.gz} {read2.fastq.gz}. The resulting SAM files were filtered using SAMtools[42] (v1.9) with flag score 2308 to prevent multi-mapping of reads. With this filter, reads mapping to more than one location are randomly assigned a single mapping location, thereby preventing mapping biases in highly identical regions. Alignments

were normalized and filtered with deepTools[43] (v3.4.3) bamCompare with the following command: bamCompare -b1 {ChIP.bam} -b2 {bulk_nucleosomal.bam} --operation ratio --binSize 1000 --minMappingQuality 1 -o {out.bw}.

The fold enrichment of CENP-A ChIP reads over bulk nucleosomal reads was calculated for all chr21 centromere haplotypes in the parent-child trio. To do this, we first selected a window encompassing the CDRs of both the proband and the corresponding parental haplotype (**Supplementary Table 2**). We then used samtools[42](v.1.9) bedcov with the option "--min-MQ 1" to calculate the total read base count over the selected window for both the CENP-A ChIP and the bulk nucleaosomal aligned reads. We divided the CENP-A ChIP reads value by the bulk nucleaosomal reads value to obtain the fold enrichment for each window.

Contig fixing and patching. Centromeric contigs that showed misassemblies or collapses during NucFreq[35] analysis (see above) were corrected following this procedure. Sequencing reads mapping to the misassembly/collapse and flanking regions were extracted and locally assembled using the module Canu from the tool HiCanu[44]. The local assemblies were validated using NucFreq[35] as described previously and inserted in the original assembly.
T21 parents Verkko[20] genome assemblies were corrected with an alternative assembly generated via hifiasm[30]. Briefly, we generated a WGA of each genome with hifiasm[30] (v0.19.5) and the following command: "hifiasm -o {output} -t {threads} –ul {ultra-long read fastq} {hifi reads fastq}". Then, we aligned the hifiasm WGA to the Verkko WGA with minimap2[31] (v2.24) and the following command: minimap2 -t {threads} -I 15G -a –eqx -x asm20 -s 5000. We identified the corresponding locations of the centromeric regions in each assembly and then determined the α-satellite HOR array composition of each with HumAS-HMMER (https://github.com/fedorrik/HumAS-HMMER_for_AnVIL) and the following command: hmmer-run.sh {path_to_directory_with_fasta} AS-HORs-hmmer3.0-170921.hmm {threads}. We

compared the α-satellite HOR compositions of each array and then replaced erroneous sequences in the Verkko assembly with the corresponding sequence in the hifiasm assembly. To ensure that we had resolved the assembly error, we aligned PacBio HiFi reads from the same genome to the new assembly and then ran NucFreq[35] to validate the assembly.

Phylogenetic analysis. For the phylogenetic reconstruction of chr21 centromere haplotypes, we first aligned a chimpanzee assembly to the T2T-CHM13[8] reference genome using Minimap2[31] with the following options "-t {threads} -I 15G -x asm5 -a –secondary=no". We identified a chr21 pericentromeric region on the q-arm where only two chimpanzee contigs mapped (chr21:11855000-11875000). The chimpanzee alignment was used as an outgroup for the analysis. We aligned all the human assemblies available against the T2T-CHM13[8] reference genome as described above and selected only the samples that have chr21 centromeric contigs spanning the selected pericentromeric window. We then extracted the sequences mapping to the pericentromeric region of interest of each sample. Using these sequences, we generated a multi-alignment with the tool MAFFT[45,46] (v7.453) using these options: "–auto –thread {threads}". The maximum-likelihood phylogeny tree was constructed with IQ-TREE[47] (v2.1.2) with the following options: "-T {threads} -m MFB -B 1000". Divergence time amongst haplotypes were estimated with IQ-TREE[47] (options: "–keep-ident –date-tip 0 –date-ci 100 -redo") using the MAFFT multi-alignment, the consensus tree, and a time estimate file containing the known divergence time between human and chimpanzee (6.4 million years) as inputs. The resulting tree was visualized with iTOL[48].

Data availability. Sequence, assembly, and native CENP-A ChIP-seq data are being submitted to the database of Genotypes and Phenotypes (dbGaP).

Supplementary Table 1. Table of all the HOR monomers identified in all chromosome 21 centromeres described and their relative percentages.

Supplementary Table 2. Summary of α-satellite coordinates, HOR array size, and CDR coordinates of all chromosome 21 centromeres described.

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. 2023.05.30.542849 Preprint at https://doi.org/10.1101/2023.05.30.542849 (2023).
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. Rhie, A. *et al.* The complete sequence of a human Y chromosome. *Nature* **621**, 344–354 (2023).
12. Hallast, P. *et al.* Assembly of 43 human Y chromosomes reveals extensive complexity and variation. *Nature* **621**, 355–364 (2023).
13. 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).
14. 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).
15. Gershman, A. *et al.* Epigenetic patterns in a complete human genome. *Science* **376**, eabj5089 (2022).
16. 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).
17. Contreras-Galindo, R. *et al.* Rapid molecular assays to study human centromere genomics. *Genome Res* **27**, 2040–2049 (2017).
18. Angell, R. R. Predivision in human oocytes at meiosis I: a mechanism for trisomy formation in man. *Hum Genet* **86**, 383–387 (1991).
19. Smith, K. N. & Nicolas, A. Recombination at work for meiosis. *Current Opinion in Genetics & Development* **8**, 200–211 (1998).
20. Rautiainen, M. *et al.* Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol* **41**, 1474–1482 (2023).
21. 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).
22. Smith, G. P. Evolution of Repeated DNA Sequences by Unequal Crossover. *Science*

191, 528–535 (1976).
23. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
24. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
25. 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).
26. Freeman, S. B. *et al.* The National down Syndrome Project: Design and Implementation. *Public Health Rep* **122**, 62–72 (2007).
27. 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).
28. 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).
29. Glennis Logsdon on protocols.io. *protocols.io* https://www.protocols.io/researchers/glennis-logsdon.
30. 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).
31. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
32. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015.
33. R Development Core Team, R. *R: A Language and Environment for Statistical Computing*. *R Foundation for Statistical Computing* vol. 1 409 (2011).
34. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer, New York, NY, 2009). doi:10.1007/978-0-387-98141-3.
35. Vollger, M. R. *et al.* Long-read sequence and assembly of segmental duplications. *Nat Methods* **16**, 88–94 (2019).
36. 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).
37. 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).
38. 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).
39. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
40. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
41. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv.org* https://arxiv.org/abs/1303.3997v2 (2013).
42. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
43. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Research* **42**, W187–W191 (2014).
44. 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.
45. 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).
46. Nakamura, T., Yamada, K. D., Tomii, K. & Katoh, K. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* **34**, 2490–2492 (2018).
47. 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).
48. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128 (2007).

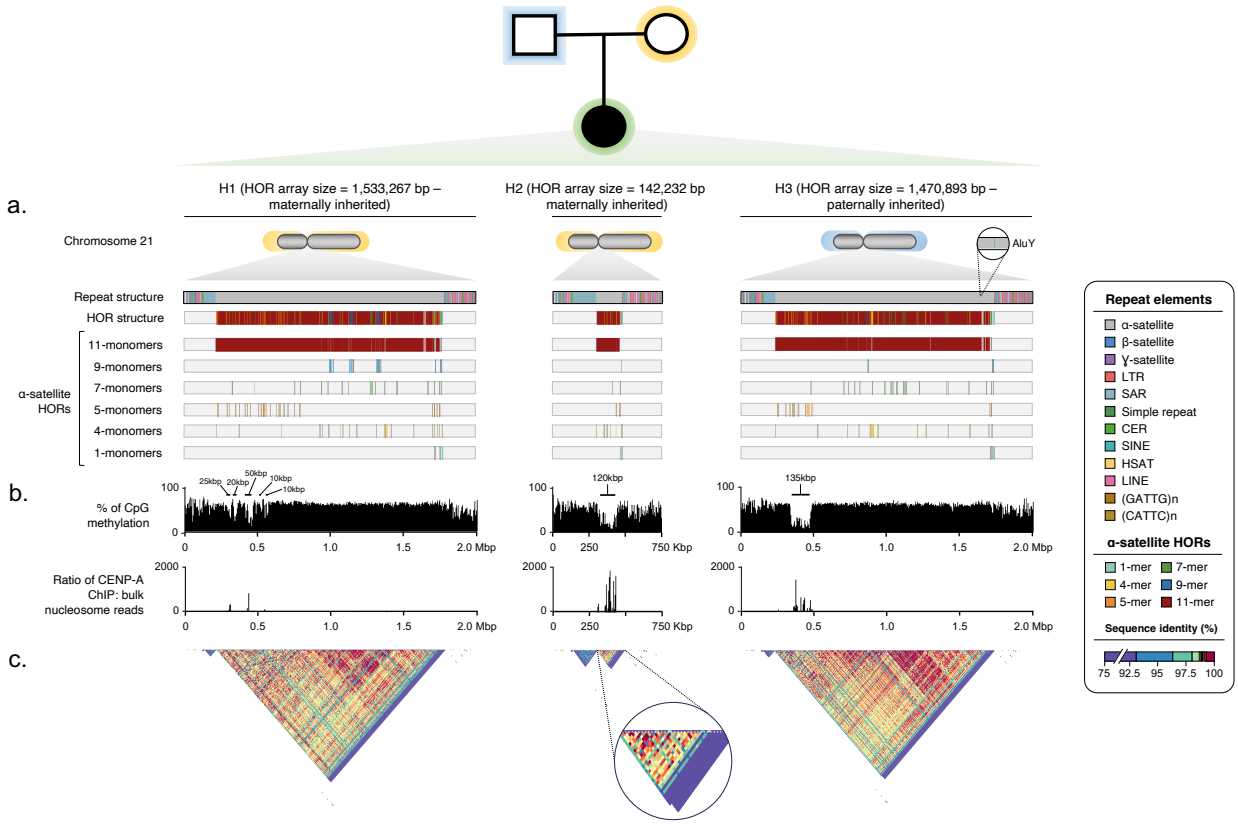


Figure 1. Genetic composition and epigenetic profile of three chromosome 21 (chr21) centromeres from a Trisomy 21 (T21) individual. a) Size and repeat composition of three centromeric haplotypes for a T21 individual (Coriell ID: AG16777). Different α-satellite higher-order repeat (HOR) monomers mapping within the α-satellite array are colored (see legend). Only α-satellite HOR monomers present more than 10 times in at least one haplotype are shown (see **Supplementary Table 1** for complete monomer composition). **b)** Epigenetics profile of each centromere showing hypomethylated pockets enriched with CENP-A chromatin based on CENP-A ChIP-seq. **c)** Pairwise sequence identity heatmaps (StainedGlass[21]) depicting sequence identity in 5 kbp windows across the centromeres reveal regions of high- and low-identity homology within each α-satellite HOR array (see legend).

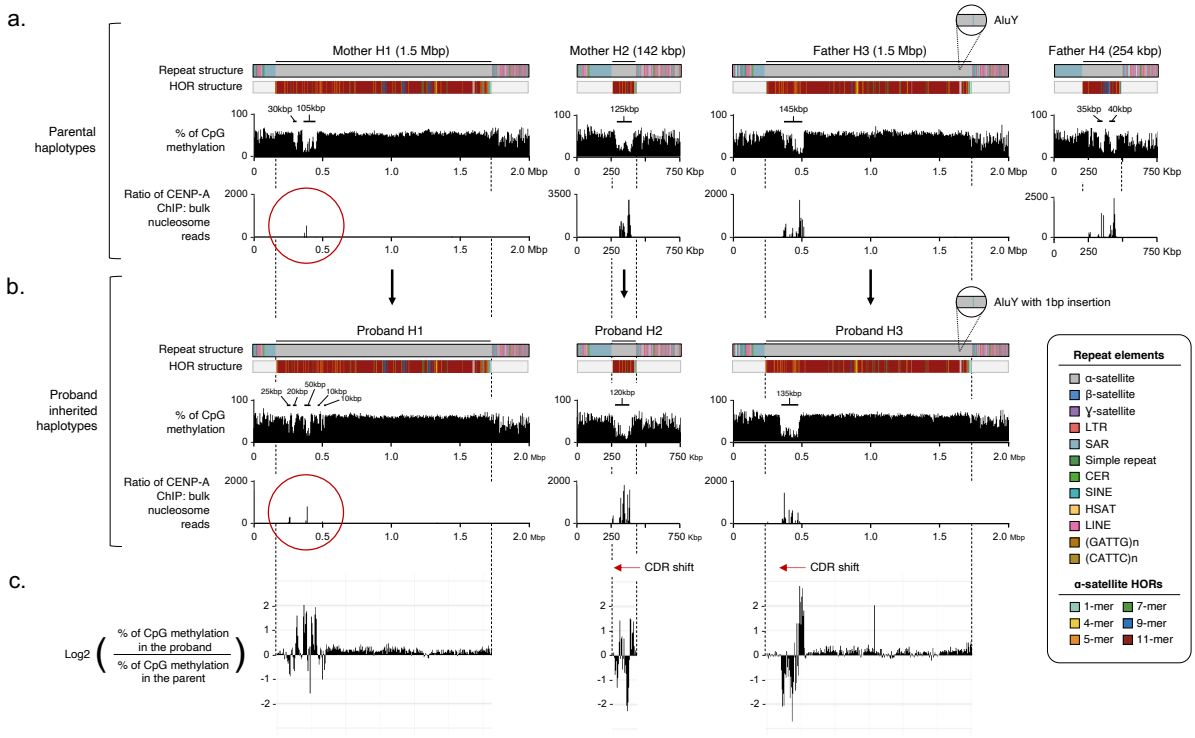
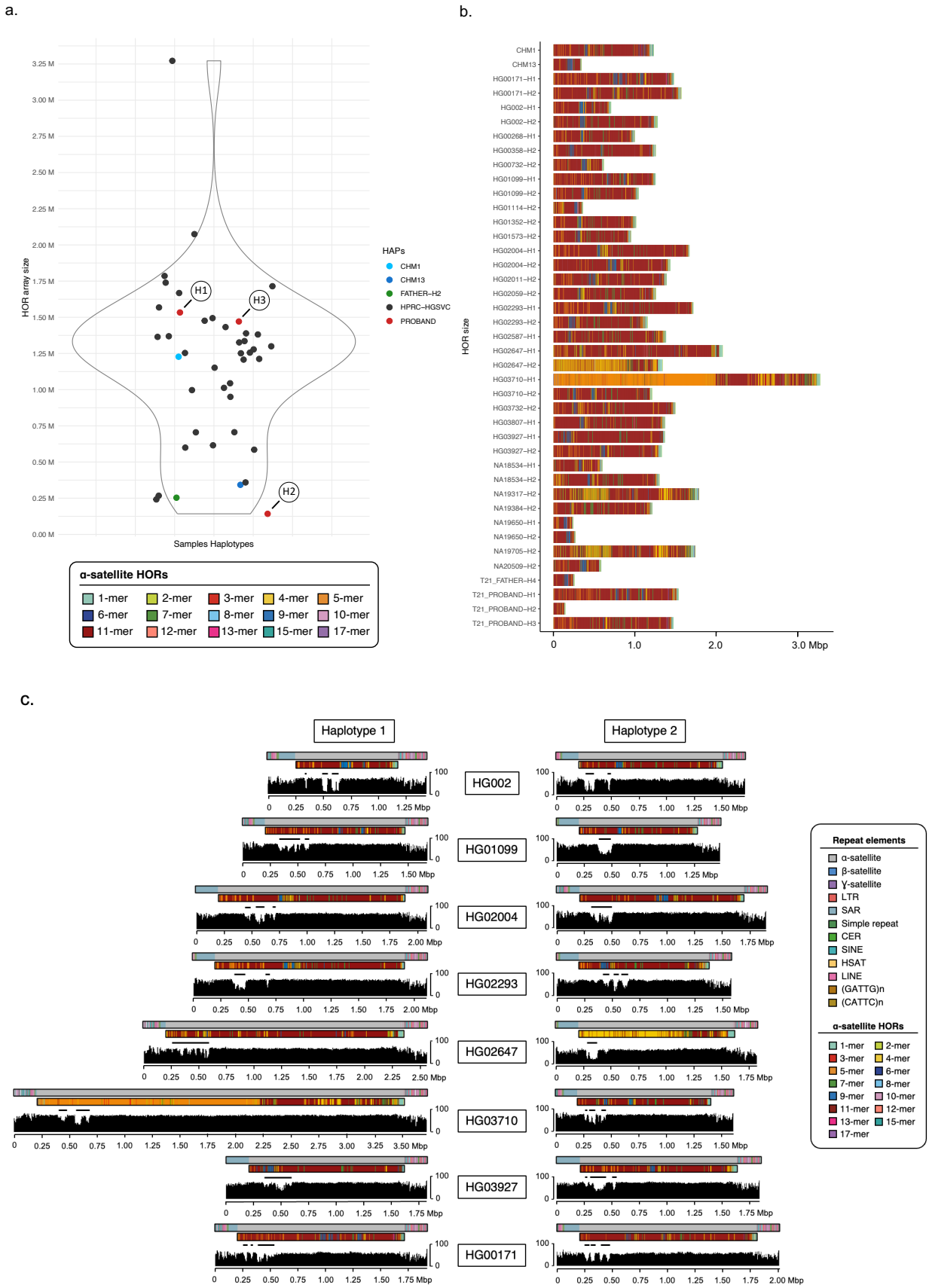


Figure 2. Genetic and epigenetic features of parent–proband transmission of chr21 centromeres. Composition and organization of α-satellite HOR arrays (only α-satellite HORs present more than 10 times in at least one haplotype are shown; see **Supplementary Table 1** for complete HOR composition), CpG methylation profiles, and CENP-A enrichment patterns of **a)** parental chr21 centromeres and **b)** transmitted proband chr21 centromeres. Low CENP-A ChIP-seq fold-enrichment is highlighted with red circles. **c)** Log_2 ratio of CpG methylation frequency between proband and parental haplotypes.


a.
HOR array size
Samples Haplotypes
H1
H2
H3
HAPs
CHM1
CHM13
FATHER–H2
HPRC–HGSVC
PROBAND
α-satellite HORs
1-mer
2-mer
3-mer
4-mer
5-mer
6-mer
7-mer
8-mer
9-mer
10-mer
11-mer
12-mer
13-mer
15-mer
17-mer
b.
HOR size
3.0 Mbp
c.
Haplotype 1
Haplotype 2
HG002
HG01099
HG02004
HG02293
HG02647
HG03710
HG03927
HG00171
Repeat elements
α-satellite
β-satellite
γ-satellite
LTR
SAR
Simple repeat
CER
SINE
HSAT
LINE
(GATTG)n
(CATTC)n
α-satellite HORs

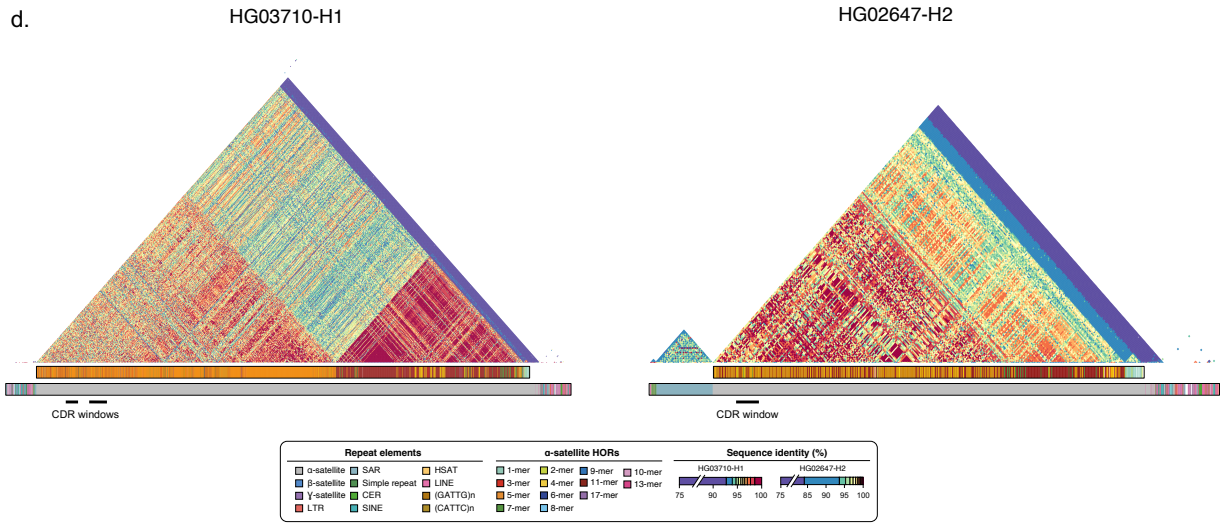


Figure 3. Diversity in α-satellite HOR array length, composition, and CpG methylation profile amongst diverse chr21 centromere haplotypes. Comparison of chr21 α-satellite HOR array **a)** length and **b)** composition amongst 35 population and 2 hydatidiform mole haplotypes compared to the T21 family haplotypes (only α-satellite HORs present more than 10 times in total are shown; see **Supplementary Table 1** for complete HOR composition). **c)** CpG methylation profiles of 16 population haplotypes (see also **Extended Figure 3c**) show a consistent location of the CDR in the first half of the centromeric array. **d)** Pairwise sequence identity heatmaps (generated with StainedGlass[21] using 5 kbp windows) across chr21 centromeres enriched with 5-monomer and 4-monomer α-satellite HORs (see legend). In both cases, the CDRs map to either the 4- or 5-monomer HOR-enriched regions as opposed to the 11-monomer HOR-enriched region.

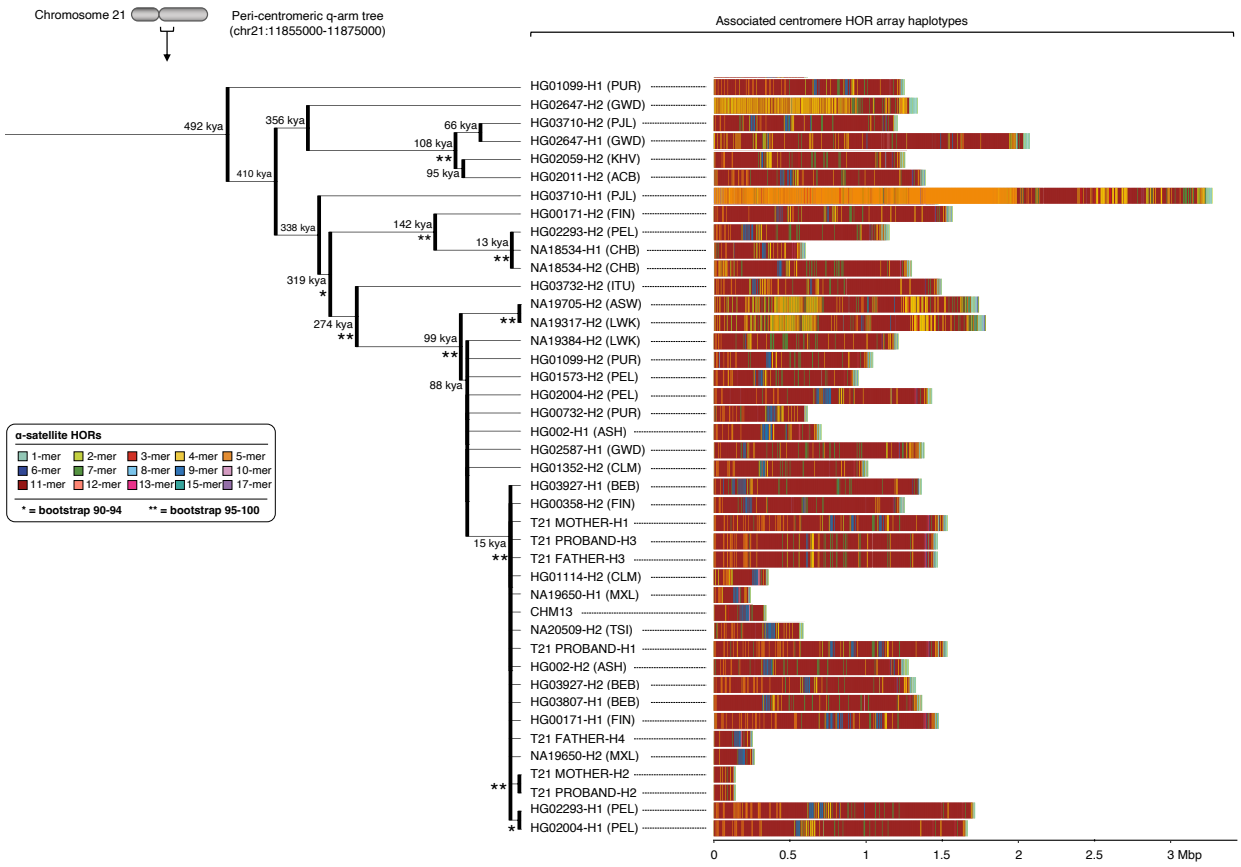


Figure 4. Phylogenetic reconstruction of chr21 centromere haplotypes. Maximum-likelihood phylogenetic tree based on a multiple sequence alignment of a 20 kbp segment from the pericentromeric q-arm regions from 38 completely sequenced chr21 centromeres. The HOR array structure of each is depicted and estimated divergence time for each cluster is projected onto the nodes. Bootstrap values 90 to 94 (*) and 95 to 100 (**) are marked with asterisks.