Alpha satellite insertions and the evolutionary landscape of centromeres

Giuliana Giannuzzi[1,2,3], Glennis A. Logsdon[4], Nicolas Chatron[3,5,6], Danny E. Miller[4,7], Julie Reversat[5], Katherine M. Munson[4], Kendra Hoekzema[4], Marie-Noëlle Bonnet-Dupeyron[8], Pierre-Antoine Rollat-Farnier[5,9], Carl A. Baker[4], Damien Sanlaville[5,6], Evan E. Eichler[4,10], Caroline Schluth-Bolard[5,6], Alexandre Reymond[3]

1. Department of Biosciences, University of Milan, Milan, Italy
2. Institute of Biomedical Technologies, National Research Council, Milan, Italy
3. Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland
4. Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA
5. Service de génétique, Hospices Civils de Lyon, Lyon, France
6. Institut NeuroMyoGène, University of Lyon, Lyon, France
7. Department of Pediatrics, Division of Genetic Medicine, University of Washington and Seattle Children's Hospital, Seattle, WA, USA
8. Service de Cytogénétique, Centre Hospitalier de Valence, Valence, France
9. Cellule Bioinformatique, Hospices Civils de Lyon, Lyon, France
10. Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA

Correspondence to:

Giuliana Giannuzzi
University of Milan
Department of Biosciences
Via Giovanni Celoria, 26
20133 Milano, Italy
giuliana.giannuzzi@unimi.it

Abstract

Human centromeres are composed of alpha satellite DNA hierarchically organized as higher-order repeats and epigenetically specified by CENP-A binding. Current evolutionary models assert that new centromeres are first epigenetically established and subsequently acquire an alphoid array. We identified during routine prenatal aneuploidy diagnosis by FISH a *de novo* insertion of alpha satellite DNA array (~50-300 kbp) from the centromere of chromosome 18 (D18Z1) into chromosome 15q26 euchromatin. Although bound by CENP-B, this locus did not acquire centromeric functionality as demonstrated by lack of constriction and absence of CENP-A binding. We characterized the rearrangement by FISH and sequencing using Illumina, PacBio, and Nanopore adaptive sampling which revealed that the insertion was associated with a 2.8 kbp deletion and likely occurred in the paternal germline. Notably, the site was located ~10 Mbp distal from the location where a centromere was ancestrally seeded and then became inactive sometime between 20 and 25 million years ago (Mya), in the common ancestor of humans and apes. Long reads spanning either junction showed that the organization of the alphoid insertion followed the 12-mer higher-order repeat structure of the D18Z1 array. Mapping to the CHM13 human genome assembly revealed that the satellite segment transposed from a specific location of chromosome 18 centromere. The rearrangement did not directly disrupt any gene or predicted regulatory element and did not alter the epigenetic status of the surrounding region, consistent with the absence of phenotypic consequences in the carrier. This case demonstrates a likely rare but new class of structural variation that we name 'alpha satellite insertion'. It also expands our knowledge about the evolutionary life cycle of centromeres, conveying the possibility that alphoid arrays can relocate near vestigial centromeric sites.

Introduction

Alpha satellite is a class of highly repetitive DNA defined by a group of related, highly divergent AT-rich repeats or 'monomers', each approximately 171 bp in length. Alpha satellite, also named alphoid DNA, comprises up to 10% of the human genome and is mostly found tandemly repeated within constitutive heterochromatin at centromeres and pericentromeric regions. At centromeric regions, satellite monomers are hierarchically organized into larger repeating units, in which a defined number of monomers have been homogenized. These units, which are named 'higher-order repeats' (HORs), are tandemly arranged into chromosome-specific, megabase-sized satellite arrays with limited nucleotide differences between repeat copies (Willard and Waye 1987; Durfy and Willard 1989; Schueler et al. 2001; McNulty and Sullivan 2018; Miga et al. 2020).

The centromere is the chromosomal locus where sister chromatids attach and the kinetochore is assembled, which is essential for proper chromosome segregation during cell division. While alpha satellite DNA constitutes the sequence of all mature centromeres, it is not sufficient nor necessary for centromere identity. This is demonstrated by dicentric chromosomes that assemble the kinetochore at only one of two alpha-satellite regions (Earnshaw and Migeon 1985) and analphoid chromosomes that possess fully functional centromeres (Voullaire et al. 1993). Centromere function appears to be epigenetically established and maintained by local enrichment of the CENP-A histone H3 variant within nucleosomes rather than presence of alphoid DNA (Palmer et al. 1991; Karpen and Allshire 1997; Panchenko and Black 2009; McKinley and Cheeseman 2016). This function can be inactivated at an original site and moved to a new position along the chromosome (Montefalcone et al. 1999). It is similarly turned off after a chromosomal fusion to ensure stability of the derived dicentric chromosome. These events determine the emergence of evolutionary new centromeres and the appearance of recognizable genomic regions where the centromere used to be positioned in the past (Amor and Choo 2002; Rocchi et al. 2009). Insights into the molecular steps of centromere repositioning from the birth of a new centromere to its maturity were uncovered by studying fly, primate, and equid chromosomes (Marshall et al. 2008; Piras et al. 2010). These analyses showed that new centromeres are first epigenetically specified and then mature by acquiring the satellite DNA array, in some cases going through intermediate configurations bearing DNA amplification (Kalitsis and Choo 2012; Nergadze et al. 2018).

Besides the main pericentromeric and centromeric locations, smaller regions of alpha satellite DNA are located in the euchromatin of the human genome, >5 Mbp from the centromeres, with

around 100 blocks annotated in the reference by the RepeatMasker program (Rudd and Willard 2004; Feliciello et al. 2020). For example, three large blocks, respectively 11, 8, and 13 kbp long, are located within cytoband 2q21 with SVA (SINE/VNTR/Alu) and LINE elements intervening between them. These alphoid sequences are the relics of an ancestral centromere that became inactive ~5 Mya after the fusion of two ancestral chromosomes in the human lineage compared to big apes (IJdo et al. 1991; Avarello et al. 1992; Baldini et al. 1993; Chiatante et al. 2017).

Here, we report an individual with a *de novo* insertion of an alpha satellite DNA array from the centromere of chromosome 18 into chromosome 15q26, the first observation of insertion of satellite DNA array into the euchromatin of the human genome that we are aware of. This case brings to light a probably rare and new class of structural variation and expands our knowledge on the evolutionary life cycle of centromeres and the origin and spread of alpha satellite in primate genome.

Results

Prenatal, postnatal and family investigations

Amniocentesis was performed at 15 weeks' gestation in a 35 years-old gravida 6 para 2 woman. She already had two healthy children, one miscarriage, and two pregnancies terminated due to fetal trisomy 21. Interphase FISH (Fluorescent *In Situ* Hybridization) on uncultured amniocytes with probes for main aneuploidies showed the presence of three signals for the alpha satellite DNA probe of chromosome 18 (D18Z1) in all cells (150/150) and three signals for chromosome 21 specific probes in 29 out of 121 cells (24%), suggesting a trisomy 18 and a mosaic trisomy 21. Karyotyping of cultured cells confirmed the presence of the mosaic trisomy 21 at 19% (12/62 cells) but showed the presence of two normal chromosomes 18. Metaphase FISH on cultured cells revealed the aberrant hybridization of the D18Z1 probe at chromosome 15q26 (**Figure 1A**). The intensity and size of the FISH signal suggested that the length of the inserted satellite DNA was ~50-300 kbp. Chromosomal microarray did not show any imbalances, except the mosaic trisomy 21 (13%). FISH analysis of both parents showed that the alphoid DNA insertion was *de novo*. Pregnancy sonographic follow-up was normal. The proband, a healthy male baby, was born at term with normal birth parameters. Post-natal karyotype and FISH confirmed the mosaic trisomy 21 (6/33 cells; 18%) and the presence of the insertion of chromosome 18 alpha satellite on the long arm of a chromosome 15. At one

115 year old, growth clinical examination (weight 10.6 kg, +1 standard deviation (SD); height 75
116 cm, +1 SD; occipito-frontal circumference 46.5 cm, +1SD) and psychomotor development
117 were normal, consistent with low level mosaic trisomy 21 and also suggesting that the alpha
118 satellite insertion had no phenotypic consequences.

119

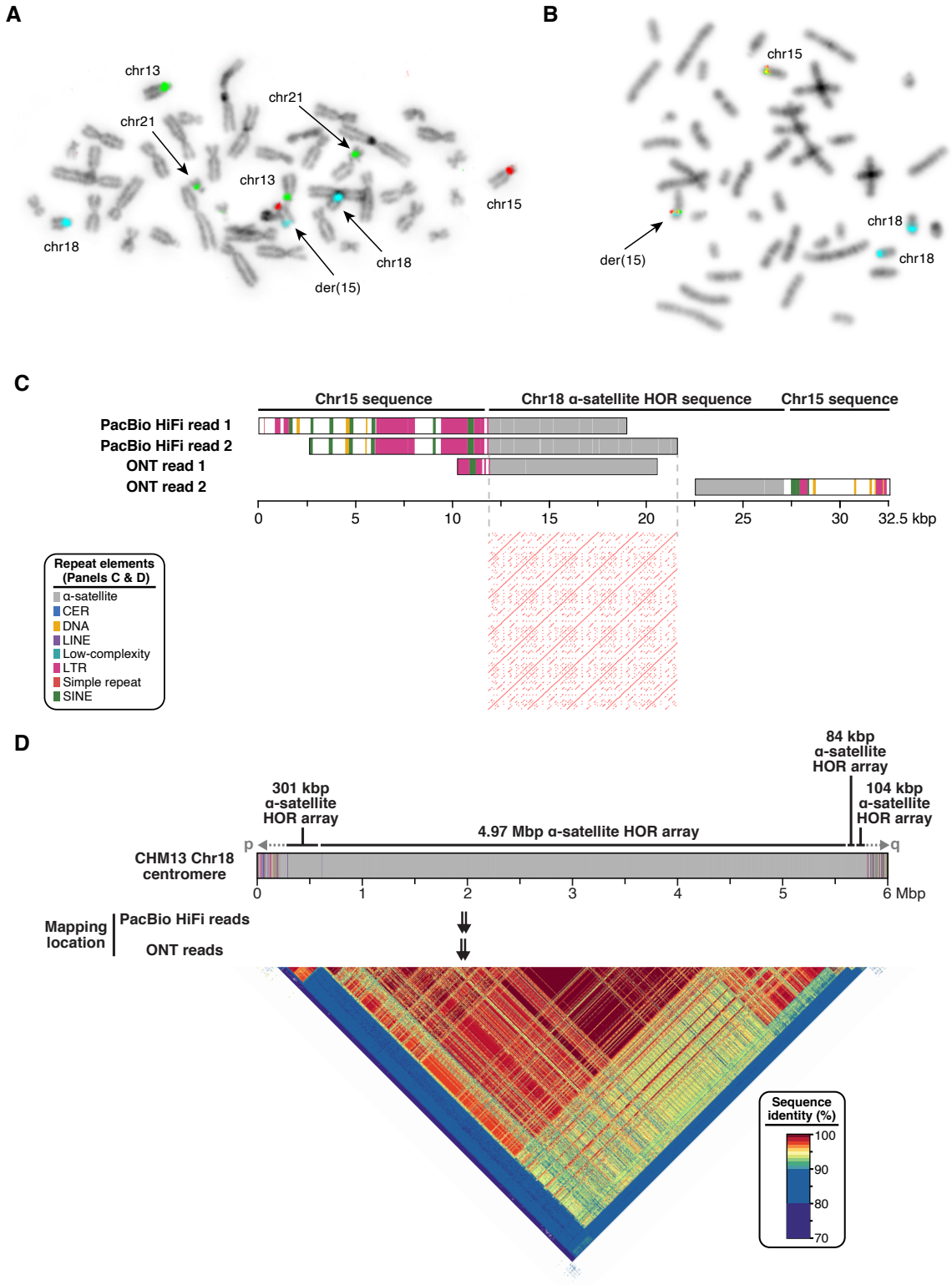

Figure 1. D18Z1 alpha satellite *de novo* insertion. **A)** FISH results of cultured amniocytes using alpha satellite DNA probes of chromosomes 15 (D15Z1, Texas-Red), 13/21 (D13/21Z1, green), and 18 (D18Z1, aqua) probes. **B)** FISH results of cultured amniocytes using the 15q25 BAC probes RP11-635O8 (red) and RP11-752G15 (green) flanking the ancestral centromere, and the D18Z1 (aqua) probe. **C)** Read length, repeat composition (color code in inset), and mapping location of the four selected HiFi and ONT reads (*top*). Dot plot (window size 20) of the longest available alpha satellite sequence (*bottom*). **D**) Schematic representation of the CHM13-T2T

chromosome 18 centromere with its repeat composition (*top*). A heatmap representation of sequence identity over the region is presented below. The mapping location of the PacBio HiFi and ONT reads is pinpointed by black arrows.

Structural characterization of the rearrangement

To characterize the alphoid DNA insertion at the sequence level, we performed WGS (whole genome sequencing) of the proband using the short-read Illumina platform. We first analyzed these data using a routine clinical analysis pipeline that did not identify any structural variant at chromosome 15q26. We then followed a customized approach, mapping reads to a library made up of the entire chromosome 15 and chromosome 18 centromeric alpha satellite DNA sequences. We isolated high-quality discordant paired reads mapped to both sequences, as well as chimeric reads anchored to chromosome 15 and containing alpha satellite DNA. These reads allowed us to define the positions of the proximal and distal breakpoints of the insertion at chr15:92,359,068 and chr15:92,361,920 (GRCh38), respectively. These coordinates, both subsequently validated by PCR, revealed the deletion of a 2,851 bp segment that was replaced by the insertion. We noted that the target site was ~10 Mbp distal from the position where an ancestral centromere was seeded and was shown to be active ~25 Mya in the common ancestor of Old World monkeys and apes, and was then inactivated sometime between 20 and 25 Mya in the common ancestor of the Hominoids (lesser apes, great apes, and humans) (Ventura et al. 2003). This was further confirmed by the co-hybridization of the D18Z1 probe with two BAC probes flanking the ancestral centromere locus (RP11-752G15 and RP11-635O8) (Giannuzzi et al. 2013). This experiment showed, at metaphase resolution, that the satellite probe signal colocalized with both BAC probes on the derivative chromosome 15 (**Figure 1B**).

We then sought to better characterize the rearrangement by generating long-read sequence information. We employed two technologies, ONT (Oxford Nanopore Technologies) with selective sampling via Read Until (Loose et al. 2016), targeting 50 kb of sequence on either side of the insertion, and PacBio HiFi sequencing. We sequenced the proband (~11.5x coverage at the targeted region), father (~20.1x), and mother (~19.8x) using readfish (Payne et al. 2020) on an ONT GridION, and the proband's genome on one PacBio SMRT cell (~6.5x coverage). We confirmed the insertion breakpoints and the 2.8 kbp deletion but were unable to assemble a contiguous sequence spanning the entire insertion. To determine which parental chromosome the event occurred on, we phased the proband, father, and mother's ONT reads and searched for diagnostic single-nucleotide variants that differed between the maternal and paternal

haplotypes. The proband is hemizygous for two maternal variants mapping within the deleted region while the father is homozygous for the alternative allele. Conversely, the proband harbored one paternal variant on the haplotype with the insertion that is absent in his mother. This demonstrated that the rearrangement occurred on the paternal chromosome. Analysis of the junctions showed that, besides the aforementioned deletion, no further rearrangements, such as a target site duplication, occurred at the boundaries. At the proximal junction, a short sequence stretch of four nucleotides (CAAA) was identified that could not uniquely be assigned to the chromosome 15 or the satellite DNA. However, due to its small size, it is unlikely that this stretch of homologous sequence had a role in the rearrangement mechanism, particularly in the determination of the target site.

We analyzed the content of interspersed repeats in 5 kb segments upstream and downstream of the rearrangement breakpoints as well as in the deleted segment on chromosome 15 sequence. These segments were enriched for LTR (long terminal repeats derived from endogenous retroviruses) content when compared to the human genome average, as assessed by simulation for the entire 13 kb segment (4.34-fold, $P = 0.035$, **Table 1**).

Table 1. Content in interspersed repeat elements of the rearranged target site on chromosome 15. The "*E*" value is the enrichment coefficient that was calculated by dividing the observed value by the mean of 10,000 genome-wide permutations (human genome average).

	Sequence upstream of the insertion (5 kb)	Deletion (2851 bp)	Sequence downstream of the insertion (5 kb)	Entire region	Human genome average	*E*, *P* ± SE
SINEs	9%	0%	12%	8%	12%	0.65, 0.57 ± 0.005
LINEs	0%	0%	0%	0%	19%	0, 1
LTR elements	62%	32%	13%	36%	8%	4.34, 0.035 ± 0.002
DNA elements	0%	0%	9%	4%	3%	1.21, 0.3 ± 0.005

Structural characterization of the alpha satellite DNA insertion

While we were unable to assemble the full sequence of the insertion, we investigated its structural properties by identifying reads with the longest content in alpha satellite DNA and unequivocally derived from this site, i.e. chimeric reads anchored to chromosome 15 sequence on either side of the insertion, spanning one breakpoint, and containing chromosome 18 centromeric alpha satellite sequences.

We selected two HiFi reads with estimated >99.9% accuracy and 7,199 bp (PacBio HiFi read 1) and 9,821 bp (PacBio HiFi read 2) of satellite DNA, both spanning the proximal junction; an ONT read with 8,618 bp of satellite DNA at the proximal junction (ONT read 1); an ONT read with 4,583 bp of satellite DNA at the distal junction (ONT read 2) (**Figure 1C**). Best alignments to the human genome reference (GRCh38) of alpha satellite segments from these four sequences showed identity with centromere reference models of chromosome 18 (Miga et al. 2014; Rosenbloom et al. 2015). Alignments to the CHM13-T2T (Telomere-to-Telomere) genome (Logsdon et al. 2020; Miga et al. 2020) resulted in unique locations for each read and pointed the origin of the transposition to a precise 10 kbp region in the centromere of chromosome 18 (chr18:17500487-17510699) (**Figure 1D**). While HiFi reads showed high identity (99%) with this region, ONT reads showed lower values (94%), mainly due to errors in their sequence. As the estimated size of the transposed segment (order of hundreds kbp) is bigger than the size of the corresponding interval within chromosome 18 centromeric sequence, we hypothesize that this region is likely expanded in the proband or alternatively in his paternal lineage and, therefore, structurally different from the CHM13 centromere. Overall, these results confirmed that the insertion originated from chromosome 18 centromeric DNA.

As chromosome 18 centromere is composed of two alpha satellite families, family I (D18Z1) and family II (D18Z2), both belonging to the suprachromosomal family 2 (SF2), whose arrays have a dimeric structure based on D1 and D2 monomers (Alexandrov et al. 1991), we assessed the similarity with deposited sequences representing both families. Local pairwise alignments showed 98% and 81% identity of both PacBio HiFi reads, respectively with D18Z1 (M65181.1) and D18Z2 (M38466.1) sequences, and 89.9% and 77.1% identity for the ONT read 2 transitioning over the distal breakpoint. These results indicate a closer relationship of the inserted satellite DNA to the D18Z1 family.

We analyzed the repetitive structure of the PacBio HiFi read 2, as it contains the longest satellite array sequence. We used the re-DOT-able tool (https://www.bioinformatics.babraham.ac.uk/projects/redotable/) and observed a higher density of matches every ~2000 bp (**Figure 1C, *bottom panel***). To further assess this periodicity, we extracted 57 monomers (size range 165-174 bp), built a multiple sequence alignment, and visualized all pairwise identity percentages by creating two heatmaps. The first one shows monomers ordered according to their position in the array, while the second heatmap depicts monomers ordered according to the dendrogram determined by the hierarchical clustering of identity percentages (**Figure 2A**). In the dendrogram-based heatmap, the

monomers cluster into two main groups as expected from the dimeric structure of the D18Z1 array (**Figure 2A**). D1-D1/D2-D2 sequence identity range from 78 to 100% (median 85%); D1-D2 identity range from 63 to 78% (median 70%) (**Figure 2B**). We then grouped every 12 monomers into ~2 kb units and obtained four repeats with 99.21-99.85% pairwise sequence identity (**Figure 2C**). These results are consistent with a 12-mer HOR structure, matching the known organization of the D18Z1 satellite array (McNulty and Sullivan 2018).

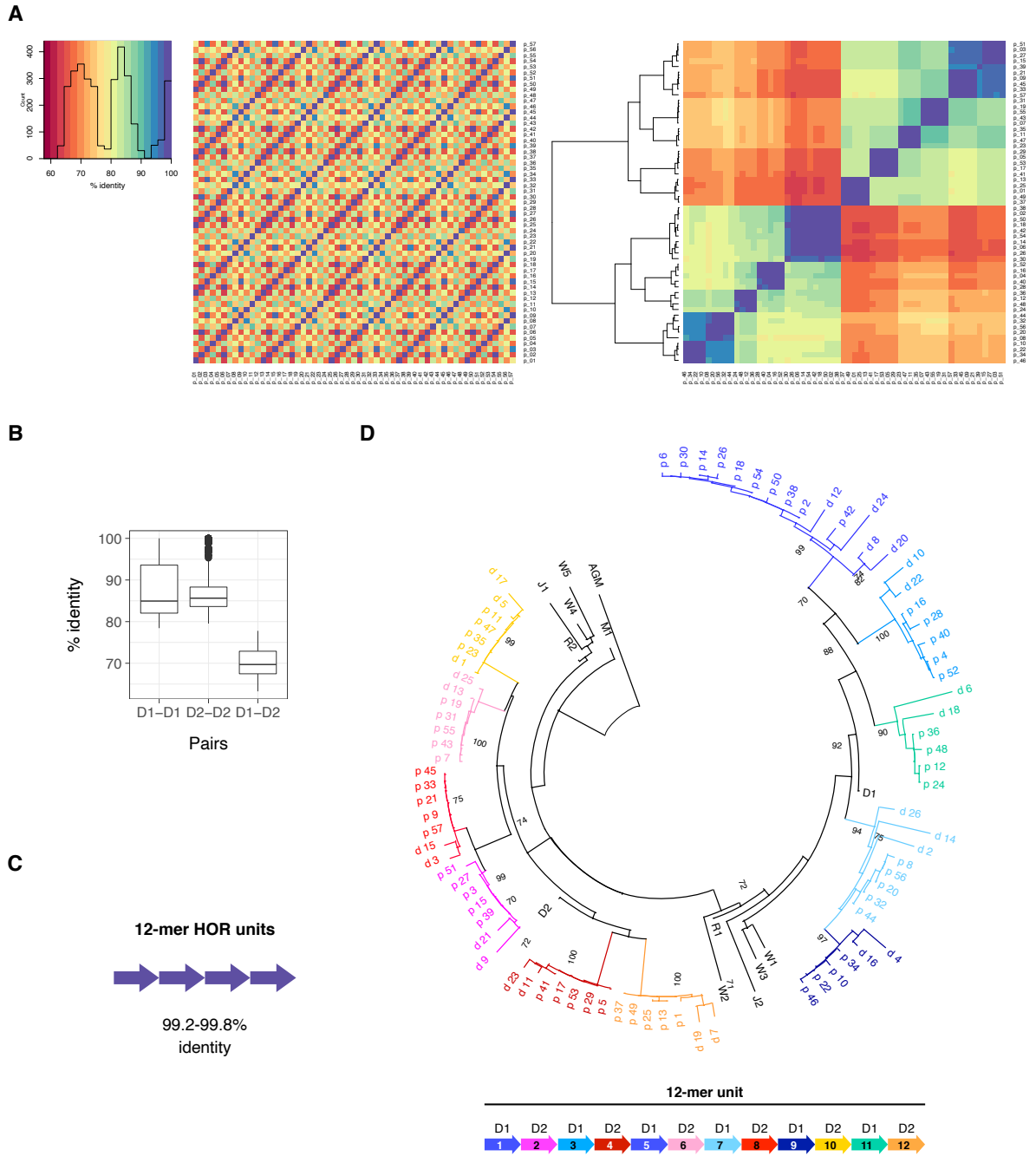

Figure 2. Organization of the alpha satellite array. A) Heatmaps of identity percentages between the 57 alpha satellite monomers of ~171 bp derived from the PacBio HiFi read 2, with monomers ordered either according to their position in the array (*left*) or as determined by clustering (*right*). **B)** Boxplots of identity percentages between D1-D1, D2-D2, and D1-D2 monomer pairs. **C)** Identity percentages between 12-mer HOR units. **D)** (*top*) Neighbor-joining tree of alphoid monomers from the PacBio HiFi read 2 transitioning over the proximal ('p') junction and from the ONT read 2 transitioning over the distal ('d') junction with sequences of the 12 human monomer types (D1, D2, J1, J2, M1, R1, R2, W1-5) and the alpha satellite from the African Green Monkey (AGM) as outgroup. Monomers are numbered according to their position in the arrays. Bootstrap values >70 are shown. (*bottom*) Schematic of the 12-mer HOR unit.

Finally, we extracted an additional 26 monomers (size range 158-177 bp) from the ONT read 2 spanning the distal breakpoint (in spite of the inherent sequencing uncertainties) and multialigned all monomers with sequences of the 12 different monomer types (D1, D2, J1, J2, M1, R1, R2, W1-5) found at all human centromeres and the alpha satellite sequence from the African Green Monkey as an outgroup. Despite differences in the accuracy of HiFi and ONT sequences, all monomers identified in the insertion clustered in two major clades formed by D1 and D2 monomers, confirming the assignment to these two monomer types. D1 and D2 monomers further grouped into 11 clades in agreement with their organization in a HOR unit of 12 monomers, with D1 monomers at positions 1 and 5 that were homogenized and formed a single clade (**Figure 2D**).

Functional profiling of the rearranged site

To assess whether this structural change is likely to have functional impact, we examined gene annotation (GENCODE v32) at the insertion breakpoints as well as in the deleted region. We find that the rearrangement did not directly disrupt any gene, with the closest one (*ST8SIA2*) annotated 32 kb distally (**Figure 3A**). We then evaluated whether the rearrangement affected other functional elements, such as regulatory DNA. To this end, we leveraged publicly available data from the ENCODE consortium of chromatin activity measured by chromatin immunoprecipitation sequencing (ChIP-seq) for three histone modifications, i.e. methylated histone 3 at lysine 4 (H3K4me1), tri-methylated histone 3 at lysine 4 (H3K4me3), and acetylated histone 3 at lysine 27 (H3K27ac), on seven cell lines. These epigenetic marks are associated with poised enhancers (H3K4me1), promoters (H3K4me3), and active enhancers (H3K27ac). Neither the deleted segment nor the breakpoints overlapped any of these chromatin features, suggesting that the rearrangement did not disrupt a regulatory element (**Figure 3A**).

Figure 3. Functional profiling of the rearrangement site. A) UCSC view of the 100 kbp region surrounding the rearrangement at 15q26.1. The deleted region is highlighted in yellow, with deletion extremes corresponding to the satellite insertion positions. The GENCODE v32 and ENCODE regulations (H3K4me1, H3K4me3, and H3K27ac) tracks are shown (hg38). No gene and no enrichment of epigenetic marks found near regulatory elements are annotated in the deleted region. The closest gene, *ST8SIA2*, is mapped 32 kb distally. **B) Methylation pattern of the insertion site in the family trio**. Methylation data obtained from the ONT selective sequencing. Methylated (red) and unmethylated (blue) CpGs are shown. The methylation profiles are similar among the family trio.

Next, we assessed whether the insertion of centromeric satellite DNA, which comes from a heterochromatic locus, modifies the epigenetic status of the 15q26 target region. We leveraged CpG methylation data of the 20 kb genomic segment surrounding the insertion site using the ONT data of the proband and his parents. Cytosine methylation is an epigenetic modification often found in CpG dinucleotides that contributes to the formation of heterochromatic regions

and leads to transcriptional modulation, in particular silencing. Comparison of the proband mutated allele with unrearranged ones, i.e. his maternal allele and the four alleles of his parents, revealed no major difference in the methylation patterns, indicating that the satellite insertion did not alter the methylation status of the surrounding region (**Figure 3B**). The absence of functional elements (gene or likely regulatory element) at the site and the maintenance of the methylation profile of the broader region suggest that the rearrangement itself has had no functional consequences. This is in line with the absence of clinical features in the proband that could not be explained by his trisomy 21 mosaicism.

Immuno-FISH with anti CENP-A and CENP-B antibodies

Cytogenetic evaluation of the derived chromosome 15 revealed no chromosomal constriction at the position where the satellite DNA sequence was inserted, suggesting that this site did not acquire properties of a functional centromere. To further demonstrate this lack of epigenetically-defined centromeric function, we performed an immuno-FISH experiment with an antibody against the CENP-A protein. We observe no colocalization of the D18Z1 probe and CENP-A staining at the satellite insertion locus on the derivative chromosome 15 (**Figure 4**). We also assessed by immuno-FISH the binding of the CENP-B box by the CENP-B protein. In 20 out of 25 mitoses, we observe a faint pattern of staining of the CENP-B antibody corresponding to the satellite insertion, whereas in the remaining five we observed no signal (**Figure 4**). Such faint signals may derive either from the smaller size of the satellite insertion compared to a centromeric satellite array or to a weaker binding of the CENP-B protein. Nevertheless, these results suggest that CENP-B proteins recognize and bind the CENP-B box on the satellite monomers of the inserted sequence. Although CENP-B is not necessary and sufficient to confer centromeric function, it was shown that it creates epigenetic chromatin states permissive for CENP-A or heterochromatin assembly (Otake et al. 2020).

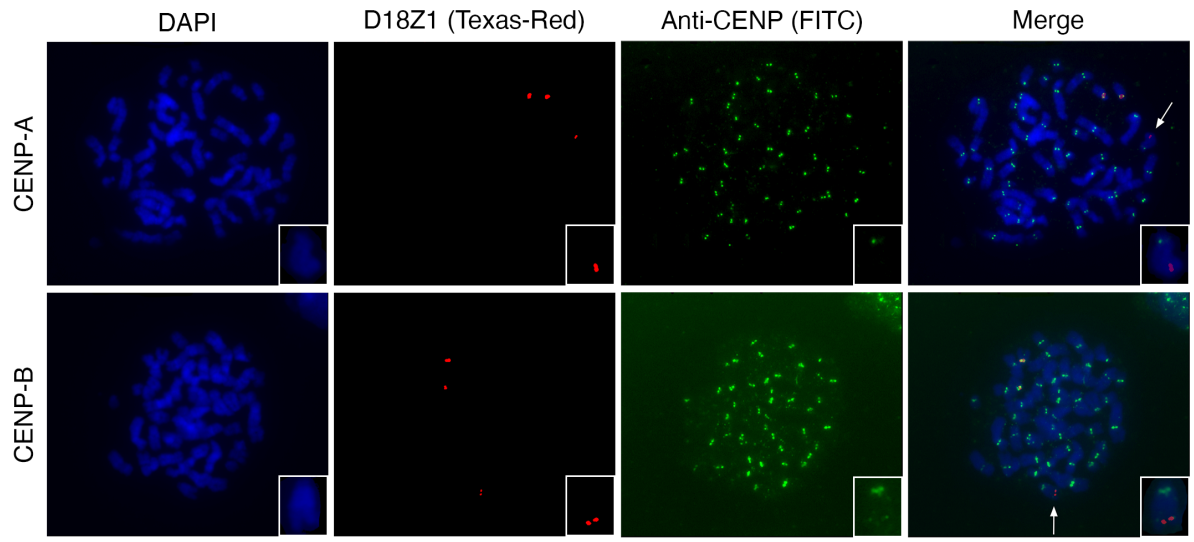

Figure 4. **CENP-A and CENP-B immuno-FISH.** Co-hybridization of the D18Z1 probe (red) with antibodies against CENP-A (*top*) and CENP-B (*bottom*) proteins (green) on chromosome metaphases from the proband. The arrows point at the derivative chromosome 15 that is also shown in larger magnification in the insets.

Discussion

During routine prenatal testing for aneuploidy by FISH, we serendipitously identified an individual carrying a *de novo* insertion of alpha satellite DNA from the centromere of chromosome 18 into cytoband 15q26 (**Figure 5**). Long-read sequencing and alignment to the CHM13-T2T genome showed that this segment transposed from a precise location of chromosome 18 centromeric HOR arrays. It also offered insights on the long-range organization of the alphoid sequence, such as homogenization of D1 monomers at positions 1 and 5.

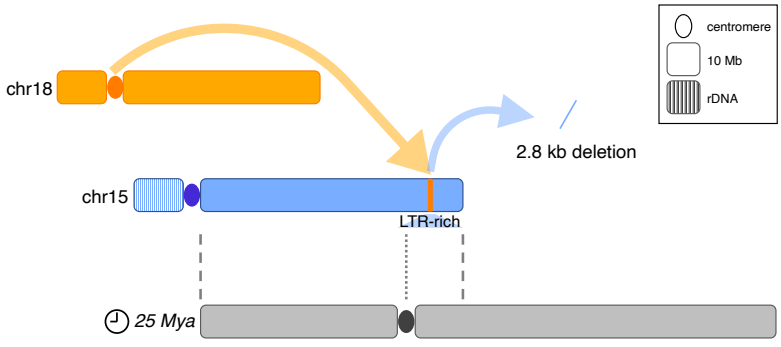

Figure 5. Schematic overview of the rearrangement. An alphoid array from the centromere of chromosome 18 inserted into an LTR-rich region of chromosome 15q26, ~10 Mbp distally from the site where an ancestral

centromere was seeded ~25 Mya. This transposition deleted 2.8 kbp of sequence. Dashed lines pinpoint the boundaries of the synteny between chromosome 15 and the ancestral submetacentric chromosome; the dotted line indicates the position of the ancestral centromere.

The lack of identification of such transposition/duplication events until now could be linked either to the fact that they are extremely rare and/or because current sequencing-based methodologies and analytical approaches aimed at genotyping structural variants are opaque to such events due to their size and highly repetitive nature. Indeed, our standard whole-genome sequencing diagnostic pipeline failed to identify this variant. Novel localizations of alphoid DNA were reported in the white-cheeked gibbon, a lesser ape with an extensively rearranged karyotype when compared to the ancestral primate karyotype. In this species, alpha satellite DNA is found not only at centromeres but also at telomeres and interstitial positions corresponding to some evolutionary breakpoints (Cellamare et al. 2009).

Our report brings to light a new class of structural variation that we call 'alpha satellite DNA insertion' (ASI) and raises questions about the frequency, structure, mechanism, and functional consequences of these events. Besides our observation, at least three additional lines of evidence suggest that duplication/transposition of alphoid DNA to a new genomic location occurs. First, several prenatal diagnostic reports describe the cross-hybridization of chromosome-specific centromeric alpha satellite probes to heterochromatic (centromeric or pericentromeric) regions of non-targeted chromosomes, i.e. the centromeres of chromosomes 19 and 22, the heterochromatin of chromosomes 1 and 9, and the pericentromeric region of chromosome 2 (Thangavelu et al. 1998; Winsor et al. 1999; Wei et al. 2007; Musilova et al. 2008; Collin et al. 2009). Of note, only centromeric probes of chromosomes 18, X, and Y are routinely used to screen for aneuploidies prenatally, therefore the ASI of other centromeric satellites would not be found in such a serendipitous manner. Moreover, ASI smaller than the standard resolution of FISH (~10 kb) would not be detected. The second line of evidence is the presence of small satellite DNA blocks that are not located at centromeric or pericentromeric regions in the human genome reference (Rudd and Willard 2004). While the presence of some of these can be explained by the evolutionary history of the locus, like the past presence of a centromere, the existence of some others may in fact be a result of fixed satellite insertion events. Thirdly, the maturation process of new centromeres, i.e. epigenetic specification before acquisition of the typical alpha satellite array (Kalitsis and Choo 2012; Nergadze et al. 2018), *per se* implies the movement of satellite DNA to other loci.

Our structural characterization of the rearrangement provides some insights on the mechanism of alphoid DNA spreading to non-centromeric locations. The coordinated deletion is reminiscent of the mechanism inferred for duplicative transposition and suggests the involvement of double-strand breakage of DNA (Cantsilieris et al. 2020), while the absence of active mobile elements adjacent to the insertion or target site duplications argues against retrotransposition (Deininger et al. 2003). It may be noteworthy that we also identified an enrichment of LTR elements in the long-range acceptor site. LTR retrotransposon activity is currently very limited or fully absent in humans (IHGSC 2001) and therefore is unlikely to have directly contributed to this form of structural variation. Such repeat-rich regions have been noted to be deleted as part of the duplicative transposition events associated with the new insertion of large (>100 kbp) blocks of segmental duplication (Johnson et al. 2006; Cantsilieris et al. 2020). Similarly, such coordinated deletions often (but not always) occur in gene-poor regions of the genome minimizing functional impacts of such massive new insertions and the fitness of the zygote/fetus.

Finally, the ASI location at 15q26 is interesting as a centromere resided at chromosome 15q25 in our past, ~10 Mbp away from the insertion site, and became inactive sometime between 20 and 25 Mya in the common ancestor of the ape lineage (Ventura et al. 2003; Giannuzzi et al. 2013). This observation raises the intriguing possibility that the alphoid array did not move to a random repeat-rich location in the genome, but instead revisited an evolutionary favored location mapping close to an ancestral centromere. Being the first observation of such events, we cannot discern between these two possibilities. However, if the latter scenario is correct, it might suggest that i) alphoid DNA preferentially moves to other extant or past centromeric locations; ii) there are genomic loci more suitable to host the centromeric function and associated alpha satellite array; iii) an alternative and opposite route to centromere repositioning and new centromere formation might in fact exist, where the region first acquires the satellite array and then the epigenetically-defined centromeric function emerges. Support for the latter comes from the observation that introduction of alpha satellite arrays in human cells can result in the formation of functional neocentromeres (Harrington et al. 1997; Ebersole et al. 2000). Some observations already demonstrate that certain regions of the genome have a memory and/or propensity to host centromeric function. For example, analphoid clinical new centromeres are often seeded at regions corresponding to ancestral centromeres, including 15q24-26, the target locus of our proband (Ventura et al. 2003; Capozzi et al. 2009), or in regions that are orthologous to positions that correspond to evolutionary new centromeres in

other primate lineages (Ventura et al. 2004; Cardone et al. 2006; Capozzi et al. 2008). Notably, the seeding position frequently maps within a variable distance (~1-14 Mb) from the region that hosts the centromeric function, as observed for the satellite insertion reported here. This suggests that centromeric function and satellite array evolution may be restricted to region rather than precise chromosomal location.

Besides suggesting a new class of structural variation and expanding current models of centromere life cycle, this case further highlights the risk of identifying false-positive aneuploidies of chromosomes 18, X, and Y when depending solely on centromeric satellite probes in rapid interphase FISH. Thus, it is critically important to follow up and confirm them by karyotyping. Lastly, this variant could be considered as a special case (as it occurs in the euchromatin) of chromosome heteromorphism, i.e. the variation in repetitive DNA content at heterochromatic regions. Although chromosome heteromorphisms are found in 2-5% of individuals and are generally considered as neutral genomic variations (Tempest and Simpson 2017), they are associated with infertility (Sahin et al. 2008). In this regard, it is possible that the aberrant presence of the alphoid array on the long arm of chromosome 15 might affect the accuracy of chromosome segregation during cell division and be causative of the recurrent trisomy 21 in the family. Future studies will clarify the prevalence of ASIs and their potential impact on chromosome aneuploidies and infertility.

Methods

Short-read sequencing and data analysis

We extracted genomic DNA from cultured amniocytes of the proband using QIAamp DNA mini kit (Qiagen, Hilden, Germany). We performed 150 bp paired-end WGS using the short-read Illumina platform. We aligned the reads to the hg38 version of the human genome using BWA-MEM version 0.7.10 (Li and Durbin 2009), run the BreakDancer version 1.4.5 (Chen et al. 2009) and ERDS version 1.1 (Zhu et al. 2012), and visually inspected the 15q24-26 region using the IGV tool. As we identified no structural variant, we re-aligned the reads to a custom library made of chromosome 15 sequence (hg38) and a deposited sequence of alpha satellite family 1 of chromosome 18 (M65181.1) (Alexandrov et al. 1991) using BWA version 0.7.17. To identify read pairs mapping at the insertion breakpoints, we selected discordant pairs with one end mapping on chromosome 15 and the other one on the satellite sequence and MAPQ>0. We removed soft and hard clipped reads and those mapping at the pericentromeric region of

418 chromosome 15. We next identified chimeric reads spanning the breakpoints among the soft
419 clipped reads using the Integrative Genomics Viewer (IGV) tool (Robinson et al. 2011).

420

421 *Long-read sequencing and data analysis*

422 We isolated PBMC (peripheral blood mononuclear cells) from the blood of the proband and
423 both parents. We extracted DNA from approximately 1-2 million cells of actively growing
424 culture by first pelleting the cells and resuspending them in 1.0 mL Cell Lysis Solution
425 (Qiagen). The samples were incubated with RNase A solution at 37°C for 40 min. Protein
426 Precipitation Solution (Qiagen) was added at 0.33x and mixed well. After a 10 min incubation
427 on ice, the precipitate was pelleted (3 min, 15000 rpm, 4°C). The supernatant was transferred
428 to new tubes, and DNA was precipitated with an equal volume of isopropanol. The DNA was
429 pelleted (2 min, 15000 rpm, 4°C) and the pellet was washed three times with 70% EtOH. The
430 clean DNA was rehydrated with DNA Hydration Solution (Qiagen) and left for two days to
431 resuspend.

432 We generated a PacBio HiFi library from the proband's genomic DNA using g-TUBE shearing
433 (Covaris) and the Express Library Prep Kit v2 (PacBio), size selecting on the SageELF
434 platform (Sage Science) to give a tight fraction of around 23 kbp by FEMTO Pulse analysis
435 (Agilent). The library was sequenced on one SMRT Cell 8M using v2 chemistry, and we
436 obtained 20.5 Gbp of HiFi reads with mean length of 20.9 kbp and median quality of Q27. We
437 assembled the data with HiCanu (Nurk et al. 2020) and Hifiasm (Cheng et al. 2021) and aligned
438 reads to the GRCh38 (hg38) reference genome using pbmm2
439 (https://github.com/PacificBiosciences/pbmm2).

440 Adaptive sampling was performed on an ONT GridION (one flow cell per sample) using
441 readfish (Payne et al. 2020). For each sample 1.5 ug of DNA was used to prepare a LSK-109
442 library according to the manufactures protocol. DNA was sheared in a Covaris g-TUBE at 6 k
443 rpm for 2 min. The region targeted was chr15:92,309,068-92,411,920 (hg38 coordinates). ONT
444 FAST5 files were basecalled using guppy 4.0.11 using the high-accuracy model. FASTQ files
445 were pooled and aligned to hg38.no_alt.fa using both minimap2 (Li 2018) and ngmlr
446 (Sedlazeck et al. 2018). We identified reads spanning the breakpoints (located at
447 chr15:92,359,068 and chr15:92,361,920) by manual inspection of the 15q26 read alignments
448 in IGV v2.4.16 (Robinson et al. 2011). We called and phased variants using Longshot (Edge
449 and Bansal 2019) and called CpG methylation using Nanopolish (Simpson et al. 2017).

Selected PacBio and ONT reads were aligned to the CHM13-T2T genome using pbmm2 and minimap2, respectively.

Analysis of repeat element content

We assessed the content in repeat elements in the deleted segment and in the 5 kb segments upstream and downstream the insertion breakpoints by using the annotation of the GRCh38 RepeatMasker track (Smit et al. 2013-2015). The null distributions were generated by performing 10,000 permutations of the entire 12,851 bp segment, excluding gaps and centromeres, by using BEDTools version v2.30.0 (Quinlan and Hall 2010). R v4.0.3 (R Core Team 2017) was used to compute empirical *P* values. Standard error (SE) was estimated using the formula SE=sqrt(*P**(1-*P*)/10,000).

Satellite monomer and HOR analysis

We created a dot plot of the PacBio HiFi read 2 using the re-DOT-able tool (https://www.bioinformatics.babraham.ac.uk/projects/redotable/). We extracted satellite monomers from the reads containing the longest satellite sequences and anchored to the proximal (PacBio HiFi read 2) or distal (ONT read 2) breakpoints by blast alignment (Altschul et al. 1990) with D1 monomer sequence (AJ130751.1). We performed multiple sequence alignments of monomers using Muscle (Edgar 2004) with default options. We created heatmaps and plots using the gplots v3.1.0 (https://CRAN.R-project.org/package=gplots) and ggplot2 v.2.2.1 (Wickham 2009) packages in the R software environment (R Core Team 2017). We used the neighbor-joining method (Saitou and Nei 1987) and the Kimura 2-parameter model distance (Kimura 1980), implemented in the MEGA X software (Kumar et al. 2018; Stecher et al. 2020), to examine phylogenetic relationships among the satellite monomers. We also included the sequences of the 12 monomer types and the alpha satellite monomer from African Green Monkey as an outgroup. All ambiguous positions were removed for each sequence pair (pairwise deletion option).

FISH and Immuno-FISH

FISH on uncultured amniocytes was performed with the Aquarius FAST FISH Prenatal kit (Cytocell, Cambridge, UK) (DXZ1, DYZ3, D18Z1, *RB1*, *DYRK1A* probes) according to manufacturer's instructions. Metaphase spreads were prepared from amniotic fluid cells and

lymphocytes according to standard procedures. FISH was further performed using BAC probes localized in 15q25.2, RP11-752G15 (FITC) (chr15:82,627,211-82,802,988, hg38) and RP11-635O8 (chr15:82,023,617-82,178,139) (TRITC) (RainbowFish, Empire Genomics, Buffalo, New York, USA) and alpha-satellites probes for chromosomes 15 (D15Z1, Texas-Red), 18 (D18Z1, Aqua) and 13/21 (D13/21Z1, Green) (Cytocell).

Immuno-FISH was performed on lymphoblastoid cells from the patient. Metaphase cells spreads were prepared according to a protocol adapted from Jeppesen (Jeppesen 2000). Briefly, lymphoblastoid cells were harvested after 44 hour culture, incubated at 37°C with colchicine (0.2µg/mL final concentration) during 2 hours, then in a 75 mM KCl hypotonic solution during 25 min. After centrifugation, cell pellet was resuspended in 75mM KCl/0.1% Tween20 and then cytocentrifuged 5 min at 1000 rpm. The slides were transferred to a Coplin jar containing KCMc solution (120 mM KCl, 20 mM NaCl, 10 mM Tris-HCl, pH 7.5, 0.5 mM EDTA, 0.1% (v/v) Triton X-100) and incubated 15 min at room temperature. Then immuno-FISH was performed with a protocol derived from Solovei et al. (Solovei et al. 2002) using as primary antibodies mouse anti-CENP-A (Abcam, Ab13939) (1/200) and mouse anti-CENP-B (5E6C1 clone, generous gift from Hiroshi Masumoto, Japan) (1/200); AlexaFluor conjugated goat anti-mouse as secondary antibody (1/1000) and D18Z1 probe (Texas-Red) (Cytocell). Images were performed with a Zeiss AxioImager Z2 fluorescence microscope equipped with a CoolCube Camera.

Acknowledgements

GG is recipient of a Pro-Women Scholarship from the Faculty of Biology and Medicine, University of Lausanne. This work was supported by the Swiss National Science Foundation grant 31003A_182632 and the Jérôme Lejeune Foundation to AR, by the National Institutes of Health grant HG010169 to EEE, and a grant from the Brotman Baty Institute for Precision Medicine to DEM and EEE. We wish to thank Emilie Chopin (Cell Biotechnology Center, Hospices Civils de Lyon) for providing lymphoblastoid cell lines, as well as Patrick Lomonte for his kind gift of antibodies and his advice.

References

Alexandrov IA, Mashkova TD, Akopian TA, Medvedev LI, Kisselev LL, Mitkevich SP, Yurov YB. 1991. Chromosome-specific alpha satellites: two distinct families on human chromosome 18. *Genomics* **11**: 15-23.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403-410.

Amor DJ, Choo KH. 2002. Neocentromeres: role in human disease, evolution, and centromere study. *Am J Hum Genet* **71**: 695-714.

Avarello R, Pedicini A, Caiulo A, Zuffardi O, Fraccaro M. 1992. Evidence for an ancestral alphoid domain on the long arm of human chromosome 2. *Hum Genet* **89**: 247-249.

Baldini A, Ried T, Shridhar V, Ogura K, D'Aiuto L, Rocchi M, Ward DC. 1993. An alphoid DNA sequence conserved in all human and great ape chromosomes: evidence for ancient centromeric sequences at human chromosomal regions 2q21 and 9q13. *Hum Genet* **90**: 577-583.

Cantsilieris S, Sunkin SM, Johnson ME, Anaclerio F, Huddleston J, Baker C, Dougherty ML, Underwood JG, Sulovari A, Hsieh P et al. 2020. An evolutionary driver of interspersed segmental duplications in primates. *Genome Biol* **21**: 202.

Capozzi O, Purgato S, D'Addabbo P, Archidiacono N, Battaglia P, Baroncini A, Capucci A, Stanyon R, Della Valle G, Rocchi M. 2009. Evolutionary descent of a human chromosome 6 neocentromere: a jump back to 17 million years ago. *Genome Res* **19**: 778-784.

Capozzi O, Purgato S, Verdun di Cantogno L, Grosso E, Ciccone R, Zuffardi O, Della Valle G, Rocchi M. 2008. Evolutionary and clinical neocentromeres: two faces of the same coin? *Chromosoma* **117**: 339-344.

Cardone MF, Alonso A, Pazienza M, Ventura M, Montemurro G, Carbone L, de Jong PJ, Stanyon R, D'Addabbo P, Archidiacono N et al. 2006. Independent centromere formation in a capricious, gene-free domain of chromosome 13q21 in Old World monkeys and pigs. *Genome Biol* **7**: R91.

Cellamare A, Catacchio CR, Alkan C, Giannuzzi G, Antonacci F, Cardone MF, Della Valle G, Malig M, Rocchi M, Eichler EE et al. 2009. New insights into centromere organization and evolution from the white-cheeked gibbon and marmoset. *Mol Biol Evol* **26**: 1889-1900.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP et al. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6**: 677-681.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**: 170-175.

Chiatante G, Giannuzzi G, Calabrese FM, Eichler EE, Ventura M. 2017. Centromere Destiny in Dicentric Chromosomes: New Insights from the Evolution of Human Chromosome 2 Ancestral Centromeric Region. *Mol Biol Evol* **34**: 1669-1681.

Collin A, Sladkevicius P, Soller M. 2009. False-positive prenatal diagnosis of trisomy 18 by interphase FISH: hybridization of chromosome 18 alpha-satellite probe (D18Z1) to chromosome 2. *Prenat Diagn* **29**: 1279-1281.

Deininger PL, Moran JV, Batzer MA, Kazazian HH, Jr. 2003. Mobile elements and mammalian genome evolution. *Curr Opin Genet Dev* **13**: 651-658.

Durfy SJ, Willard HF. 1989. Patterns of intra- and interarray sequence variation in alpha satellite from the human X chromosome: evidence for short-range homogenization of tandemly repeated DNA sequences. *Genomics* **5**: 810-821.

Earnshaw WC, Migeon BR. 1985. Three related centromere proteins are absent from the inactive centromere of a stable isodicentric chromosome. *Chromosoma* **92**: 290-296.

Ebersole TA, Ross A, Clark E, McGill N, Schindelhauer D, Cooke H, Grimes B. 2000. Mammalian artificial chromosome formation from circular alphoid input DNA does not require telomere repeats. *Hum Mol Genet* **9**: 1623-1631.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792-1797.

Edge P, Bansal V. 2019. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nat Commun* **10**: 4660.

565 Feliciello I, Pezer Z, Kordis D, Bruvo Madaric B, Ugarkovic D. 2020. Evolutionary History of Alpha
566 Satellite DNA Repeats Dispersed within Human Genome Euchromatin. *Genome Biol Evol* **12**:
567 2125-2138.
568 Giannuzzi G, Pazienza M, Huddleston J, Antonacci F, Malig M, Vives L, Eichler EE, Ventura M. 2013.
569 Hominoid fission of chromosome 14/15 and the role of segmental duplications. *Genome Res*
570 **23**: 1763-1773.
571 Harrington JJ, Van Bokkelen G, Mays RW, Gustashaw K, Willard HF. 1997. Formation of de novo
572 centromeres and construction of first-generation human artificial microchromosomes. *Nat*
573 *Genet* **15**: 345-355.
574 IHGSC. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.
575 IJdo JW, Baldini A, Ward DC, Reeders ST, Wells RA. 1991. Origin of human chromosome 2: an
576 ancestral telomere-telomere fusion. *Proc Natl Acad Sci U S A* **88**: 9051-9055.
577 Jeppesen P. 2000. Immunofluorescence in cytogenetic analysis: method and applications. *Genetics and*
578 *Molecular Biology* **23**: 1003-1014.
579 Johnson ME, National Institute of Health Intramural Sequencing Center Comparative Sequencing P,
580 Cheng Z, Morrison VA, Scherer S, Ventura M, Gibbs RA, Green ED, Eichler EE. 2006.
581 Recurrent duplication-driven transposition of DNA during hominoid evolution. *Proc Natl Acad*
582 *Sci U S A* **103**: 17626-17631.
583 Kalitsis P, Choo KH. 2012. The evolutionary life cycle of the resilient centromere. *Chromosoma* **121**:
584 327-340.
585 Karpen GH, Allshire RC. 1997. The case for epigenetic effects on centromere identity and function.
586 *Trends Genet* **13**: 489-496.
587 Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through
588 comparative studies of nucleotide sequences. *J Mol Evol* **16**: 111-120.
589 Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics
590 Analysis across Computing Platforms. *Mol Biol Evol* **35**: 1547-1549.
591 Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094-3100.
592 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.
593 *Bioinformatics* **25**: 1754-1760.
594 Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovykh MA, Koren S, Nurk S, Mercuri L, Dishuck PC,
595 Rhie A et al. 2020. The structure, function, and evolution of a complete human chromosome 8.
596 *bioRxiv* doi:10.1101/2020.09.08.285395: 2020.2009.2008.285395.
597 Loose M, Malla S, Stout M. 2016. Real-time selective sequencing using nanopore technology. *Nat*
598 *Methods* **13**: 751-754.
599 Marshall OJ, Chueh AC, Wong LH, Choo KH. 2008. Neocentromeres: new insights into centromere
600 structure, disease development, and karyotype evolution. *Am J Hum Genet* **82**: 261-282.
601 McKinley KL, Cheeseman IM. 2016. The molecular basis for centromere identity and function. *Nat*
602 *Rev Mol Cell Biol* **17**: 16-29.
603 McNulty SM, Sullivan BA. 2018. Alpha satellite DNA biology: finding function in the recesses of the
604 genome. *Chromosome Res* **26**: 115-138.
605 Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D,
606 Logsdon GA et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome.
607 *Nature* **585**: 79-84.
608 Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ. 2014. Centromere reference models
609 for human chromosomes X and Y satellite arrays. *Genome Res* **24**: 697-707.
610 Montefalcone G, Tempesta S, Rocchi M, Archidiacono N. 1999. Centromere repositioning. *Genome*
611 *Res* **9**: 1184-1188.
612 Musilova P, Rybar R, Oracova E, Vesela K, Rubes J. 2008. Hybridization of the 18 alpha-satellite probe
613 to chromosome 1 revealed in PGD. *Reprod Biomed Online* **17**: 695-698.
614 Nergadze SG, Piras FM, Gamba R, Corbo M, Cerutti F, McCarter JGW, Cappelletti E, Gozzo F,
615 Harman RM, Antczak DF et al. 2018. Birth, evolution, and transmission of satellite-free
616 mammalian centromeric domains. *Genome Res* **28**: 789-799.
617 Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy
618 AM, Koren S. 2020. HiCanu: accurate assembly of segmental duplications, satellites, and
619 allelic variants from high-fidelity long reads. *Genome Res* **30**: 1291-1305.

Otake K, Ohzeki JI, Shono N, Kugou K, Okazaki K, Nagase T, Yamakawa H, Kouprina N, Larionov V, Kimura H et al. 2020. CENP-B creates alternative epigenetic chromatin states permissive for CENP-A or heterochromatin assembly. *J Cell Sci* **133**.

Palmer DK, O'Day K, Trong HL, Charbonneau H, Margolis RL. 1991. Purification of the centromere-specific protein CENP-A and demonstration that it is a distinctive histone. *Proc Natl Acad Sci U S A* **88**: 3734-3738.

Panchenko T, Black BE. 2009. The epigenetic basis for centromere identity. *Prog Mol Subcell Biol* **48**: 1-32.

Payne A, Holmes N, Clarke T, Munro R, Debebe BJ, Loose M. 2020. Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nat Biotechnol* doi:10.1038/s41587-020-00746-x.

Piras FM, Nergadze SG, Magnani E, Bertoni L, Attolini C, Khoriauli L, Raimondi E, Giulotto E. 2010. Uncoupling of satellite DNA and centromeric function in the genus Equus. *PLoS Genet* **6**: e1000845.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

R Core Team. 2017. R: A language and environment for statistical computing., https://www.r-project.org/.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24-26.

Rocchi M, Stanyon R, Archidiacono N. 2009. Evolutionary new centromeres in primates. *Prog Mol Subcell Biol* **48**: 103-152.

Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M et al. 2015. The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res* **43**: D670-681.

Rudd MK, Willard HF. 2004. Analysis of the centromeric regions of the human genome assembly. *Trends Genet* **20**: 529-533.

Sahin FI, Yilmaz Z, Yuregir OO, Bulakbasi T, Ozer O, Zeyneloglu HB. 2008. Chromosome heteromorphisms: an impact on infertility. *J Assist Reprod Genet* **25**: 191-195.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**: 406-425.

Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF. 2001. Genomic and genetic definition of a functional human centromere. *Science* **294**: 109-115.

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. 2018. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* **15**: 461-468.

Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* **14**: 407-410.

Smit AFA, Hubley R, Green P. 2013-2015. RepeatMasker Open-4.0.

Solovei I, Cavallo A, Schermelleh L, Jaunin F, Scasselati C, Cmarko D, Cremer C, Fakan S, Cremer T. 2002. Spatial preservation of nuclear chromatin architecture during three-dimensional fluorescence in situ hybridization (3D-FISH). *Exp Cell Res* **276**: 10-23.

Stecher G, Tamura K, Kumar S. 2020. Molecular Evolutionary Genetics Analysis (MEGA) for macOS. *Mol Biol Evol* **37**: 1237-1239.

Tempest HG, Simpson JL. 2017. Why are we still talking about chromosomal heteromorphisms? *Reprod Biomed Online* **35**: 1-2.

Thangavelu M, Chen PX, Pergament E. 1998. Hybridization of chromosome 18 alpha-satellite DNA probe to chromosome 22. *Prenat Diagn* **18**: 922-925.

Ventura M, Mudge JM, Palumbo V, Burn S, Blennow E, Pierluigi M, Giorda R, Zuffardi O, Archidiacono N, Jackson MS et al. 2003. Neocentromeres in 15q24-26 map to duplicons which flanked an ancestral centromere in 15q25. *Genome Res* **13**: 2059-2068.

Ventura M, Weigl S, Carbone L, Cardone MF, Misceo D, Teti M, D'Addabbo P, Wandall A, Bjorck E, de Jong PJ et al. 2004. Recurrent sites for new centromere seeding. *Genome Res* **14**: 1696-1703.

673 Voullaire LE, Slater HR, Petrovic V, Choo KH. 1993. A functional marker centromere with no
674 detectable alpha-satellite, satellite III, or CENP-B protein: activation of a latent centromere?
675 *Am J Hum Genet* **52**: 1153-1163.
676 Wei S, Siu VM, Decker A, Quigg MH, Roberson J, Xu J, Adeyinka A. 2007. False-positive prenatal
677 diagnosis of trisomy 18 by interphase FISH: hybridization of chromosome 18 alpha-satellite
678 DNA probe (D18Z1) to the heterochromatic region of chromosome 9. *Prenat Diagn* **27**: 1064-
679 1066.
680 Wickham H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
681 Willard HF, Waye JS. 1987. Hierarchical order in chromosome-specific human alpha satellite DNA.
682 *Trends in Genetics* **3**: 192-198.
683 Winsor EJ, Dyack S, Wood-Burgess EM, Ryan G. 1999. Risk of false-positive prenatal diagnosis using
684 interphase FISH testing: hybridization of alpha-satellite X probe to chromosome 19. *Prenat*
685 *Diagn* **19**: 832-836.
686 Zhu M, Need AC, Han Y, Ge D, Maia JM, Zhu Q, Heinzen EL, Cirulli ET, Pelak K, He M et al. 2012.
687 Using ERDS to infer copy-number variants in high-coverage genomes. *Am J Hum Genet* **91**:
688 408-421.

689