OXFORD UNIVERSITY PRESS | Molecular Biology and Evolution

The Targeted Sequencing of Alpha Satellite DNA in Cercopithecus pogonias Provides New Insight into the Diversity and Dynamics of Centromeric Repeats in Old World monkeys

Journal:	*Molecular Biology and Evolution*
Manuscript ID	Draft
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Cacheux, Lauriane; Museum National d'Histoire Naturelle, Adaptations du vivant Ponger, Loic; Museum National d'Histoire Naturelle, Adaptations du vivant Gerbault-Seureau, Michèle; Muséum national d'Histoire naturelle, Institut de Systématique, Evolution, Biodiversité Loll, François; Museum National d'Histoire Naturelle, Adaptations du vivant Gey, Delphine; Museum National d'Histoire Naturelle, Origines et Evolution Richard, Florence; Museum National d'Histoire Naturelle, Origines et Evolution Escude, Christophe; Museum National d'Histoire Naturelle, Adaptations du vivant
Key Words:	Alpha satellite DNA, Centromere genomics, chromosomal evolution, Higher order repeats, acrocentric chromosomes, Cercopithecini

ARTICLE

DISCOVERIES

The Targeted Sequencing of Alpha Satellite DNA in *Cercopithecus pogonias* Provides New Insight into the Diversity and Dynamics of Centromeric Repeats in Old World monkeys

Lauriane Cacheux,[1,2] Loïc Ponger,[*,1] Michèle Gerbault-Seureau,[2] François Loll,[1] Delphine Gey,[3] Florence Anne Richard[2,4] and Christophe Escudé[*,1]

[1]Département Adaptations du Vivant, Structure et Instabilité des Génomes, INSERM U1154, CNRS UMR7196, Sorbonne Universités, Muséum national d'Histoire naturelle, Paris, France.

[2]Département Origines et Evolution, Institut de Systématique, Evolution, Biodiversité, UMR 7205 MNHN, CNRS, UPMC, EPHE, Sorbonne Universités, Muséum national d'Histoire naturelle, Paris, France.

[3]Service de Systématique moléculaire, UMS 2700 CNRS, Sorbonne Universités, Muséum national d'Histoire naturelle, Paris, France.

[4]Université Versailles St-Quentin, Montigny-le-Bretonneux, France.

* **Corresponding authors**: E-mails: loic.ponger@mnhn.fr and christophe.escude@mnhn.fr

Abstract

Alpha satellite is the major repeated DNA element of primate centromeres. Specific evolutionary mechanisms have led to a great diversity of sequence families with peculiar genomic organization and distribution, which have till now been studied mostly in great apes. Using high throughput sequencing of alpha satellite monomers obtained by enzymatic digestion followed by computational and cytogenetic analysis, we compare here the diversity and genomic distribution of alpha satellite DNA in two related Old World monkey species, *Cercopithecus pogonias* and *Cercopithecus solatus*, which are known to have diverged about seven million years ago. Two main families of monomers, called C1 and C2, are found in both species. A detailed analysis of our datasets revealed the existence of numerous subfamilies within the centromeric C1 family. Although the most abundant subfamily is conserved between both species, our FISH experiments clearly show that some subfamilies are specific for each species and that their distribution is restricted to a subset of chromosomes, thereby pointing to the existence of recurrent amplification/homogenization events. The pericentromeric C2 family is very abundant on the short arm of all acrocentric chromosomes in both species, suggesting specific mechanisms that lead to this distribution. The use of two restriction enzymes allowed us to show that, while the monomeric organization is the predominant structural motif, higher order organization patterns can be detected in the *Cercopithecus pogonias* genome. Our study suggests a high dynamics of alpha satellite DNA in Cercopithecini, with recurrent apparition of new sequence variants and frequent interchromosomal sequence transfer.

Key words

Alpha satellite DNA; Centromere genomics; chromosomal evolution; Higher order repeats; acrocentric chromosomes; Cercopithecini

Introduction

In eukaryotes, centromeric DNA is made of large tracts of tandemly repeated sequences, also called satellite DNA. Satellite DNAs can differ significantly between closely related species and their evolution is driven by molecular processes that differ from those that affect other parts of genomes. Changes in satellite DNA content and distribution can alter heterochromatin and centromere function and therefore can accompany speciation (Palomeque and Lorite 2008; Plohl et al. 2008).In Primates, the most abundant satellite DNA, called alpha satellite, is made of AT-rich monomers that are about 170 bp in length (Schueler and Sullivan 2006). Alpha satellite monomers represent a very large sequence family that has been classified into distinct subfamilies, which differ in their DNA content but also in their organization and genomic distribution (Waye and Willard 1986; Alexandrov et al. 1988; Vissel and Choo 1991; Shepelev et al. 2009; Hayden 2012; Catacchio et al. 2015). Alpha satellite DNA displays two types of organization in primate genomes: a so-called monomeric organization, where arrays of adjacent monomers belong to the same family, and a higher-order repeats (HOR) organization that involves highly conserved repeated motifs where each motif is made of several monomers, possibly belonging to different families (Schueler and Sullivan 2006).

Several studies have addressed the evolutionary dynamics of alpha satellite DNA, most of them focusing on comparing human DNA sequences with those of great apes (Schueler and Sullivan 2006; Cellamare et al. 2009; Shepelev et al. 2009; Catacchio et al. 2015; Chiatante et al. 2017). The observation of an age gradient when going from centromere towards chromosome arms has led to suggest that, on a single chromosome, alpha satellite families emerge and expand at the centromere core, thereby splitting and displacing older families distally onto each chromosome arm, where they are found in so-called pericentromeric regions (Schueler et al. 2005). In addition, a certain amount of evidence points to the existence of transfer of alpha satellite DNA between chromosomes. For example, some families are found preferentially on certain subsets of human chromosomes (Alexandrov et al. 1988). Finally, some families and/or HORs are conserved within great apes, but they usually span non-homologous centromeres (Jorgensen et al. 1987; Archidiacono et al. 1995; Warburton et al. 1996; Rudd et al. 2006; Catacchio et al. 2015). Numerous mechanisms have been called upon to underpin these observations at the molecular level, such as unequal crossing over or sister chromatid exchange, transposition, gene conversion, rolling circle replication and reinsertion, and transposon-mediated exchange (Schindelhauer and Schwarz 2002; Rudd et al. 2006; Palomeque and Lorite 2008; Plohl et al. 2008; Garrido-ramos 2017). Nevertheless, how concerted evolution leads to appearance and accumulation of species-specific sequence variations in short evolutionary periods and drives satellite DNA divergence remains largely unknown (Dover 1982; Pérez-Gutiérrez et al. 2012; Feliciello et al. 2014; Utsunomia et al. 2017).

In contrast to apes, information gathered on alpha satellite families is relatively limited in monkeys (Alkan et al. 2007). Cercopithecini represent a large clade of Old World monkeys containing 35 species that have diverged over the last 10 million years (Tosi 2008; Guschanski et al. 2013). The numerous chromosomal rearrangements that are observed in this clade can be associated to centromere repositioning or emergence of new centromeres (Dutrillaux et al. 1980; Moulin et al. 2008). The evolutionary history of alpha satellite DNA has never been studied in relation to such events. In a recent study, we have characterized alpha satellite DNA in *Cercopithecus solatus* (CS), using deep sequencing of enzymatically obtained monomers and dimers of alpha satellites, combined with computational and cytogenetic analyzes. Our results provided evidence for the existence of at least four alpha satellite families that differed from those previously described in the ape lineage (Cacheux et al. 2016).We present here investigations into the alpha satellite component of another species, *Cercopithecus pogonias* (CP), whose genome contains 72 chromosomes, as compared with 60 for CS (Dutrillaux et al. 1980; Moulin et al. 2008). The chosen experimental approach was similar to the previous one, except that monomers were obtained using two different restriction enzymes, XmnI and HindIII. A thorough investigation of our datasets allowed us to refine the identification of alpha satellite DNA families and to compare the diversity, structural organization and chromosomal distribution of alpha satellite DNA in both species, thereby providing unprecedented information regarding the dynamics of alpha satellite families during evolution.

Results

Identification of alpha satellite DNA families from the CP XmnI dataset

Alpha satellite monomers were isolated from the CP genome using the XmnI restriction enzyme, then sequenced and parsed as previously described for CS (Cacheux et al., 2016) (see Materials and Methods). The recovered 112,575 sequences were first analyzed with a principal component analysis (PCA) using the 5-mer nucleotide composition. Visualization of sequences into the plane formed by the two first components of the PCA revealed a pattern that differed slightly from the one obtained for CS. We distinguished a large group whose structure suggests it could contain several subgroups (left of fig. 1*A*) and a smaller well separated group (right of fig. 1*A*). We decided to combine the data from the two species into the same projection space (fig. 1*B*). The obtained graph shows that the group that appears on the right overlaps quite well with the C2 group of CS. On the other hand, the group that appears on the left occupies a larger space on the graph, extending both above and below the C1 group of CS and thereby suggesting that two additional groups of sequences may be present in CP. We hypothesized that the two previously identified families C1 and C2 coexist in CP with two new families that we decided to call C5 and C6, respectively (fig. 1*C*). After having assigned all sequences to each of the four families by using a combination of hierarchical clustering (HCA) and linear discriminant analysis (LDA), as previously described, we confirmed, using phylogenetic trees, the existence of the four families as well as the identities of the C1 and C2 families of CP and CS (supplementary fig. S1). The abundance of each family is reported on table S1 together with some of their properties, and their consensus are depicted on figure 2. The consensus sequences of the C1 and C2 families of CP are identical to those that were established for CS, at the exception of a single nucleotide at position 167 that was ambiguous in the C2 consensus of CS. The consensus sequences of both C5 and C6 differed only by three single nucleotide variations from that of C1 (fig. 2), although in the case of C5 the N at position 28 reflects the presence of abundant sequences containing either a G (as in C1) or a T within the dataset. The C5 and C6 families exhibit a high sequence homogeneity (95 % and 98 % mean sequence identity, respectively) which is in the same of order as that of C1 (95%) and much greater than that of C2 (85 %). All families contained a pJalpha box and no CENP-B box, as observed for CS (see table S1). Some sequences were observed to be present a high number of times in the dataset (see table 1). They will be described in more details further.

Chromosomal distribution of the alpha satellite families analyzed by FISH

We were next interested in designing oligonucleotide probes for studying the chromosomal distribution of the four families of alpha satellite DNA identified within the CP monomer dataset by FISH. Our previous results and a new in silico analysis suggested that the specific detection of the C5 and C6 families should be possible using the C5a and C6a probes, respectively (table S2 and fig. S2), because binding of these probes to other families would be prevented by the presence of at least two mismatched base pairs (Cacheux et al 2016). The same in silico analysis revealed that there is no oligonucleotide probe design that will allow for the specific detection of the C1 family. The C1a and C1b probes are expected to detect either both C1 and C5 (for C1a) or C1 and C6 (for C1b) in CP (tableS2). FISH experiments were then performed on metaphases prepared from cells that came from the same male specimen as the one used for the sequencing. The use of the C1a/C1b and C2a/C2b probe sets revealed hybridization patterns on CP chromosomes that resembled those observed on CS chromosomes. The probes targeting C1 produced intense signals covering the centromeres of all but two chromosomes (fig. 3*A* and 3*B*). On some chromosomes, the signal appeared very large, extending towards pericentromeric regions. The probes targeting C2 stained intensely the acrocentric chromosomes on their shorter arm and produced weaker signal in the pericentromeric regions of numerous non acrocentric chromosomes (fig. 3*A*,3*B* and S3). The absence of alpha satellite DNA on two chromosomes was confirmed using a probe designed to bind all alpha satellite sequences, and the use of a chromosome banding technique allowed us to identify these chromosomes as the Y and a single chromosome 6 (fig. S4).

We next investigated the labeling pattern generated using the C5a and C6a probes. Intense signals were observed on 11 chromosomes and 32 chromosomes for the C5a and C6a probes, respectively (fig. 3*C,D*). Only a single chromosome pair displayed both signals (fig. 3C, see arrows). The identity of these chromosomes was also established using cytogenetic experiments (fig. S5 and S6): probe C5a labeled 5 pairs of autosomes and the X chromosome, while C6a produced intense signals at the centromere of all 12 pairs of acrocentric chromosomes, slightly lighter signals at the centromere of one pair of submetacentrics, and 6 additional weaker signals which were shown to belong to four chromosome pairs, two of which displaying a heterozygote signal (fig. S6).Except for one chromosome pair, identified as chromosome 20, the C5a probe provided a signal that was located on both sides of the primary constriction, but absent from the central part, which is still labeled by the C1a probe. This suggests that the C5 family occupies a slightly pericentromeric localization (see arrows on fig. 3*D*). The C6a probe always provided a signal that was located at the centromere core. The hybridization patterns of C5a and C6a were clearly distinct from each other and from those of the C1 targeting probes, which validates our probe design strategy. As expected, the signals obtained using the C1a and C1b probes were found to overlap with those of C5a and C6a, respectively. Finally, FISH experiments were performed on CS chromosomes using the C5a and C6a probes. C6a did not provide any signal, as expected from the absence of the C6 family in the CS genome (not shown). A slight signal was observed using the C5a probe, that was removed by increasing the temperature, suggesting a light non specific hybridization (see fig. S7). Based on all these observations, we conclude that in addition to the C1 and C2 families previously described in CS, the genome of CP contains two additional alpha satellite families, named C5 and C6, that display specific chromosomal distribution patterns.

The HindIII dataset reveals additional families and organizational patterns

Digestion of CP DNA using the HindIII restriction enzyme resulted in a ladder pattern similar to the one obtained using the XmnI enzyme (not shown). We therefore decided to implement on HindIII monomers a similar experimental and analytical approach as the one described for XmnI. A total of 84,485 alpha satellite monomers were recovered. Sequences were visualized into the plane formed by the two first components of the PCA, alone and in combination with the sequences from the XmnI dataset (fig. 4*A,B*). The obtained graph provided evidence for four families and suggested that three of them are identical to those found in the XmnI dataset. We decided to name C1', C2' and C5' the three families that overlap with C1, C2 and C5 on the PCA graph (fig. 4*C*). Using once again a HCA/LDA approach, all sequences were sorted for their belonging to one of the four families, and consensus sequences were computed. The strict identity of the consensus sequences between C1 and C1', C2 and C2', and C5 and C5', except for a phase shift, suggests that these monomers are the digestion results of tandemly repeated sequences from the same family by the two different enzymes.

The fourth family of HindIII sequences appeared on the PCA graph as a group of points with a size similar to that of C6 and a slightly shifted position (fig. 4*B* and 4*C*); this family was called C6'. Comparison of the consensus sequences of C6 and C6' showed that they were identical in the overlapping 106 bp HindIII-XmnI fragment but that they differed by a substitution, C2A, and a deletion, G17Del, within the non-overlapping but homologous 66 bp XmnI- HindIII fragment (supplementary fig. S8). We designed an oligonucleotide probe targeting the C2A-G17Del variation that distinguishes the C6' consensus from the C6 consensus. The hybridization pattern of this probe in FISH experiments overlapped quite well with the signals provided by probe C6a (fig. 5). We only noticed additional very weak signals on a few additional chromosomes using the C2A-G17Del probe, which will be discussed further. These results provide strong support for the existence of a higher order repeat structure containing at least two monomers, where a monomer from the C6 family is followed by another monomer whose sequence is only partially known.

Highly repeated sequence variants provide insights into additional alpha satellite families

In both CS and CP, the monomer datasets contained several sequences that were repeated a high number of times. A detailed investigation of the 30 most abundant sequences from CP was performed for both enzymes. As observed in our previous study (Cacheux et al 2016), several sequences containing a deletion within a homopolymer tract were identified as sequencing errors based on strong biases in the orientation of the sequencing reads (table 1 and S3). These sequences, that are shown with a grey shadow in the tables, were not considered in the forthcoming analysis. The most abundant sequence for both enzymes was the exact sequence consensus of the C1 family. Other highly repeated sequences corresponded to sequences that differed from the latter one by a single nucleotide variation (such as in 2, 3, 13, 14, etc), two single nucleotide variations (such as in 12, 15, 16, etc), three single nucleotide variations (such as in 5, 10, 11, etc), and also to sequences combining one or two single nucleotide variations with one single nucleotide deletion (see sequences 7 and 8; all examples are taken from the XmnI dataset). Similar variation patterns were observed with both enzymes and in general, identical variations were found with similar frequency within both datasets. Interestingly, the absence in the XmnI dataset of a highly repeated sequence from the HindIII dataset (number 13) could be explained by nucleotide variations that abolish the cleavage site for XmnI. Although a slight bias for read orientation (see for example sequences 10, 13, 16 and 19 in table 1) was sometimes observed, probably due to sequence-dependent differential efficiency of the Ion torrent technology, we reasoned that all these sequences represented homogenous sets of identical sequences directly recovered from the CP genome.

We noticed that sequence 5 from the XmnI dataset (455 repeats) matched to the consensus of the C6 family, and that sequences 11 (116 repeats) and 30 (53 repeats) matched to the consensus of C5 with a G or a T at position 28, respectively. On the graph showing the two first principal components of the PCA (fig. 6*A*), the points corresponding to these sequences were located at the left end of elongated groups of points, displaying a "comet-like" structure. For these three highly repeated sequences, we decided to plot sequences from the XmnI dataset that were identical to these sequences except for one single nucleotide difference (fig. 6*B*). Interestingly, the distributions of these sequences overlaps quite well with the beginning of the tails of the comets, suggesting that comets are traces of mutation events that have affected sequences that were initially present in a high number of identical copies. This observation therefore establishes a link between highly repeated sequences, comet-like structures and potential families or sub-families of alpha satellite DNA. The highly repeated sequence variants can be viewed as the signature of faithful homogenization/amplification events affecting a single alpha satellite monomer, while the tails of the comets represent the divergence of sequences following mutation events. In this view, each different amplified sequence variant and the closely related sequences define an independent alpha satellite DNA family. On the graph shown in figure 6, whose transparency was chosen greater than the one shown on figure 1, additional comet-like structures can be distinguished. Repeating the process shown in figure 6*B* for additional highly repeated sequences, such as 3, 7 and 19 (fig. 6*C*), or 2, 8, 12 and 15 (fig. 6*D*), we showed that some of the observed comets seem to derive from identified highly repeated sequences (see for example the sequences shown in blue on fig. 6*C* and 6*D*, which correspond to sequences 19 and 12, respectively). As dispersed points corresponding to mutated sequences are observed next to all highly repeated sequence, it is likely that the large cloud of points that was attributed till now to a single C1 family may in fact represent numerous subfamilies, each one deriving from a previously amplified sequence. These subfamilies, which derive from sequences that differ from each other by only few nucleotides, generally overlap on the PCA graph. Interestingly, we were able to detect comet-like structures when we plotted the results of the PCA for the CS XmnI dataset in the CP axis system (fig. S9), which shows that the C1 family of CS has an internal distribution which may be more complex than previously anticipated (Cacheux et al 2016).

The comparison of highly repeated sequence variants found in both species reveals that, besides the most abundant sequence variant, i.e. sequence 1, only very few sequence variants are found in both species (for example sequences 14 and 20). Moreover, numerous relatively abundant sequence variants exist that are found only in one species. One striking observation is that many of the abundant sequence variants of CP contain the C158G single nucleotide variation (sequences 2, 8, 10, 12, 15, 28), while this variation was barely detected within the CS sequences (a sequence strictly identical to sequence 2 of CP was found repeated only 28 times). Using a probe set that was designed in order to distinguish sequences containing a C or a G at position 158 in FISH experiments, we showed that strong signals were observed on all CP chromosomes using the 158C-detecting probe, while the 158G-detecting probe stained only a subset of CP chromosomes, with strong signals observed at the centromere of all 12 pairs of acrocentric chromosomes while weaker signals were located at the core centromere of a few other chromosomes (fig. S10). In CS, FISH experiments had also shown that distribution of one of the highly abundant sequence was limited to 4 chromosome pairs (Cacheux et al 2016). These observations support the hypothesis by which amplification events lead to the local accumulation of new sequence variants, whose detailed analysis provides a new approach for the comparative genomics of alpha satellite DNA between species.

Alpha satellite DNA and karyotypic structure in Cercopithecini

The evolutionary history of chromosomes has been previously studied in Cercopithecini (Moulin et al. 2008). Alignment of chromosomes from different species and building of the karyotype of a presumed common ancestor allows one to track the finite number of events that are supposed to have occurred during evolution. Without getting into all the details of these processes, we decided here to investigate if the distribution of the so-called C5 and C6 families, that have emerged in the CP lineage, was obviously associated to the emergence of new centromeres or with centromere repositioning events that would have occurred in this specific lineage. The scheme shown on figure 7 represents the alignments of all CP and CS chromosomes, using the human genome as reference, where the position of FISH signals obtained in figures S3, S5 and S6 are reported. We observed that the C5 family was sometimes located on chromosomes that seem to have undergone rearrangement but whose centromere was not obviously repositioned (for example CPO8 or CPO20, see HSA11 and HSA2q). Interestingly it is also found on chromosome X whose structure and centromere position seem to be preserved. The C6 family was observed on all acrocentric chromosomes and on five pairs of non acrocentric chromosomes, three of which (CPO2, CPO11 and CPO12) do not show obvious changes in chromosome structure or centromere positioning (see HSA6, HSA13 and HSA5). Among the 12 pairs of acrocentric chromosomes from CP (named 24 to 35), three (CPO24, CPO25 and CPO28) have homologs that are also acrocentric in CS and therefore are not expected to have required the emergence of a new centromere (see HSA5, HSA7, HSA22). All these observations demonstrate that the sequences that have emerged in the CP lineage can be found in some centromeric regions that have not been obviously reorganized. Of course, the low resolution of our cytogenetic approach does not exclude that some rearrangements have occurred in these centromeric regions. On the other hand, the finding of the C6 sequences on all acrocentrics supports the existence of frequent exchange of genetic material between these chromosomes. Our approach also made possible to study the dynamics of sequences from the C2 family, that are found to be very abundant on the short arm of acrocentric chromosomes in both species. Only three out of the twelve pairs of acrocentrics in CP are also acrocentrics in CS. Sequences from the C2 family may have been present on the common ancestors of those chromosomes and then have invaded all acrocentrics on their short arm side. C2 signals are also present on non acrocentric chromosomes, albeit with a lower intensity. Although it is difficult to establish quantitative comparison of signal intensities between homologous chromosomes, observation of the karyotypes suggests that the distribution of the pericentromeric C2 signals differs between CP and CS. In particular, strong signals observed on CSO6, CSO9, and CSO10 (see Cacheux et al, 2016) are not found on the homologous CPO1, CPO4 and CPO17 (fig . S3).

Discussion

Identification of alpha satellite DNA families

In the present study, we have analyzed the content and genomic distribution of alpha satellite DNA in the CP genome, implementing an experimental strategy that was similar to the one we previously applied to another Cercopithecini species, CS (Cacheux et al. 2016). Our final aim was to compare the diversity and distribution of this important genome component in these two related species. In both species, analysis of XmnI monomers revealed a very abundant family, called C1, and a more diverged and less abundant family, called C2, which have a centromeric and pericentromeric localization, respectively, as shown by FISH experiments using oligonucleotide probes. The analysis of CP monomers obtained using a different restriction enzyme, HindIII, did not reveal any important alpha satellite family that would not be cleaved by XmnI. Moreover, it helped in the analysis of the structural organization of monomers (see further).

Comparing the sequence distribution of both species on PCA graphs led us to distinguish two additional alpha satellite families, which were shown by FISH experiments to be localized only on specific chromosomes from the CP genome. We also noticed that the axis system that emerged from the PCA analysis of the monomers from CP revealed comet-like structures in the graphical representation, and that highly-repeated sequences were found at the "head" of each comet, while the "tail" of the comets contained mutated versions of the highly repeated sequences. We propose here that this pattern reveals the evolutionary processes that underlie the evolution of alpha satellite DNA. At a certain time, a sequence variant is amplified through a recombination-based or rolling-circle mechanism, giving rise to multiple identical copies which are later modified by mutations, thereby forming a new family. Most of these families differ from each other by only a few nucleotides in their consensus sequence, making their identification through a PCA analysis a highly difficult task. The identification of highly repeated sequence variants provides an alternative approach for their identification. In this view, the so-called C5 and C6 families may represent at least three subfamilies of C1 instead of the initially proposed two independent families.

Structural organization of alpha satellite DNA in Cercopithecini

In our previous study (Cacheux et al. 2016), sequencing of CS XmnI monomers and dimers had led to the identification of two families of sequences with a monomeric organization (C1 and C2) and of two additional families, called C3 and C4, that were part of a higher organization. The distribution of these two last families was restricted to the Y chromosome and to the pericentromeres of a few other chromosomes. The low number of dimers that could be analyzed led us to abandon the sequencing of dimers in the present study and to choose instead to perform the global analysis of monomers using two restriction enzymes, which should help in studying the relative organization of monomers from each family relative to each other. The results obtained with both enzymes were consistent with each other and with the tandem organization of monomers from the C1, C2, and C5 families. The distinction, within the HindIII dataset, of an alpha satellite DNA family that shares an identical nucleotide composition with C6 over the HindIII-XmnI fragment but diverges over the XmnI-HindIII fragment, was interpreted as a hint towards the existence of a higher order repeat structure in which monomers from the C6 family are associated to monomers that are known to carry the C2A-G17Del variation but whose complete sequence is not available. The PCA analysis of the XmnI monomer dataset did not reveal an obvious family that would carry this variation. Nevertheless, observation of the highly repeated sequences revealed two potential candidates, which are the subfamilies defined by the consensus sequences 7 and 8, that both carry the observed variation. Interestingly, the repeat number of sequence 5 (455), which is at the origin of the C6 family, is very close to the sum of repeat numbers observed for sequences 7 and 8 (458). It is therefore tempting to speculate that at least two different higher order repeats, each containing at least two monomers, are present in the CP genome. The sequences of these HOR, which are located at or very close to centromeres, are very homogenous, as shown by the high sequence identity observed for the C6 family. These features are very different from those of the C3-C4 dimers previously described in CS, which had a much lower sequence identity and were located in pericentromeric regions. Homogenous alpha satellite HORs have long been considered to be specific to hominoid centromeres before recent studies proved the existence of such organizations in New World monkeys (Terada et al. 2013; Sujiwattanarat et al. 2015). The present observation supports the idea that HORs may be more common and more diverse than initially thought. Whether the newly discovered HORs from the CP genome are involved in centromere function or not remains to be investigated.

Emergence of new alpha satellite DNA families during Cercopithecini evolution

CP and CS share two main families of alpha satellite DNA, the centromeric C1 family and the pericentromeric C2 family. The sequence identity level is higher in C1 than in C2 for both species, which supports the hypothesis by which C1 has appeared more recently than C2. Interestingly, the structure of the PCA graph led us to postulate the existence in both species of many subfamilies of C1, each one deriving from a highly repeated sequence. The most abundant subfamily, which is derived from a sequence that exactly reflects the consensus of the C1 family, is conserved between both species, while most of the others are not conserved. These non-conserved subfamilies have probably emerged after the divergence of the CP and CS lineages, i.e. in a few million years of evolution.

Our FISH strategy, which makes use of oligonucleotide probes to distinguish localized sequence variations, cannot be used for labeling all these subfamilies specifically, because many of these subfamilies share one or several nucleotide variations and because oligonucleotide probes may hybridize to targets despite the presence of a single mismatch in the absence of carefully designed competitors (Cacheux et al. 2016). For example, the non specific hybridization of the C2A-G17Del probe on families derived from sequences 13, 25 or 28 may provide an explanation for the observed non perfect overlap between signals obtained with this probe and those obtained with the C6a probe. Nevertheless, in some cases, FISH experiments could be used for confirming the species specific distribution of the families and for investigating how this emergence proceeds. In particular they clearly showed that the distribution of some subfamilies is restricted to a subset of chromosomes. This suggests the existence of local amplification mechanisms which may be eventually followed by interchromosomal transfer. The observation of monomers from the so-called C5 family on both sides of centromeres supports the existence of successive events involving different subfamilies. In this specific case, amplification of C5 monomers would have been followed by amplification or integration, in the middle of the series of C5 monomers, of another sequence variant, which is still detected by the C1a probe but no more by the C5a probe.

Our data therefore point to the existence of recurrent amplification events affecting alpha satellite DNA. The amplification mechanism may lead to a monomeric organization, i.e. succession of monomers belonging to the same family, as demonstrated for C5, or to a higher order organization, as shown in the case of C6 and its associated monomers. The sequences that have been amplified never differ from the consensus sequence of the C1 family by more than 3 or 4 nucleotides. This property may be caused by the amplification mechanism itself, which would not act on divergent sequences, or from the elimination of amplified sequences that have excessively diverged. The consensus sequence of C1 has been itself the substrate of a major amplification event, probably before the divergence of Cercopithecini, but one cannot exclude that this sequence has been a substrate for amplification mechanisms after this divergence, i.e. concomitantly with the amplification of mutated sequence variants. The abundance of a specific variation, C158G, which was found in many of the subfamilies uncovered from the genome of CP, raises the question of a potential selective pressure that would favor the amplification or maintenance of this variation only in the CP lineage.

Dynamics of alpha satellite DNA in relation to chromosome evolution in Cercopithecini

The data we present here provide for the first time the opportunity to compare the chromosomal distribution of various alpha satellite DNA families in two Old World monkey species and to investigate how emergence of new alpha satellite DNA families is influenced by structural rearrangements of centromeres or chromosomes. Focusing our attention on two easily detectable subfamilies of C1, the so-called C5 and C6, we did not find evidence for a strong association between emergence of new sequence variants and centromere emergence or repositioning events, which suggests that evolution of centromeric DNA occurs independently from centromere rearrangement. The most prominent feature was the abundance of the C6 family on all acrocentric chromosomes. We also observed that chromosomes 2 were heterozygous for the presence of this family, suggesting some ongoing evolutionary processes.

We also compared the chromosomal distribution of the C2 family, which sequence composition was similar in both species. Our observations show that this family is highly abundant on the short arm of acrocentric chromosomes in both species. A similar observation has already been made in a New World monkey, *Aotus azarae* (Prakhongcheep et al. 2013).Therefore, exchange of genetic material between chromosomes that share a specific structural organization, such as acrocentric chromosomes, may be favored both at centromere and pericentromere. The bouquet chromosome configuration, occurring in prophase I, may play a role in these processes, as previously suggested (Paço et al. 2014). C2 was also found in the pericentromeric regions of several non acrocentric chromosomes in both species. Slight differences in the distribution of these sequences may result from differential elimination of these sequences from pericentromeric regions in different lineages.

Finally, one unexpected feature was the inability to detect any alpha-satellite repeat on a single chromosome 6. This feature was not observed on another metaphase preparation obtained from a female specimen, where both chromosomes 6 were equally labeled (not shown). This peculiar observation may be the result of a chromosomal rearrangement during cell culture, but may also reflect an heterozygotic individual carrying a chromosome 6 without any satellite DNA at its centromere, as it has been observed for example in orangutan or equids (Piras et al. 2010; Locke et al. 2011). Further sampling will be required to answer this question.

Conclusion

The characterization of the alpha satellite component of the CP genome provides for the first time the opportunity to compare the diversity and distribution of alpha satellite DNA in two related Old World monkey species, CP and CS. The major families of alpha satellite DNA, called C1 and C2, are conserved between both species as well as their gross distribution, but a detailed investigation led us to envision the presence of highly repeated sequences in our datasets as revealing numerous subfamilies of C1 that differ between both species. Each family is the result of evolutionary mechanisms that involve local amplification of a specific sequence variant followed by mutations of the amplified sequences. While most alpha satellite DNA is characterized by a monomeric organization in both species, higher organization patterns were detected specifically in CP. Emergence of new families seems to occur independently from rearrangements of centromeres. Our cytogenomic approach suggests different types of transfer or loss of genetic material that may explain the peculiar distribution of centromeric and pericentromeric sequences. Future work addressing other species within the Cercopithecini clade will help elucidating the evolutionary mechanisms as well as the functional significance of alpha satellite DNA variation.

Supplementary Material

Supplementary data are available at Molecular Biology and Evolution online.

Authors' Contributions

C.E. and F.A.R. conceived the study. L.C. and D.G. prepared the samples for sequencing. L.C. and L.P. performed the computational analysis. L.C, F.L. and M.G-S. performed the FISH experiments and image acquisitions. L.C. and M.G-S. reconstructed the karyotypes. L.C., L.P. and C.E. drafted the manuscript. All authors except F.A.R., who passed away during the study, read and approved the final manuscript.

Acknowledgments

We are indebted to Florence Anne Richard, to whom we dedicate this manuscript. We thank Anne-Marie and Bernard Dutrillaux for their advice in metaphase preparation and analysis. This research was supported by the Actions Thématiques Muséum "Génomique et Collections" and "Emergence".

Materials and methods

DNA collection and metaphase preparations

Fibroblast cell samples of CP (ID: 2001-027, male sample) from a cryo-conserved living cell bank (https://www.mnhn.fr/fr/collections/ensembles-collections/ressources-biologiques-cellules-vivantes-cryoconservees/tissus-cellules-cryoconserves-vertebres) were used for DNA extraction, which was performed using the Omega Biotek Tissue DNA Kit. Fibroblast cell samples of this same specimen were used for metaphase preparations. Cell cultures and metaphase preparations were achieved according to (Moulin et al. 2008).

Alpha satellite DNA isolation and sequencing

XmnI or HindIII were used to digest CP DNA in vitro. 10 µg of CP genomic DNA were incubated for 6h at 37°C with 70 units of XmnI or HindIII (New England Biolabs) in a total volume of 35 µL. The restriction enzymes were then inactivated for 20 min at 65°C. Both samples were loaded on a 1% agarose gel after addition of 7 µL loading buffer (50% glycerol) and electrophoresis was performed in 0.5X Tris-borate-EDTA buffer, at room temperature for 3 h at 100 V. The gel was briefly stained with ethidium bromide and then imaged by UV transillumination. Bands corresponding to alpha satellite monomers (~170 bp) were cut and DNA was extracted from the gel with the Omega Biotek Gel extraction kit and resuspended in 100 µl of elution buffer. About 250 ng were obtained for each of the XmnI and HindIII monomers.

Sequencing was performed on a PGM sequencing platform (Ion Torrent technology) using the 400 bp sequencing kit. HindIII DNA sample wasblunted according to the Quick Blunting Kit (E1201S, NEB). Twolibraries were generated using 50 ng of the twoblunt digest pools and the Ion Plus Fragment Library

Kit (4471252, Life Technologies) and tagged with Ion Xpress barcode adapters (4471250, Life Technologies). After purification (1.8X) with Ampure XP Beads (A63880, Agencourt Technology), the libraries were quantitated using a Sybr Green qPCR assay (SsoAdvanced supermix, Biorad) based on a custom *E. coli* reference library. After a dilution of each library down to 26 pM, 0.22 fmol of each library were pooled as templates for the clonal amplification on Ion Sphere particles during the emulsion PCR, performed on a One Touch2 emPCR robot according to the Ion PGM Template OT2 400 Kit user guide (4479878, Life Technologies). The amplification products were loaded onto an Ion 316v2 chip (4483324, Life Technologies), and subsequently sequenced according to the Ion PGM Sequencing 400 Kit user guide (4482002, Life Technologies). After standard filtration of the raw reads (polyclonal and low quality removal), the Ion Torrent sequencing yielded 210,527 sequences for the XmnI sample and 166,099 sequences for the HindIII sample. They were deposited in the NIH Short Read Archive (SRA accession numbers SRX1959818 and SRX1959815).

Alpha satellite sequence filtering

All XmnI sequences with an average Phred score lower than 25, a length outside the range 162-182 bp, and sequences without the XmnI digested sites at the extremities (5'-NNTTC … GAANN-3') were not considered for further analysis. Alpha satellite sequences were identified with a BLAST search against a reference alpha satellite sequence from *Chlorocebus aethiops* (AM23721) (Altschul et al. 1990). Using default BLAST parameters, all sequences exhibiting a hit longer than 80 bp were considered as alpha satellite sequences and conserved for the following analysis. All sequences were then reoriented if necessary in order to match the orientation of the reference alpha satellite sequence. The orientation information was preserved for investigations regarding reading biases.

All HindIII sequences with an average Phred score lower than 25, a length outside the range 166-186 bp (the blunting step added 4 nucleotides to the classic monomer length), and sequences without the HindIII digested or blunted sites at the extremities (5'-ATGC … ATGC-3') were not considered for further analysis. Alpha satellite sequences were identified with the same BLAST search as above. All sequences were then reoriented if necessary in order to match the orientation of the reference alpha satellite sequence. The four supplementary nucleotides added to the HindIII monomers during the blunting step (3'ATGC) were discarded.

Alpha satellite sequence characterization

Monomeric sequences were compared using their 5-mer composition in order to identify putative alpha satellite families without direct alignment. For each set of monomers, the 5-mer frequency table was analyzed using a principal component analysis (PCA) to reduce the space complexity and enable data visualization on the first factorial planes. Sequences were classified into groups by using a hierarchical clustering method (HCA) based on the Ward criterion (Ward 1963) applied to the Euclidean distances calculated from the 100 first principal components of the PCA. Because of the size of the monomer dataset, direct classification of the sequences using HCA was not possible. Instead, HCA was applied on 2500 randomly selected sequences which were used to train a linear discriminant model. This model has been finally used to classify all the other monomers.

Because of the size of the datasets, the consensus sequences and the sequence distance analysis were conducted with different subsets of randomly selected sequences. The selected sequences were aligned using MUSCLE (Edgar 2004) and analyzed with Seaview (Gouy et al. 2010). CENP-B and pJalpha boxes were searched with the patterns TTCGTTGGAARCGGGA and TTCCTTTTYCACCRTAG respectively (Rosandić et al. 2006) by using the program Fuzznuc (Rice et al. 2000) and allowing 2 mismatches. All statistical analyses were conducted with R (R Core Team 2014). Our R scripts and other programs are available upon request.

Oligonucleotide probes

Short oligonucleotide probes (18 or 19 nucleotides) were designed in order to target specifically the different alpha satellite families identified in CP, by systematic prediction of binding frequencies based on the sequencing results. Sequences and binding frequencies are available in supplementary table S2, which also provides details about the positions of locked nucleic acid (LNA) modifications in the probes. These positions were selected based on previous experience in order to achieve a good binding affinity and specificity (Ollion et al. 2015; Cacheux et al. 2016). When possible, we selected probes that were perfectly complementary to more than 20% of the sequences from the target group and to less than 3% of the sequences from the other groups. Table S2 also provides the expected binding frequencies if hybridization is possible despite the presence of one mismatch between the probe and its targets. Additional probes were used to localize specific sequence variants, such as C2A-G17Del (5'CaTTtTcCcTtCaAgAaTcC3', 3'Biotin), 158C (5'CaCaAgAaCAgCcTtAgC3',3'Digoxygenin) and 158G (5'CaCaAgAaGAgCcTtAgC3', 3'Biotin). All probes were purchased from Eurogentec (Seraing, Belgium).

FISH experiments

FISH were performed on metaphase chromosome preparations. Hybridization solutions were prepared by diluting the oligonucleotide probes to a final concentration of 0.1 µM in a hybridization solution consisting of 2X SSC pH 6.3, 50% deionized formamide, 1X Denhardt solution, 10% dextran sulfate, and 0.1% SDS. 20 µL of the hybridization solution were deposited on each slide and covered with a coverslip. The slides were then heated for 3 min at 70 °C and hybridized for 1 h at 37 °C in a Thermobrite apparatus (Leica Biosystems). Then, each slide was washed twice in 2X SSC at 63°C. Preparations were then incubated in blocking solution (4% bovine serum albumin (BSA), 1X PBS, 0.05% Tween 20) for 30 min at 37 °C to reduce nonspecific binding. Then, depending on the combination of probes, the following antibodies were used for subsequent revelations: Alexa 488-conjugated streptavidin (1:200; Life Technologies), Cy5-conjugated streptavidin (1:200; Caltag Laboratories), FITC-conjugated sheep anti-digoxigenin (1:200; Roche), and Rhodamine-conjugated sheep anti-digoxigenin (1:200; Roche). All antibodies were diluted in blocking solution containing 1X PBS, 0.05% Tween 20, and 4% BSA. Antibody incubation lasted for 30 min at 37 °C. All washings were performed in 2X SSC, 0.05% Tween 20. Chromosomes were counterstained with DAPI (4',6-diamidino-2-phenylindole) by pipetting 40 μL of a 5 μg/mL solution onto the slides, incubating for 5 min and then briefly washing in 1X PBS. Slides were mounted by adding a drop of Vectashield Antifade Mounting Medium (Vector Laboratories) and covering with a coverslip. Metaphases were imaged using an Axio Observer Z1 epifluorescent inverted microscope (Zeiss) coupled to an ORCA R2 cooled CDD camera (Hamamatsu). The Axio Observer Z1 was equipped with a Plan-Apochromat 63x 1.4 NA oil-immersion objective and the following filters set: 49 shift free for DAPI (G365 / FT395 / BP445/50), 38 HE shift free for FITC/Alexa488 (BP470/40 / FT495 / BP525/50), homemade sets for Rhodamine (BP546/10 / FF555 / BP 583/22) and for Cy5 (BP643/20 / FF660 / BP684/24). The light source was LED illumination (wavelengths: 365nm, 470nm or 625nm) except for Rhodamine, for which a metal halide lamp HXP120 was preferred. Immersion oil of refractive index 1.518 at 23°C was used.

References

Alexandrov I a, Mitkevich SP, Yurov YB. 1988. The phylogeny of human chromosome specific alpha satellites. Chromosoma 96:443–453.

Alkan C, Ventura M, Archidiacono N, Rocchi M, Sahinalp SC, Eichler EE. 2007. Organization and evolution of primate centromeric DNA from whole-genome shotgun sequence data. PLoS Comput. Biol. 3:1807–1818.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. J. Mol. Biol.:403–410.

Archidiacono N, Antonacci R, Finelli P, Lonoce A, Rocchi M. 1995. Comparative Mapping of Human Alphoid Sequences in Great Apes Using Fluorescence. Genomics 484:477–484.

Cacheux L, Ponger L, Gerbault-Seureau M, Richard FA, Escudé C. 2016. Diversity and distribution of alpha satellite DNA in the genome of an Old World monkey: Cercopithecus solatus. BMC Genomics [Internet] 17:916. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-3246-5

Catacchio CR, Ragone R, Chiatante G, Ventura M. 2015. Organization and evolution of Gorilla centromeric DNA from old strategies to new approaches. Sci. Rep. [Internet] 5:14189. Available from: http://www.nature.com/doifinder/10.1038/srep14189

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.

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. [Internet] 34:1669–1681. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msx108

Dover G. 1982. Molecular drive: a cohesive mode of species evolution. Nature 299:111–117.

Dutrillaux B, Couturier J, Chauvier G. 1980. Chromosomal evolution of 19 species of sub-species of Cercopithecinae. Ann Genet. 23:133–143.

Edgar RC. 2004. MUSCLE : a multiple sequence alignment method with reduced time and space complexity. 19:1–19.

Feliciello I, Akrap I, Brajkovi J, Zlatar I, Ugarkovi Ur Ica. 2014. Satellite DNA as a Driver of

Population Divergence in the Red Flour Beetle Tribolium castaneum. Genome Biol. Evol. [Internet] 7:228–239. Available from: http://gbe.oxfordjournals.org/cgi/doi/10.1093/gbe/evu280

Garrido-ramos MA. 2017. Satellite DNA : An Evolving Topic.

Gouy M, Guindon S, Gascuel O. 2010. SeaView Version 4 : A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. 27:221–224.

Guschanski K, Krause J, Sawyer S, Valente LM, Bailey S, Finstermeier K, Sabin R, Gilissen E, Sonet G, Nagy ZT, et al. 2013. Next-Generation Museomics Disentangles One of the Largest Primate Radiations. Syst. Biol. 62:539–554.

Hayden KE. 2012. Human centromere genomics: Now it's personal. Chromosom. Res. 20:621–633.

Jorgensen a L, Jones C, Bostock CJ, Bak a L. 1987. Different subfamilies of alphoid repetitive DNA are present on the human and chimpanzee homologous chromosomes 21 and 22. 6:1691–1696.

Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth L V, Muzny DM, Yang S-P, Wang Z, Chinwalla AT, Minx P, et al. 2011. Comparative and demographic analysis of orang-utan genomes. Nature 469:529–533.

Moulin S, Gerbault-Seureau M, Dutrillaux B, Richard FA. 2008. Phylogenomics of African guenons. Chromosom. Res. 16:783–799.

Ollion J, Loll F, Cochennec J, Boudier T, Escudé C. 2015. Cell cycle-dependent positioning of individual centromeres in the interphase nucleus of human lymphoblastoid cell lines. J. Cell Biol. (submitted.

Paço A, Adega F, Mestrovic N, Plohl M, Chaves R. 2014. Evolutionary Story of a Satellite DNA from Phodopus. 6:2944–2955.

Palomeque T, Lorite P. 2008. Satellite DNA in insects: a review. Heredity (Edinb). 100:564–573.

Pérez-Gutiérrez M a., Suárez-Santiago VN, López-Flores I, Romero AT, Garrido-Ramos M a. 2012. Concerted evolution of satellite DNA in Sarcocapnos: A matter of time. Plant Mol. Biol. 78:19–29.

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.

Plohl M, Luchetti A, Meštrović N, Mantovani B. 2008. Satellite DNAs between selfishness and functionality: Structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. Gene 409:72–82.

Prakhongcheep O, Hirai Y, Hara T, Srikulnath K, Hirai H, Koga A. 2013. Two types of alpha satellite DNA in distinct chromosomal locations in Azara's owl monkey. DNA Res. 20:235–240.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. [Internet] 16:276–277. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0168952500020242

Rosandić M, Paar V, Basar I, Glunčić M, Pavin N, Pilaš I. 2006. CENP-B box and pJα sequence distribution in human alpha satellite higher-order repeats (HOR). Chromosom. Res. 14:735–753.

Rudd MK, Wray G a, Willard HF. 2006. The evolutionary dynamics of alpha -satellite. Genome Res. 16:88–96.

Schindelhauer D, Schwarz T. 2002. Evidence for a fast, intrachromosomal conversion mechanism from mapping of nucleotide variants within a homogeneous ??-satellite DNA array. Genome Res. 12:1815–1826.

Schueler MG, Dunn JM, Bird CP, Ross MT, Viggiano L, Rocchi M, Willard HF, Green ED. 2005. Progressive proximal expansion of the primate X chromosome centromere. Proc. Natl. Acad. Sci. U. S. A. 102:10563–10568.

Schueler MG, Sullivan B a. 2006. Structural and functional dynamics of human centromeric chromatin. Annu. Rev. Genomics Hum. Genet. 7:301–313.

Shepelev V a., Alexandrov A a., Yurov YB, Alexandrov I a. 2009. The evolutionary origin of man can be traced in the layers of defunct ancestral alpha satellites flanking the active centromeres of human chromosomes. PLoS Genet. 5.

Sujiwattanarat P, Thapana W, Srikulnath K, Hirai Y, Hirai H, Koga A. 2015. Higher-order repeat structure in alpha satellite DNA occurs in New World monkeys and is not confined to hominoids. Sci. Rep. [Internet] 5:10315. Available from: http://www.nature.com/doifinder/10.1038/srep10315

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Terada S, Hirai Y, Hirai H, Koga A. 2013. Higher-order repeat structure in alpha satellite DNA is an attribute of hominoids rather than hominids. J. Hum. Genet. [Internet] 58:752–754. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23945983

Tosi AJ. 2008. Forest monkeys and Pleistocene refugia: A phylogeographic window onto the disjunct distribution of the Chlorocebus lhoesti species group. Zool. J. Linn. Soc. 154:408–418.

Utsunomia R, Ruiz-ruano FJ, Silva DMZA, Serrano ÉA, Rosa IF, Scudeler PES, Hashimoto DT. 2017. A Glimpse into the Satellite DNA Library in Characidae Fish (Teleostei , Characiformes). 8:1–11.

Vissel B, Choo KH. 1991. Four distinct alpha satellite subfamilies shared human. 19:271–277.

Warburton PE, Haaf T, Gosden J, Lawson D, Willard HF. 1996. Characterization of a chromosome-specific chimpanzee alpha satellite subset: evolutionary relationship to subsets on human chromosomes. Genomics 33:220–228.

Ward JH. 1963. Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. 58:236–244.

Waye J, Willard H. 1986. Structure, organization, and sequence of alpha satellite DNA from human chromosome 17: evidence for evolution by unequal crossing-over and an ancestral pentamer repeat shared with the human X chromosome. Mol. Cell. Biol. [Internet] 6:3156–3165. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=367051&tool=pmcentrez&rendertype=abstract

Table

Id	**Sequence**	**Number**	**Forward(%)**
1	Consensus C1	2983	46
2	C158G	848	48
3	C116T	568	41
4	C114Del	508	1
5	C137A-CC149AA	455	34
6	T101Del	323	98
7	C2A-G17Del	250	66
8	C2A-G17Del-C158G	208	70
9	C114Del-C158G	145	0
10	A3741T-G64A-C158G	136	15
11	A40C-C42G	116	44
12	C116T-C158G	112	46
13	C2A	103	73
14	T121A	100	43
15	C137A-C158G	100	51
16	A3741T-G64A	100	24
17	C137A-CC149AA-C114Del	89	1
18	C2A-G17Del-C114Del	81	1
19	T38G	77	29
20	A110G	76	56
21	A86T	74	39
22	T80Del-T101Del	67	100
23	A41G	65	38
24	T101Del-C158G	62	98
25	G17Del	59	47
26	G17C	58	54
27	C144A	57	46
28	C2A-C158G	54	54
29	C137A	54	65
30	A40C-C42G-G28T	53	49

Table 1. Analysis of alpha satellite sequences found in high copy number inthe CP XmnI monomer dataset. The sequences are named according to the “Id” column. The “Sequence” column indicates how each sequence variant differs from the consensus sequence of the C1 family, using standard notations. The “Number” column displays the number of identical copies of the sequence in the monomer dataset. The “Forward” column displays the percentage of reads obtained in the forward orientation (i.e. the orientation of our reference sequence). Strong biases for read orientation reveal artifactual sequences which are indicated on a grey background.

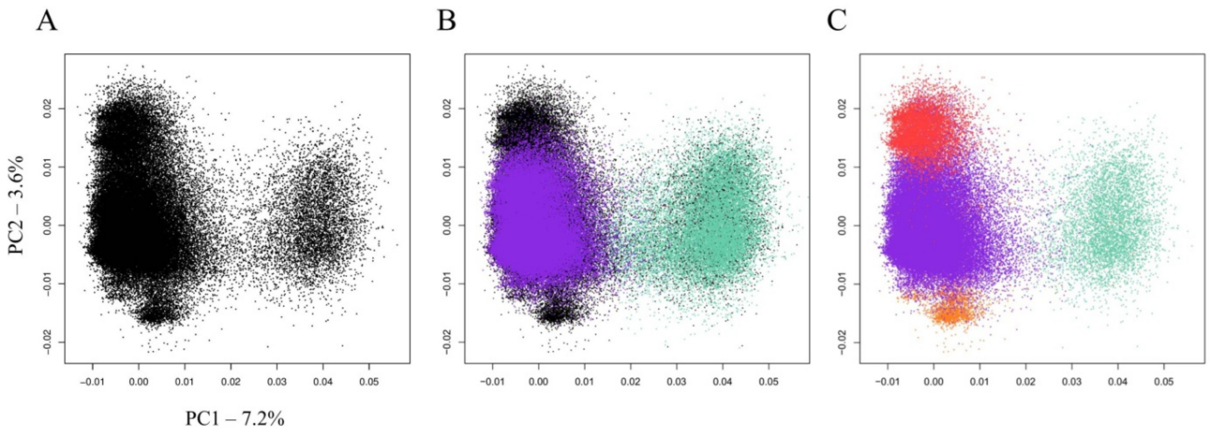
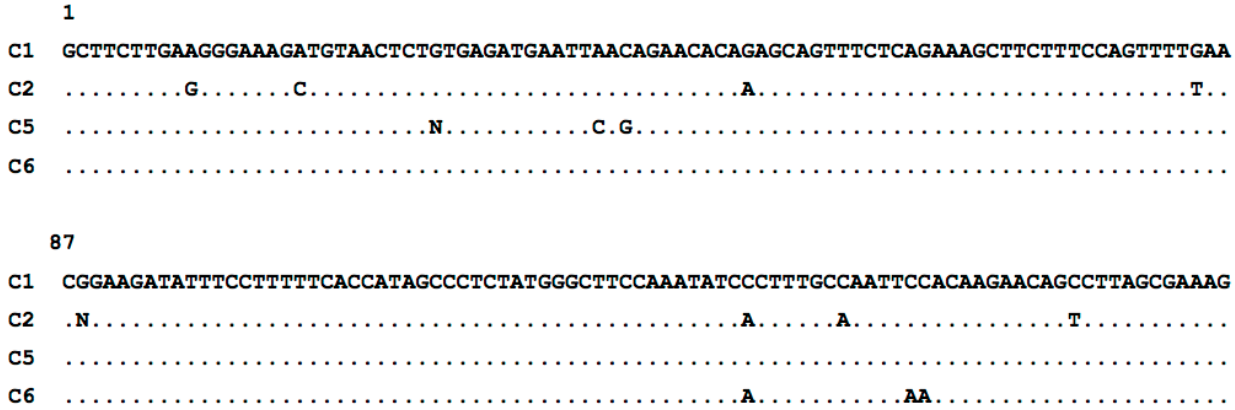


Fig. 1. Characterization of alpha satellite DNA diversity in the *XmnI* monomer dataset. (*A*) PCA projection on principal components 1 and 2 of the normalized 5-mer frequency vectors for all sequences from the CPXmnImonomer dataset. Each point represents a monomer sequence. (*B*) Prediction of the C1 (purple) and C2 (pastel green) sequences from CS XmnI monomer dataset by using the PCA projection of CP monomers. (*C*) PCA projection of CP XmnI monomer dataset with sequences colored according to their assignment to the C1 (purple), C2 (pastel green), C5 (red) or C6 (orange) alpha satellite family, based on a hierarchical classification method (see Materials and methods).



```
      1
C1    GCTTCTTGAAGGGAAAGATGTAACTCTGTGAGATGAATTAACAGAACACAGAGCAGTTTCTCAGAAAGCTTCTTTCCAGTTTTGAA
C2    .........G.......C................................A................................T..
C5    ...........................N...........C.G............................................
C6    ......................................................................................

     87
C1    CGGAAGATATTTCCTTTTTCACCATAGCCCTCTATGGGCTTCCAAATATCCCTTTGCCAATTCCACAAGAACAGCCTTAGCGAAAG
C2    .N................................................A......A................T...........
C5    ......................................................................................
C6    ..................................................A...........AA......................
```

Fig. 2. Consensus sequences of the alpha satellite families identified in the CP XmnI dataset. The consensus sequences were determined following the alignment of 500 randomly selected sequences for the C1, C2, C5 and C6 families. Each position was considered unambiguous if more than 60% of monomers had the same nucleotide at this position. A point at a position replaces a nucleotide identical to the nucleotide at the homologous position in the C1 consensus.

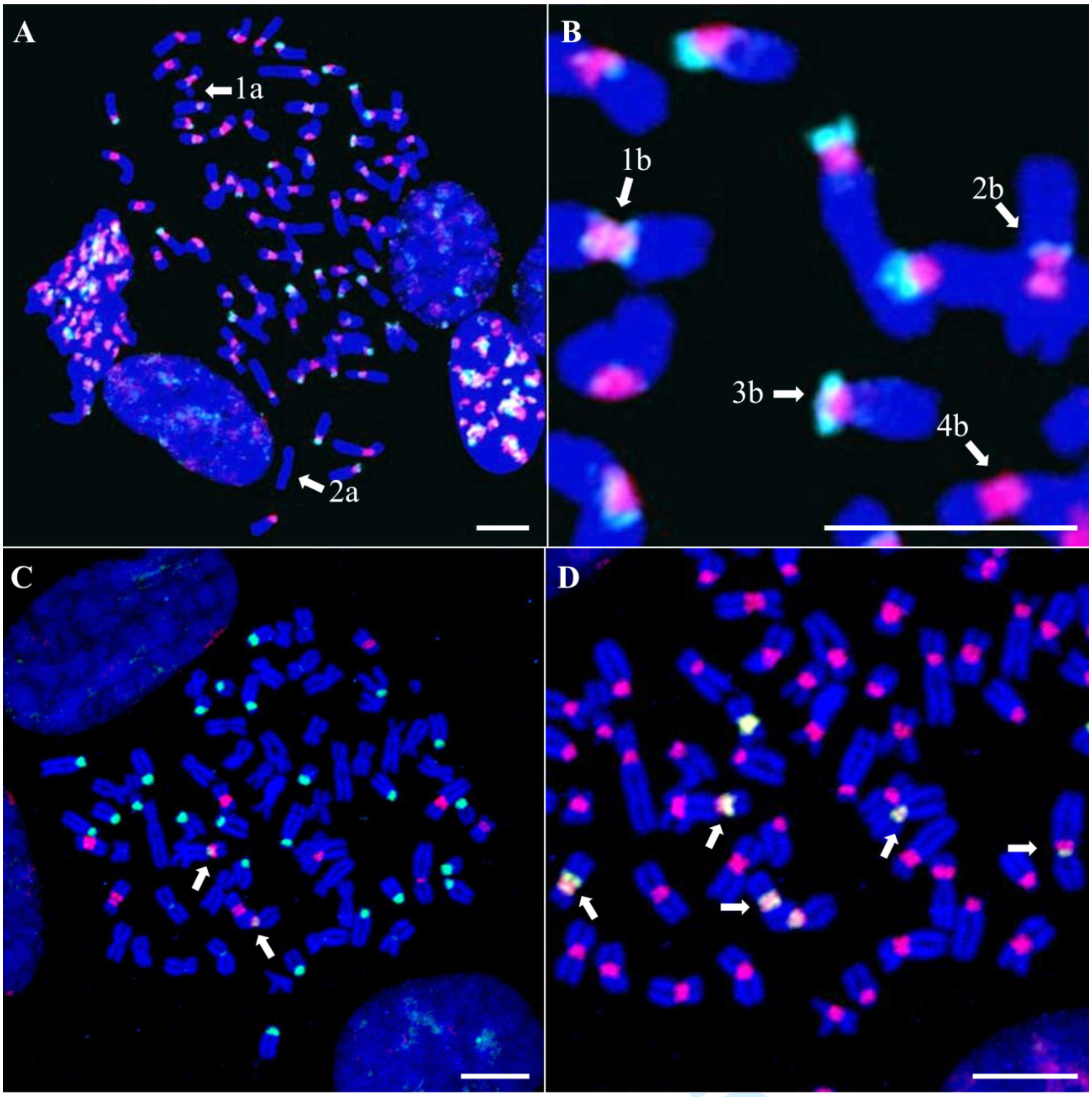


Fig. 3. FISH analysis of the C1, C2, C5 and C6 alpha satellite families on CP chromosomes.CPmetaphase chromosomes are colored in blue. (*A,B*) Probes C1a, C1b and C2b are hybridized simultaneously. (*A*) Hybridization of probes C1a and C1b (red) and probe C2b (green). 1a and 2a: unlabeled chromosomes. (*B*) Focus on image (*A*) showing in details the different types of distribution of the C2b signals. 1b: both pericentromeric regions, 2b: one pericentromeric region toward the long arm, 3b: one pericentromeric region toward the short arm of an acrocentric chromosome, 4b: no signal. (*C, D*) Probes C1a, C5a and C6a are hybridized simultaneously. (*C*) Hybridization of probe C5a (red) and probe C6a (green). Arrows: two chromosomes where both probes produce signals. (*D*) Focus on the metaphase shown in (*C*) but with hybridization of probe C1a (red) and probe C5a (green). Arrows: pericentromeric hybridization of probe C5a on several chromosomes. Scale bar = 10 µm.

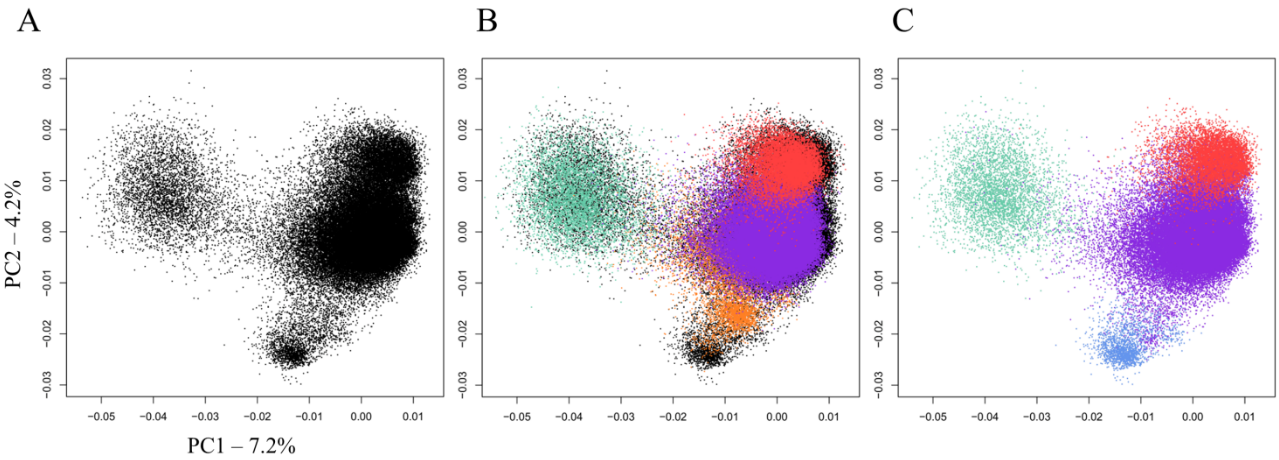


Fig. 4. Characterization of alpha satellite DNA diversity in the *HindIII* monomer dataset. (*A*) PCA projection on principal components 1 and 2 of the normalized 5-mer frequency vectors for all sequences from the *HindIII* monomer dataset. Each point represents a monomer sequence. (*B*) Prediction of the position of the XmnI monomer sequences on the graph shown in (*A*). Sequences are colored according to their assignment to the C1 (purple), C2 (pastel green), C5 (red) or C6 (orange) alpha satellite families. (*C*) PCA projection shown in (*A*) with sequences colored according to their assignment to the C1' (purple), C2' (pastel green), C5' (red) or C6' (blue) alpha satellite families, using a hierarchical classification method (see Materials and methods).

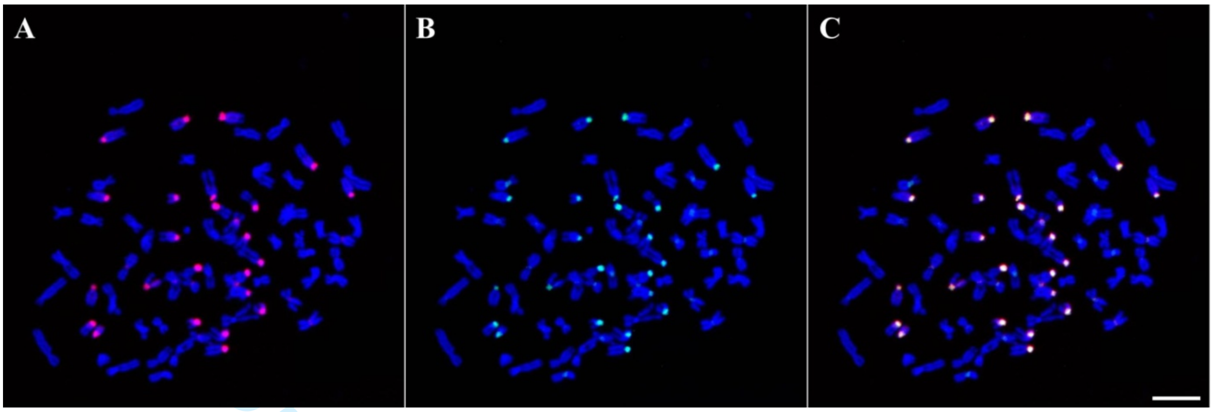


Fig. 5. FISH detection of the C2A-G17Del sequence variant and relative distribution with respect to the C6a probe.

Probes C2A-G17Del and C6a are hybridized simultaneously to CP metaphase chromosomes, which are colored in blue. (*A*) Hybridization of probe C6a is shown in red. (*B*) Hybridization of probe C2A-G17Del is shown in green. (*C*) Combined signals from (*A*) and (*B*). Scale bar = 10 µm.

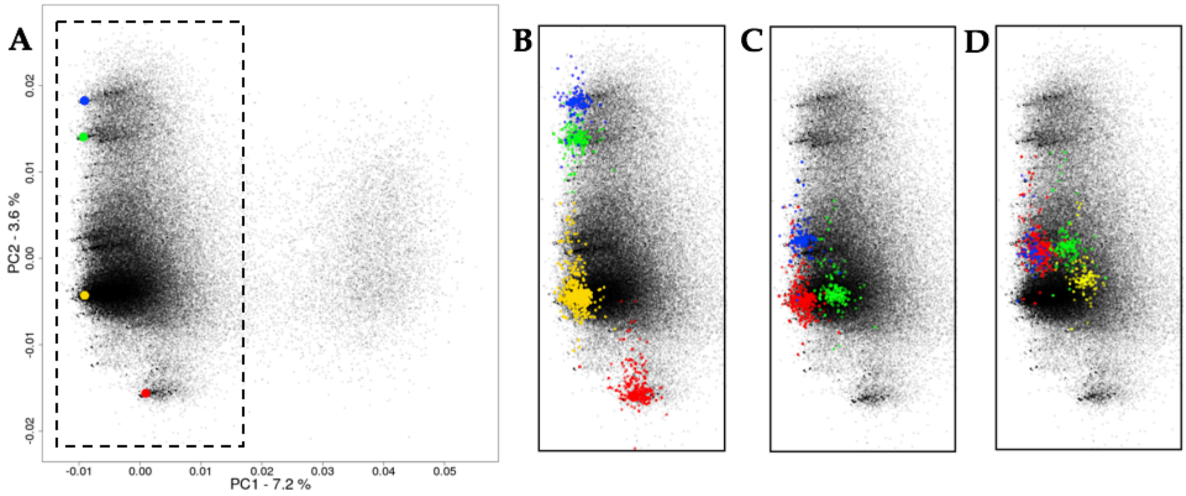


Fig. 6. Distribution of sequences into comet-like clusters near abundant sequence variants.

(*A*) PCA projection on principal components 1 and 2 of the normalized 5-mer frequency vectors for all sequences from the XmnI monomer dataset is shown here with a lower point density than the one shown on Figure 1. Sequences corresponding to highly repeated sequences 1 (yellow), 5 (red), 11 (green) and 30 (blue) are highlighted. (*B*), (*C*) and (*D*) Only the region of the PCA projection corresponding to the dotted rectangle (i.e. to the C1 family) is shown. (*B*) Sequences from the dataset that correspond to single nucleotide variations from sequences 1, 5, 11 and 30 are shown using the same color code as in (A). (*C*) Sequences from the dataset that correspond to single nucleotide difference from sequences 3, 7, and 19 are shown in red, green and blue, respectively. (D) Sequences from the dataset that correspond to single nucleotide difference from sequences 2, 8, 12, 15 are shown in red, green, blue and yellow, respectively.

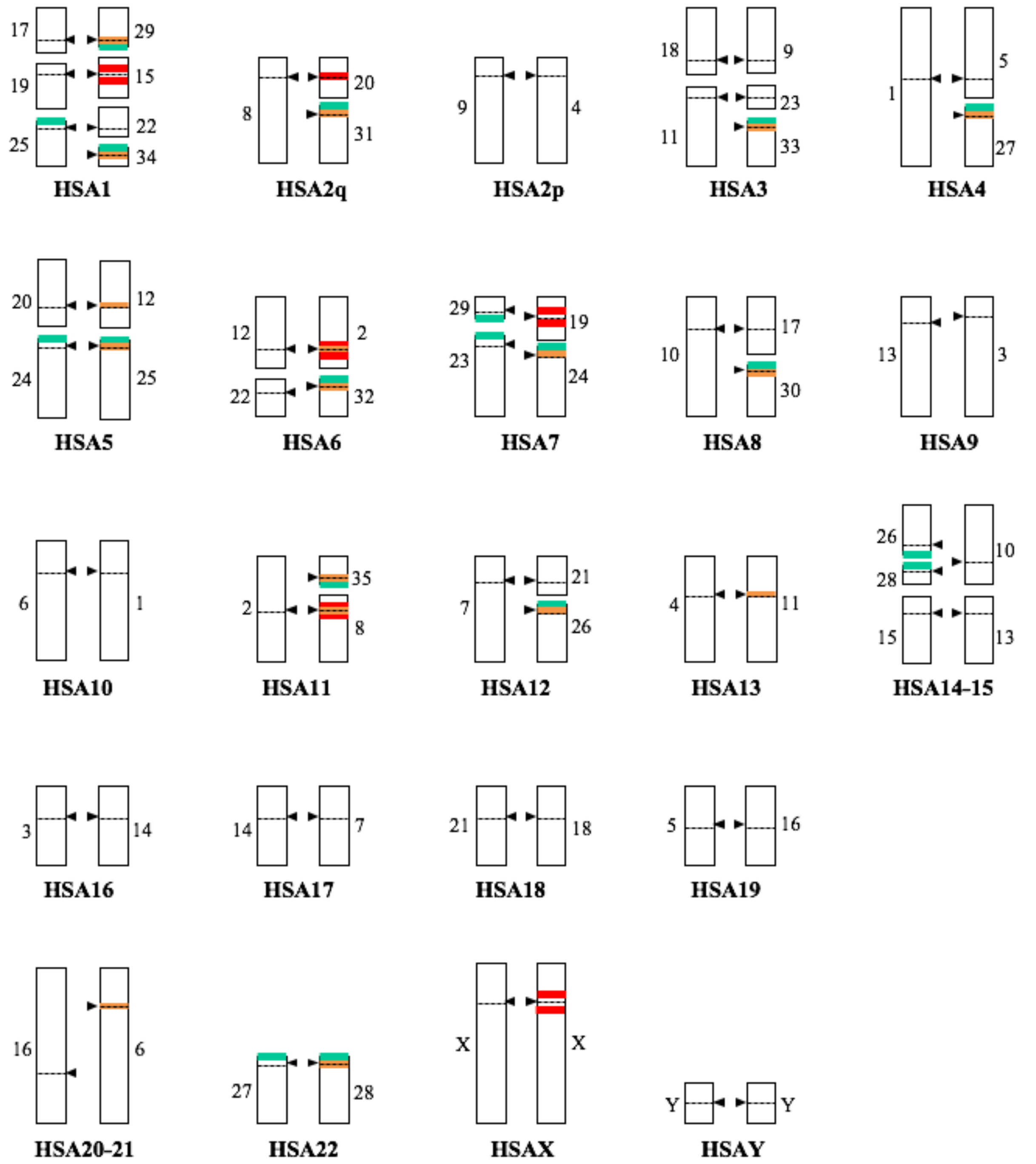


Fig. 7. Scheme representing the distribution of alpha satellite families C2, C5 and C6 on CS and CP chromosomes. Homologous chromosomes have been aligned using human chromosomes as references. CS and CP chromosomes are always shown on the left and right hand side, respectively. Numbering below each set of chromosomes refers to human chromosome numbers. Homologies are taken from Moulin et al. (2008). Arrows and dotted lines point to centromere positions. Distribution is shown in pastel green for C2, red for C5 and orange for C6. For C2, only the strong signals located on acrocentric chromosomes are shown.