UNIVERSITA' DEGLI STUDI DI PAVIA

Dipartimento di Biologia e Biotecnologie "Lazzaro Spallanzani"

Dynamics and evolution of mammalian centromeres and satellite DNA

Eleonora Cappelletti

Dottorato di Ricerca in
Genetica, Biologia Molecolare e Cellulare
Ciclo XXXI – A.A. 2015-2018

UNIVERSITA' DEGLI STUDI DI PAVIA

Dipartimento di Biologia e Biotecnologie "Lazzaro Spallanzani"

Dynamics and evolution of mammalian centromeres and satellite DNA

Eleonora Cappelletti

Supervised by Prof. Elena Giulotto

Dottorato di Ricerca in
Genetica, Biologia Molecolare e Cellulare
Ciclo XXXI – A.A. 2015-2018

Abstract

The centromere is a specialized nucleoprotein structure of the eukaryotic chromosome whose role is ensuring proper segregation of sister chromatids during cell division. In centromeric chromatin, despite the evolutionary conservation of proteins, DNA sequences are highly variable. This paradox is now explained by the well-established knowledge that the centromeric function is epigenetically specified and CENP-A, the centromere-specific histone H3 variant, is the major determinant. Although dispensable for centromeric function, satellite DNA has been proposed to contribute to centromere stability and organization. In spite of their high divergence, centromeric satellites share the presence of a common motif, the so-called CENP-B box, which is recognized by CENP-B, the only known centromeric protein that exhibits unequivocal DNA binding specificity. However, the role of CENP-B in the epigenetic establishment of centromeric chromatin remains controversial.

Although satellite DNA (highly repetitive DNA) is a common feature of mammalian centromeres, in our laboratory we proved that, in *Equus* species (horses, asses and zebras), satellite DNA is uncoupled from centromeric function: beyond the classical satellite-based ones, several centromeres are completely satellite-free, whereas many satellite DNA loci are not centromeric, representing a powerful model system to investigate the epigenetic centromeric function in relationship with satellite DNA.

Following the previous discovery of the only satellite-less centromere of the horse (Wade et al. 2009, Purgato et al. 2015), we identified and characterized by ChIP-seq with an anti-CENP-A antibody an extraordinarily high number of satellite-less centromeres (16 out 31) in the donkey, demonstrating that the presence of more than half of centromeres void of satellite DNA is compatible with genome stability and species survival. The presence of amplified DNA at some centromeres suggests that these arrays may represent an intermediate stage toward satellite DNA formation during evolution. As expected from the absence of satellite DNA, these satellite-less centromeres lack any recognition site for CENP-B.

We characterized at the molecular level, the satellite-based centromeres of the horse, identifying 37cen satellite as the major centromeric satellite DNA sequence, which is organized in a head-to-tail fashion and is transcriptionally active. Surprisingly, this satellite does not contain any recognition site for CENP-B, suggesting a peculiar pattern of interaction between CENP-B and centromeres in the equid species. Using a combination

of ChIP-seq and cytogenetic approaches, we demonstrated that CENP-B binds a novel satellite DNA family, the CENPB-sat, and that the genus *Equus* is characterized by marked uncoupling between CENP-B and CENP-A. In the horse, CENP-B domains are restricted to a subset of pericentromeres and are excluded from the centromeric core. In the donkey and the Burchell's zebra, the progressive reduction and degeneration of binding sites have led to the disappearance of detectable levels of CENP-B binding at all chromosomes. On the other hand, in the Grevy's zebra CENP-B is present mainly at non centromeric chromosomal termini, interpreted as the relics of ancestral inactivated centromeres, while CENP-B is undetectable at most active centromeres. Taken together, our results suggest that the uncoupling between the centromeric function and CENP-B that marked equid phylogeny could explain the exceptional plasticity of equid centromeres.

These conclusions are supported by our finding in another model organism, the rodent species Chinese hamster. As for the genus *Equus*, the CENP-B binding motif is not contained in the major centromeric satellite, which in this species corresponds to telomeric-like TTAGGG repeats. The karyotype of Chinese hamster derives from chromosomal fusions and fission events during karyotype evolution, which led to centromeric localization of telomeric arrays. It is tempting to speculate that these rearrangements were facilitated by the uncoupling between CENP-B and the centromeric function.

During my thesis work, I also studied another aspect of centromere biology that is the localization of centromeres in the tridimensional nuclear architecture. As previously described in the literature, a prominent feature of the mammalian nucleus is the clustering of centromeres at the nuclear and nucleoli periphery. However, it is a matter of debate whether centromere clustering depends on the presence of satellite repeats or on the centromeric function. Taking advantage of the variable satellite DNA localization in the genus *Equus*, we demonstrated that the clustering phenomenon relies on the presence of satellite repeats and not on the centromeric function.

Finally, I investigated the basis of the inhibition of meiotic recombination which is exerted by the centromere, taking advantage of the genus *Equus*. In particular, we demonstrated that a satellite-less centromere exerts the same inhibitory effect on meiotic recombination as a classical satellite-based centromere. This result suggests that the "centromere effect" on meiotic recombination does not depend on the presence of satellite DNA. During this analysis, we observed a peculiar phenomenon in horse spermatocytes at the pachytene phase of meiosis: double-spotted centromeres were detected on a few chromosome bivalents by immunofluorescence. The number of these peculiar centromeres varied from 0 to 7 and inter- and intra-

individual variability of their frequency was found. This observation could be explained by different mechanisms: positional variation of the centromeric domain of the two homologous chromosomes, misalignment of pericentromeric and centromeric satellite DNA arrays during homolog pairing or a combination of both.

Acknowledgements

I would like to express my gratitude to my supervisor, Prof. Elena Giulotto, for her guidance during all these years. Her teaching was crucial to improve myself and proceed in the right direction. She encouraged me in growing as a researcher.

I would specially like to thank Maria Francesca Piras, who became a precious friend, for her essential teaching, patience and continuous support from the first steps in the laboratory and in "our CENP-B project". Without her this work would not have been possible and I learned a lot from her.

I would like to thank Solomon Nergadze for his guide in the first months in the laboratory and for his help during all these years.

A special mention to the other members of the laboratory, in particular Marco Santagostino, Lela Khoriauli, Alexandra Smirnova, Claudia Badiale, Riccardo Gamba, Marco Corbo, Rosa Maria Di Mauro and Francesco Gozzo for their help, assistance and friendship. I would also like to remember Federico Cerutti, whose bright personality will never be forgotten. All of them made these years extraordinary.

I would also like to thank all the other past and present members of the laboratory and, in particular, Demetrio Turati and Marina Bambi, whose help was very important for different parts of this thesis work.

Finally, I would like to thank Dr. Irina Solovei, for having accepted me in her laboratory and for her teaching, Prof. Aurora Ruiz-Herrera, for the valuable contribution and advice, and all the collaborators, in particular Prof. Kevin Sullivan, Prof. Giulio Pavesi, and Prof. Elena Raimondi, for helpful discussion and suggestions.

Contents

Abbreviations

BAC: bacterial artificial chromosome
bp: base pair
CDS: coding sequence
cen: centromere
CENP: centromere protein
CGR: *Cricetulus griseus*
ChIP-seq: chromatin immunoprecipitation-sequencing
chr: chromosome
CREST: Calcinosis, Raynaud phenomenon, esophageal dysmotility, sclerodactyly, telangiectasia
Da: Dalton
dist: distal region
dNTP: deoxynucleotide triphosphate
EAS: *Equus asinus*
EBU: *Equus burchelli*
ECA: *Equus caballus*
EGR: *Equus grevyi*
ENC: evolutionary new centromere
ERE: equine repetitive element
FISH: fluorescence *in situ* hybridization
HOR: high order repeat
HSA: *Homo sapiens*
kb: kilobase
Mb: megabase
MMU: *Mus musculus*
MYA: million years ago
NCBI: National Center for Biotechnology Information
NGS: next generation sequencing
p: short arm of a chromosome
PAK: perissodactyl ancestral karyotype
prox: proximal region
PCR: polymerase chain reaction
q: long arm of a chromosome
UTR: untranslated region

Introduction

1. THE CENTROMERE

The centromere is a specialized nucleoprotein structure of the eukaryotic chromosome whose role is ensuring proper segregation of sister chromatids during cell division. Actually, the centromere is the site of kinetochore assembly and spindle fiber attachment.

Among Eukaryotes, three different types of centromeres have been identified: regional centromeres, holocentric centromeres and point centromeres (Clarke 1998, Choo 2000, Nagaki et al. 2005). Regional centromeres, characteristic of higher eukaryotes, extend over large regions (from tens to a few thousand of kilobase pairs) and cytologically appear as distinct primary constrictions in metaphase chromosomes. These centromeres generally consist of long stretches of highly reiterated DNA (satellite DNA) and/or retrotransposable elements, spanning from tens of kilobases to several megabases (Kalitsis and Choo 2012). On the other hand, holocentric centromeres, peculiar of some plants, nematodes and insects, span the entire chromosome and the whole chromosome acquires centromeric function (Choo 2000, Nagaki et al. 2005). Finally, point centromeres, typical of *S. cerevisiae*, cover only few hundred nucleotides and associated kinetochores bind a single microtubule (Clarke 1998, Cleveland et al. 2003).

S. cerevisiae is the only eukaryotic organism in which the centromeric function is entirely determined by the sequence (Clarke 1998, Cleveland et al. 2003). As a matter of fact, in all eukaryotes, with the exception of this yeast, we can describe the relationship between centromeric DNA sequence and function with the so-called “centromere paradox” (Henikoff et al. 2001). Although the centromeric function is well conserved along the evolutionary tree, centromeric DNA sequences are highly divergent among taxa and also between chromosomes of the same cell (Choo 2000, Henikoff et al. 2001, Cleveland et al. 2003, Plohl et al. 2008,). Moreover centromeric DNA sequences are not intrinsically sufficient to nucleate the centromeric function (Choo 2000, Cleveland et al. 2003, Kalitsis and Choo 2012). Although, mitotically stable human isodicentric chromosomes contain two identical, well-separated regions of centromeric DNA, only one active centromere is formed, suggesting that the centromeric sequence is not enough for centromere establishment (Choo 2000, Marshall et al. 2008). Finally, centromere formation can occur in hitherto non-centromeric chromosomal

regions that are usually devoid of canonical centromeric repeated DNA (Choo 2000).

All these observations support the model of the "epigenetic centromere" (Cleveland et al. 2003). According to this model, the centromere is an epigenetic locus which behaves as a self-replicating protein complex that resides on centromere DNA but is not determined by it (Cleveland et al. 2003, Allshire and Karpen 2008).

2. CENTROMERIC AND PERICENTROMERIC SATELLITE DNA

Satellite DNA consists of tandem arrays of a repeated sequence, which represents the monomer unit. As mentioned before, regional centromeres of higher eukaryotes are classically associated with satellite sequences in spite of the lack of a strict DNA sequence dependency. The presence of satellite DNA is not restricted to the centromere but spread in the extended pericentromeric region. In particular, pericentromeric satellite DNA far surpasses centromeric satellite sequences in abundance, constituting up to 50% of genomes in certain cases (Jagannathan and Yamashita 2017, Garrido-Ramos 2017).

Satellite sequences represent the most rapidly evolving DNA sequences in eukaryotic genomes and are highly divergent even among related species (Plohl et al. 2014). Their extraordinary variability is due to the fact that sequence variants are easily fixed by expansion and contraction and can arise *de novo* at new sites (Henikoff et al. 2001). In particular, it is well accepted that satellite DNA, as well as many other repetitive sequences, evolves through the so-called "concerted evolution". Concerted evolution is defined as the non-independent evolution of repetitive DNA sequences resulting in a sequence similarity of repeating units that is greater within than among species (Dover 1982, Elder and Turner 1995). This cohesive evolution, resulting in intra-specific similarity and inter-specific divergence, can be explained through the "molecular drive" model (Dover et al. 1982, Garrido-Ramos 2017): new variants that appear by mutation of individual unit are expanded by unequal crossing-over, gene conversion, transposition or rolling-circle reinsertion of replicated extrachromosomal forms and subsequently fixed in the population and finally in the species (Garrido-Ramos 2017). The above-mentioned mechanisms lead also to sequence homogenization of unit repeats in the genome. Actually, once a mutation

appears in a unit, it can be spread and become predominant over the other variants or it is eliminated while other variants expand (Plohl et al. 2012).

Despite the high divergence of satellite DNA, closely related species often share satellite families. Actually, according to the "library hypothesis", related species may share an ancestral set of different satellite families which can be differently expanded in each species, fluctuating in copy number (Salser et al. 1976, Fry and Salser 1977). During the karyotype evolution events that mark the phylogeny of related species sharing a common set of satellite families, it has been demonstrated that satellite loci do not necessarily evolve orthologously. (Warburton et al. 1996, Schueler and Sullivan 2006).

Surprisingly, although centromeric satellites are highly divergent, they usually share unit repeat lengths that tend towards multiples of the nucleosomal repeat length (Henikoff et al. 2001, Cleveland et al. 2003). It was suggested that selection for nucleosomal length might sometimes constrain evolution of centromeric satellites, consistent with their structural role in the genome (Henikoff et al. 2001).

2.1. Repeat organization: the case of alpha satellite

Satellite sequences can be organized in stretches of single repeated monomers or can form higher-order-repeat (HOR) units, which are tandem arrays of larger units consisting of multiple basic repeat monomers. The formation of HOR has been reported for different eukaryotes, ranging from plants to metazoans (Navrátilová et al. 2008, Garrido-Ramos 2017).

One of the best studied family of satellite DNA is the alpha satellite (AS). AS was initially isolated from the genome of the African Green Monkey *Cercopithecus aethiops* and then demonstrated to be the major centromeric satellite family of simian primates (Maio 1971, Willard 1991, Sujiwattanarat et al. 2015). AS is made by tandemly repeated AT-rich monomers of about 170 bp, arranged in a head-to-tail fashion. In the human genome, there are two different organizations of AS: the monomeric pattern and the high-order-repeat pattern (Figure 1). The HOR pattern is found at the centromere core, made by 2 to 34 head-to-tail HORs that reiterate tandemly with 95-99% identity between copies (Willard and Waye 1987, Alexandrov et al. 2001, Garrido-Ramos 2017). The HOR arrays characterize the primary constriction of every chromosome with the exception of Y chromosome. The HOR centromeric core is flanked by the AS arranged in the unordered monomeric pattern. These pericentromeric monomers share 50-100%

sequence identity and are frequently interrupted by interspersed elements (Schueler and Sullivan 2006, Fukagawa and Earnshaw 2014, Garrido-Ramos 2017).

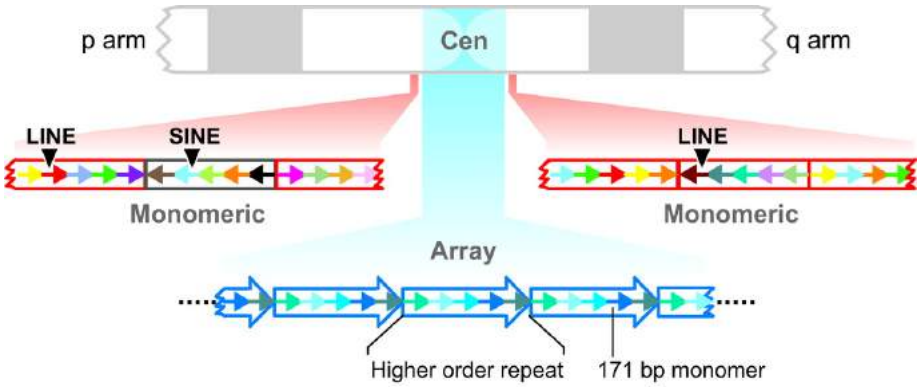


Figure 1. Organization of human centromeres. A typical human chromosome is schematically represented. Each small arrow represents a single satellite monomer. In the centromeric core (Cen, cyan), AS is organized in an HOR array. In the pericentromeric flanking regions (red), AS shows the monomeric unordered organization, frequently interrupted by interspersed elements (SINEs and LINEs) (from Schueler and Sullivan 2006).

This structure is due to the progressive proximal expansion which occurred during the evolution of the AS (Figure 2). Actually, primate centromeres were demonstrated to evolve by amplification of AS sequences in the inner core, which expands and moves the peripheral sequences sideways, forming layers of different age in the pericentromeric area (Shepelev et al. 2009). Thus, flanking monomeric satellite sequences represent the remnants of ancestral centromeres of primate progenitors (Alexandrov et al. 2001, Shepelev et al. 2009). It has been demonstrated that homogenization of satellite sequences is limited to the centromeric core while the ancestral units become more and more divergent. In particular, the homogeneization is intrachromosomal and different chromosomes evolved a specific type of HOR array (Alexandrov et al. 2001, Shepelev et al. 2009). Initially, this organization was not detected in primates other than apes, suggesting that the other primates carry only the monomeric organization of the AS at their centromeres (Alexandrov et al. 2001). However, recently, it has been demonstrated that this HOR organization of centromeric DNA is not limited to human and hominoids and it has been proposed that the establishment of high-order-repeat is a general event that can occur

occasionally or frequently in the centromeres of all simian primates (Sujiwattanarat et al. 2015, Cacheux et al. 2016).

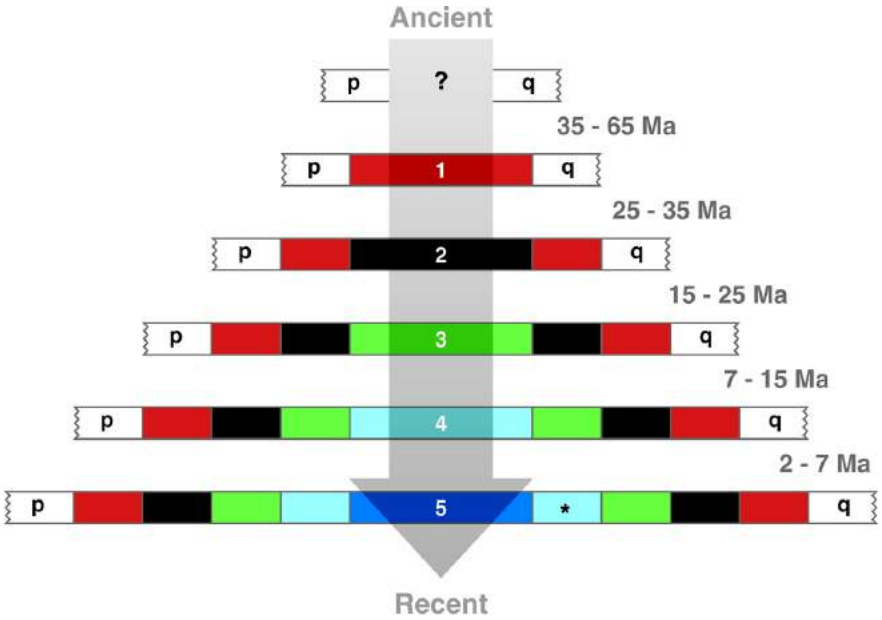


Figure 2. Progressive proximal expansion model for AS evolution. Successive additions (colored rectangles) to the centromere occurred during primate evolution. Each addition of new layers of AS moves previous centromeric DNA outward. An asterisk (∗) indicates the region of monomeric AS with functional centromere protein (CENP)-B boxes. The image is based on data from the human X chromosome and dates are derived from the accepted primate tree and from the phylogeny of L1 elements within AS sequences (from Schueler and Sullivan 2006).

2.2. Bridging telomeres and centromeres: the Chinese hamster example

It is well known that simple repetitive sequences, such as $(TTAGGG)_n$ in mammals and other vertebrate species, form the ends of linear chromosomes, namely the telomeres, and are essential for the preservation of chromosome integrity (Blackburn 1991). Interestingly, telomeric repeats can be present also at intrachromosomal sites, probably as the result of chromosomal fusions and fission events during karyotype evolution (Meyne et al. 1990, Nanda et al. 2002).

In several organisms, ranging from vertebrates to plants, intrachromosomal arrays of telomeric-like repetitions were expanded during

evolution and invaded the centromeric domain, becoming a considerable fraction of the centromeric satellite DNA (Bertoni et al. 1994, Nanda et al. 2002, He et al. 2013).

A remarkable example is considered by the rodent species, in which TTAGGG repetitions constitute the telomeres of all chromosomes but are also localized at most centromeres (Meyne et al. 1990). In particular, previous works from our laboratory demonstrated that in *Cricetulus griseus* (Chinese hamster, *2n* = 11) all the centromeres, with the exception of Y chromosome, comprise telomeric-like repetitions, detectable by FISH (Bertoni et al. 1996). In addition, TTAGGG repetitions are also found at several interstitial sites. A similar situation was detected also in a CHO-K1 derived cell line (CHO-PV), where 17 out of 19 chromosomes contain telomeric-like repeats at the primary constrictions (Bertoni et al. 1996).

2.3. Function of satellite DNA

The function of satellite DNA remains poorly understood in eukaryotes, although it is important to distinguish between centromeric and pericentromeric satellite sequences.

A large body of evidence demonstrated that both centromeric and pericentromeric satellite DNA is transcribed in eukaryotes from yeast to mammals and several roles of these transcripts in the epigenetic establishment of centromeric chromatin have been identified in a number of species (Rošić and Erhardt 2016). In particular, pericentromeric transcripts mainly contribute to the maintenance of the heterochromatin environment in which centromeres are embedded, while centromeric transcripts are involved in CENP-A loading and kinetochore assembly (Rošić and Erhardt 2016, McNulty et al. 2017).

In addition, although dispensable for centromere specification, centromeric satellite DNA might contribute to the stability of centromeres, as predicted from the "centromere drive" theory (Malik and Bayes 2006, Kursel and Malik 2018). This model is based on the asymmetry of female meiosis, in which only one of the four meiotic products is retained in the egg. Homologous chromosomes may compete for their inclusion in the egg via their "centromere strength", defined as the ability of their kinetochores for recruiting microtubules. Since its proposal, this "strength" was attributed to the presence of extended arrays of satellite DNA: the higher number of repeats were present, the stronger recruitment of centromeric protein would

occur (Malik and Bayes 2006, Fishman and Saunders 2008). Furthermore, this hypothesis was supported by the work of Iwata-Otsubo and collaborators, which demonstrates in mouse cell lines that "strong" and "weak" centromeres in meiosis differed according to the extension of centromeric satellite DNA arrays (Iwata-Otsubo et al. 2017, Kursel and Malik 2018). Nonetheless, excessive accumulation of repeated arrays by "selfish" centromeres may be harmful because of deleterious fitness consequences, directly as a result of expanded or mismatched centromeric strengths or indirectly due to the hitchhiking of deleterious alleles with driving centromeres (Kursel and Malik 2018). These deleterious effects were predicted to be counteracted by the co-evolution between centromeric proteins and centromeric DNA, which avoid an excessive expansion of the functional centromere. According to this model, the pericentromeric satellites become more and more degenerated and cannot be bound by centromeric proteins, since they have lost their transmission advantage and evolve neutrally (Malik and Bayes 2006).

However, the extreme abundance of pericentromeric satellite DNA compared to the centromeric one does not support the common view that pericentromeric satellites are simply selfish parasitic sequences remained as "fossils of centromere evolution" (Malik 2009, Jagannathan and Yamashita 2017). Actually, the maintenance of such extended arrays of satellite DNA would be a too high burden for the cell (Jagannathan and Yamashita 2017). To solve this controversy, Jagannathan and Yamashita recently proposed a structural role in the tridimensional nuclear organization for pericentromeric DNA, which could drive the formation of chromocenters in the nucleus (Jagannathan and Yamashita 2017, see Paragraph "The tridimensional nuclear architecture of centromeres").

3. NEOCENTROMERES

The centromere has so far escaped comprehensive molecular analysis due to its typical association with tandemly repeated DNA. It was clearly demonstrated that, although satellite DNA is usually associated to centromeres, it is not necessary for specifying centromeric function. Actually, functional satellite-free centromeres resulting from a centromerization event have been described (Voullaire 1993, Choo 2000, Amor and Choo 2002, Marshall et al. 2008, Piras et al. 2010, Purgato et al. 2015).

The term "centromerization" was coined by Choo to define the process of centromere formation in a chromosomal region. Centromerization

normally concerns the propagation of an existing centromere during replication. Rarely, this phenomenon occurs in regions which are normally non-centromeric. The ectopic centromere that appears occasionally in hitherto non-centromeric chromosomal regions is called "neocentromere" (Choo 2000, Amor and Choo 2002, Kalitsis and Choo 2012). Two different types of neocentromeres have been identified: clinical neocentromeres and evolutionary new centromeres. While clinical neocentromeres are sporadic cases that are not fixed in the population, evolutionary new centromeres are fixed in the species and represent an aspect of karyotype evolution.

Such neocentromeres must not be confused with the "classical" plant neocentromeres first described by Rhoades and Vilkomerson (Rhoades and Vilkomerson 1942). Actually, plant neocentromeres are accessory centromeres coexisting with the functional normal centromere, their activity is confined to meiosis and they do not form a typical kinetochore (Rhoades and Vilkomerson 1942, Amor and Choo 2002, Dawe and Hiatt 2004).

3.1. Human clinical neocentromeres

Since the discovery of the first neocentromere (Voullaire et al. 1993), more than 90 cases of human neocentromeres have been described (Marshall et al. 2008, Kalitsis and Choo 2012). Generally, neocentromerization is a rare rescue mechanism to avoid the loss of an acentric chromosomal fragment originating from a chromosomal rearrangement. The majority of human neocentric chromosomes derive from inverted duplications or interstitial deletions. Beyond neocentromere formation, these chromosomal rearrangements result in karyotype instability and are usually detrimental to the individual, explaining why human neocentromeres are unusual and not fixed in the population (Amor and Choo 2002, Marshall et al. 2008).

Human clinical neocentromeres are functional centromeres which are completely devoid of satellite DNA (Amor and Choo 2002, Marshall et al. 2008, Kalitsis and Choo 2012). They typically arise in gene-poor euchromatic regions, with the exception of a small number of cases located in the heterochromatic region of the long arm of chromosome Y. However, heterochromatic markers have been detected at neocentromeres emerged in euchromatic regions, suggesting that neocentromeres carry certain features of heterochromatin (Amor and Choo 2002, Kalitsis and Choo 2012). Despite the absence of sequence preference for neocentromere seeding, centromerization does not occur apparently at random sites along chromosomes. It has been

hypothesized that genomic "hotspots" for centromerization exist in certain region of the genome. These genomic locations may favor neocentromerization because of specific epigenetic hallmarks or the persistence of recombinogenic duplicons. Actually, it has been proposed that regions of the genome with a high content of duplications are predisposed to rearrangements, which then lead to neocentromere formation through epigenetic changes in the chromatin after DNA repair (Marshall et al. 2008).

Rarely, human neocentromeres arise in an intact chromosome with the pre-existing centromere still present, but inactivated. These neocentromerization events do not result from chromosomal rearrangements followed by centromerization. The active centromere has been repositioned leading to the formation of a pseudodicentric chromosome (Marshall et al. 2008). Few cases have been reported in literature but, considering that they do not cause clinical problems, they have been discovered serendipitously and they could be more common than the statistics indicate (Marshall et al. 2008). These pseudodicentric cases are very interesting because they could follow the mechanism of formation of evolutionary new centromeres during evolution.

3.2. Evolutionary new centromeres

Evolutionary new centromeres (ENCs) originate from centromere repositioning. Centromere repositioning is the movement of the centromere along the chromosome without marker order variation. This phenomenon was described for the first time by Montefalcone and collaborators in primates (Montefalcone et al. 1999). Since its discovery, it has become clear that centromere repositioning is an important mechanism of karyotype evolution ranked on equal ground with traditional chromosome rearrangements such as inversion, translocation, deletions and insertions (Rocchi et al. 2012).

The majority of evolutionary new centromeres so far discovered are associated to highly repetitive DNA (Montefalcone et al. 1999, Cardone et al. 2006, Rocchi et al. 2012). The first example of a satellite-less evolutionary neocentromere was described in the horse by our laboratory (Wade et al. 2009). Later, other examples of satellite-free evolutionary new centromeres have been reported in other mammalian species by our laboratory (Piras et al. 2010, Nergadze et al. 2018) and other groups (Shang et al. 2010, Locke et al. 2011, Tolomeo et al. 2017). These satellite-free neocentromeres are likely to represent an "immature" stage of centromerization, suggesting a possible

mechanism for centromere formation and maturation in higher eukaryotes (Wade et al. 2009, Piras et al. 2010).

In figure 3 (Piras et al. 2010) the current four-step model explaining neocentromere formation and maturation during evolution is depicted. The first step would consist in the shift of the centromeric function in a new position lacking satellite DNA, while the satellite DNA from the old centromere remains in its original position. A subsequent step would be the loss of leftover satellite sequences, relics of the old centromere. Finally, the new centromere could reach its maturity by acquiring satellite DNA as in the numerous evolutionary new centromeres described in several species (Piras et al. 2010). Since several evolutionary neocentromeres, fixed within species, are satellite-free, the accumulation of satellite sequences may simply be a neutral process driven by the presence of heterochromatin in the centromeric DNA (Piras et al. 2010). Recently, we found satellite-less centromeres comprising novel tandem repetitions, suggesting that arrays may represent the intermediate stage toward satellite DNA acquisition during evolution (Nergadze et al. 2018).

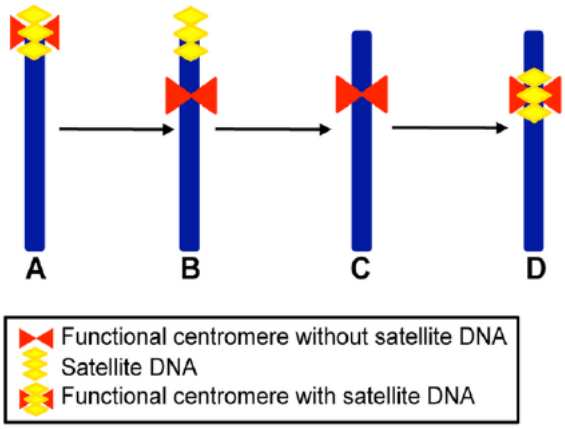


Figure 3. Schematic representation of the four-step model for neocentromere formation during evolution. (A) Acrocentric ancestral chromosome carrying satellite DNA (yellow) at its terminal centromere (red). (B) Submetacentric chromosome derived from centromere repositioning; this chromosome maintained satellite DNA sequences (yellow) at the terminal position, coinciding with the old centromere site, while the neocentromere (red) is devoid of repetitive sequences. (C) Submetacentric chromosome derived from (B) in which the terminal satellite sequences have been lost. (D) Submetacentric chromosome in its full "maturation" stage carrying satellite DNA (yellow) at the centromere (from Piras et al. 2010).

In the scenario depicted by this model, evolutionarily immature centromeres, lacking satellite DNA, might be expected to be found in rapidly evolving species (Piras et al. 2010). Works from our laboratory demonstrated that centromere repositioning played an important role in rapid karyotype evolution of the species belonging to the genus *Equus* (horses, asses and zebras) (Carbone et al. 2005, Piras et al. 2009, Piras et al. 2010). Moreover, in these species, several centromeres were proved to be completely satellite-free (Wade et al. 2009, Piras et al. 2010). Therefore, the rapidly evolving *Equus* species gave us the opportunity to catch snapshots of evolutionarily new centromeres in different stages of maturity (Piras et al. 2010, Purgato et al. 2015, Nergadze et al. 2018).

3.2.1. The centromeres of the genus *Equus*

The order Perissodactyla (odd-toed ungulate mammals) consists of Ceratomorpha and Hippomorpha suborders. The suborder Ceratomorpha includes Tapiridae (tapirs) and Rhinocerotidae (rhinoceroses) families, while the suborder Hippomorpha comprises only the Equidae (horses, asses and zebras) family. The Equidae family is now represented by the only extant genus *Equus*, encompassing eight species: two horses (*E. caballus* and *E. przewalskii*), two Asiatic asses (*E. hemionus onager* and *E. kiang*), one African ass (*E. asinus*) and three zebras (*E. grevyi*, *E. burchelli* and *E. zebra hartmannae*) (Piras et al. 2009) (Figure 4).

The karyotype of still extant species of Ceratomorpha is characterized by high chromosome number (*2n* ranging from 52 to 84) and mostly acrocentric elements. This arrangement is believed to correspond to the Perissodactyl ancestral karyotype (Trifonov et al. 2008, Piras et al. 2009). Actually, it was demonstrated that the rate of evolutionary rearrangements in the Ceratomorpha was extremely low (Trifonov et al. 2008). The scenario changed remarkably during the radiation of the genus *Equus*, emerged about 4-4.5 MYA according to recent phylogenetic studies (Orlando et al. 2013). Speciation events occurred very rapidly in the evolutionary time scale and were accompanied by extensive karyotype rearrangements. In fact, karyotypes of equid species are extremely variable in chromosome numbers (*2n* ranging from 32 to 66) and structure, with variable numbers of metacentric and submetacentric chromosomes (Musilova et al. 2007, Trifonov et al. 2008, Piras et al. 2009, Trifonov et al. 2012, Musilova et al. 2013).

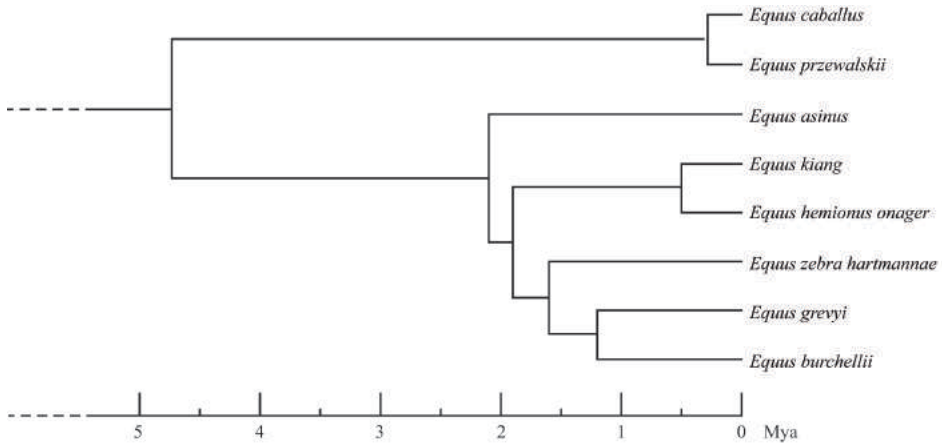


Figure 4. Phylogenetic tree of the genus *Equus*. Based on Trifonov et al. 2012 and Orlando et al. 2013.

Centromere repositioning occurred at surprisingly high frequency in the genus *Equus* and several evolutionary new centromeres were identified (Carbone et al. 2006, Piras et al. 2009).

In 2009, following the sequencing of the horse whole-genome, the centromere of chromosome 11 (ECA11) was proved to be completely devoid of satellite DNA. This was the first satellite-free evolutionary neocentromere, stably fixed within a species, to be discovered and characterized at the molecular level (Wade et al. 2009). Recent studies show that the position of ECA11 centromere is not fixed but slides within an about 500 kb gene desert regions among individuals, giving rise to different positional alleles, defined "epialleles" (Figure 5). This phenomenon is called "centromere sliding" and we recently proved that the positions of centromeric domains are inherited as Mendelian traits, but their position can slide in one generation being stable during mitotic propagation of cultured cells (Purgato et al. 2015; Nergadze et al. 2018). These data confirm the epigenetic nature of centromeres and prove that centromeric domains are characterized by positional instability (Purgato et al. 2015). The fact that CENP-A binding domains can move within relatively restricted regions suggests that the centromeric function is physically limited by epigenetic boundaries.

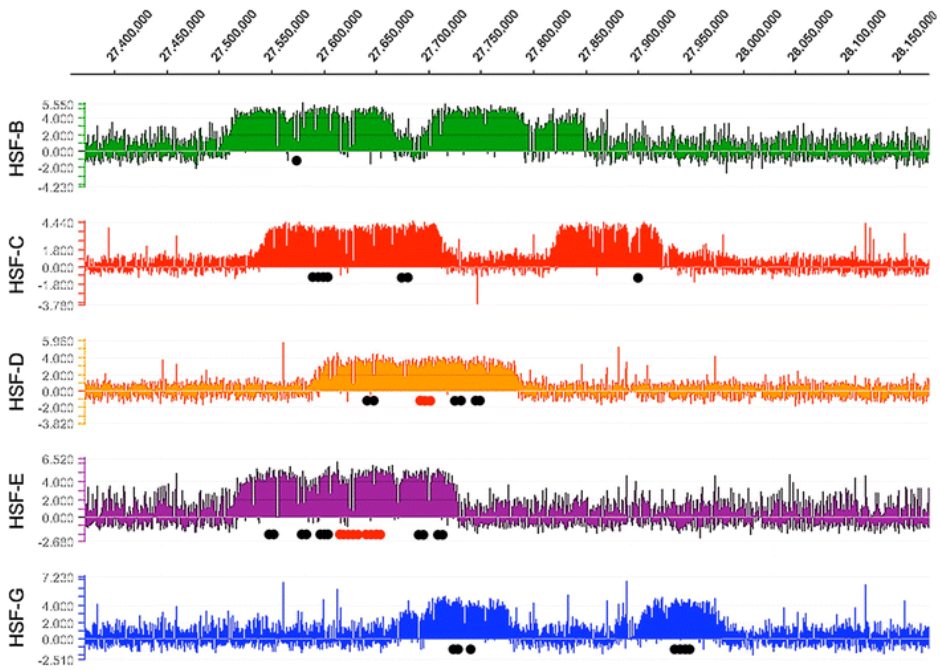


Figure 5. Variable position of the centromere of horse chromosome 11 among different individuals. DNA obtained by chromatin immunoprecipitation using an anti-CENP-A antibody, from five different horse fibroblast cultures, was hybridized to a tiling array covering the centromere region. y-axis, the log2 ratio of the hybridization signals obtained with immunoprecipitated DNA versus input DNA; x-axis, genomic coordinates on ECA11. Positions of informative SNPs are indicated as black dots (a single nucleotide of the SNP is enriched in immunoprecipitated DNA) and red dots (both SNP alleles are present in immunoprecipitated DNA). Adapted from Purgato et al. 2015.

Following the identification of ECA11 neocentromere, a cytogenetic analysis was carried out to investigate the distribution of satellite tandem repeats in *E. caballus*, *E. asinus*, *E. grevyi* and *E. burchelli* (Piras et al. 2010). In particular, the distribution of 37cen and 2PI (Anglana et al. 1996), the major satellite DNA families in the four analyzed *Equus* species, was analyzed by fluorescence *in situ* hybridization (FISH). In order to rule out the possibility that other satellite families could not be detected, hybridization with total genomic DNA was carried out as well. Figure 6 (Piras et al. 2010) shows the results of these hybridization experiments. In *E. caballus* the majority of centromeres contained both satellites, five chromosomes (1, 4, 5, 12 and X) showed only 37cen signals and chromosome 2 showed only 2PI

signal. As expected from previous molecular results, the centromere of ECA11 was the only one lacking any signal. The situation was quite different in *E. asinus*, *E. grevyi* and *E. burchelli*. Actually, several centromeres were devoid of satellite DNA at the cytogenetic level and satellite signals were detected at several non-centromeric termini, probably corresponding to relics of ancestral now inactive centromeres (Piras et al. 2010). In particular, nine previously identified evolutionary new centromeres, namely the centromeres of ECA11, EAS8, EAS9, EAS11, EAS13, EAS15, EAS16/EBU19, EAS18/EBU20 and EAS19, were proved to be satellite-free at the cytogenetic level (Carbone et al. 2006, Piras et al. 2009, Piras et al. 2010).

All these studies indicate that the genus *Equus*, providing both satellite-free and classical centromeres, is a unique model for the study of centromere function, organization and evolution.

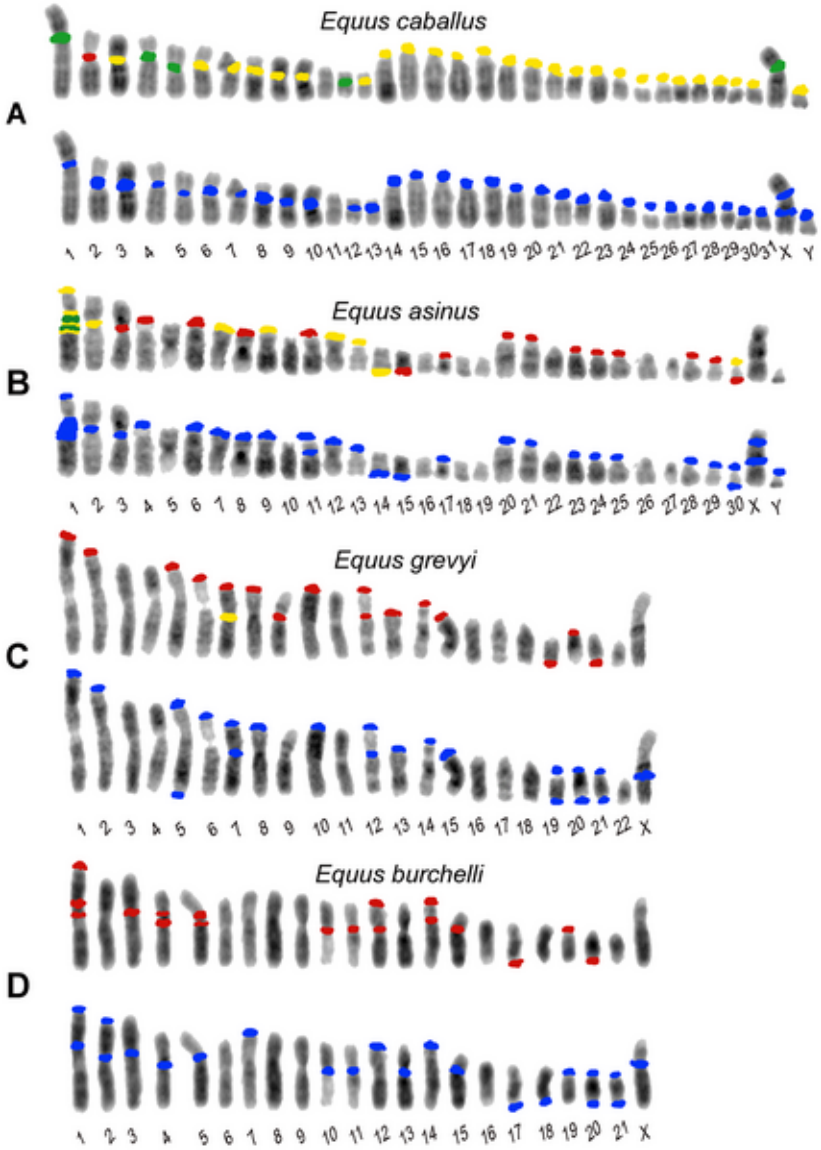


Figure 6. Schematic representation of satellite DNA distribution on metaphase chromosomes of Equids. Distribution of FISH signals on horse (A), donkey (B), Grevy's (C), and Burchell's (D) zebras chromosomes. Hybridization positive loci have been marked in different colors on banded karyotypes from each species: loci hybridizing with the 37cen probe only are labelled in green, 2PI positive loci are labelled in red and loci hybridizing with both 37cen and 2PI are labelled in yellow. Hybridization with genomic DNA probes, detecting total satellite DNA, is marked in blue (from Piras et al. 2010).

4. CENTROMERIC PROTEINS

It is well accepted that the centromeric DNA sequences are unable to specify centromeric function, resulting in high divergence of centromeric sequences and failure to detect common motifs. On the other hand, there are proteins that are specifically found only at centromeres and exhibit conservation among different taxa (Henikoff et al. 2001, Westermann and Schleiffer 2013).

Centromeric proteins can be generally classified as constitutive or transient proteins. Constitutive proteins reside at centromeres at all stages of the cell cycle. On the contrary, transient proteins associate with the centromere during specific stages of the cell cycle (Saxena et al. 2002, Przewloka and Glover 2009). Constitutive centromeric proteins are major candidates for maintaining the centromeric function (Henikoff et al. 2001).

Centromeric proteins display dynamic behaviour during the cell cycle. Actually, they serve as beacons that mark locations on chromosomes to which kinetochore proteins recruit. At a certain point, centromeres become kinetochores (Przewloka and Glover 2009). Indeed, centromeric proteins first acquire the ability to engage key kinetochore components that in turn attract proteins responsible of microtubule binding (Przewloka and Glover 2009). The overall kinetochore comprises an inner layer, which assembles over centromeric chromatin, a middle layer and an outer layer, which contacts spindle microtubules (Cleveland et al. 2003, Santaguida and Musacchio 2009).

The history of centromeric proteins started when rheumatologists identified patient sera that recognized the centromere regions of chromosomes giving the so-called “speckled nuclear” pattern (Moroi et al. 1980). Those patients had a scleroderma-related syndrome known as Calcinosis, Raynaud’s phenomenon, Esophageal dysmotility, Sclerodactyly and Telengiectasia (CREST) syndrome (Earnshaw 2015). Thus, the antibodies were termed CREST antibodies (Earnshaw 2015). Then it turned out that other patients with only Raynaud’s phenomenon had antibodies that recognized centromeres (Earnshaw 2015). This is the reason why now these antibodies are called only anti-centromere antibodies (ACA) (Earnshaw 2015).

By immunoblotting with ACA sera, three major antigens were identified and proved to be centromeric proteins. These were the first known centromeric proteins to be recognized. They were designated CENPs (CENtromere Proteins) and named CENP-A (the 17 kDa antigen), CENP-B

(the 80 kDa antigen) and CENP-C (the 140 kDa antigen) (Earnshaw and Rothfield 1985, Earnshaw 2015). The proteins were referred to as a "family" not just because they were all at centromeres: ACA recognized some epitopes shared by CENP-A and CENP-B and others shared by CENP-B and CENP-C (Earnshaw and Rothfield 1985, Earnshaw 2015). The nature of these mutual determinants is still unknown and may rely on yet-unidentified post-translational modifications, in view of lack of sequence similarity (Earnshaw 2015).

4.1. CENP-A and CENP-C

CENP-A is the centromere-specific variant of the histone H3. In all eukaryotes, CENP-A is the hallmark of functional centromeres, including satellite-free centromeres, but is absent from centromeres that are mutated or inactivated (Henikoff et al. 2001, Allshire and Karpen 2008). Moreover, CENP-A depletion results in mislocalization of most kinetochore proteins, explaining why *CENP-A* knock-out mice are not viable and show severe mitotic problems (Howman et al. 2000, Allshire and Karpen 2008). On the other hand, depletion of most kinetochore proteins has no effect on CENP-A localization. In addition, overexpression of CENP-A results in its mislocalization to normally non-centromeric regions and the formation of ectopic kinetochores (Allshire and Karpen 2008). Thus, some have argued that the ability to be bound by CENP-A is the epigenetic mark of centromere function (Henikoff et al. 2001, Allshire and Karpen 2008, Piras et al. 2010, Fachinetti et al. 2013).

Like all histones, CENP-A contains a globular histone fold domain, which is highly conserved among eukaryotes and similar to the core domain of histone H3, and a N-terminal tail, which is highly divergent among different species (Henikoff et al. 2001) (Figure 7). Centromere targeting of CENP-A is directed by the CENP-A targeting domain (CATD) located in the histone-fold region (Allshire and Karpen 2008, Musacchio and Santaguida 2009, Fukagawa and Earnshaw 2014). The CENP-A nucleosome core is rigid, but overall the DNA wraps less tightly than in conventional nucleosomes, suggesting that CENP-A chromatin may organize in a distinct conformation (Fukagawa and Earnshaw 2014). Conflicting models have been proposed for the structure of CENP-A containing nucleosomes, and a spirited ongoing debate concerns whether they are octameric or tetrameric (Fukagawa and Earnshaw 2014). However, although the debate is still active, emerging data

appear to support the existence of octameric CENP-A nucleosomes *in vivo* (Fukagawa and Earnshaw 2014).

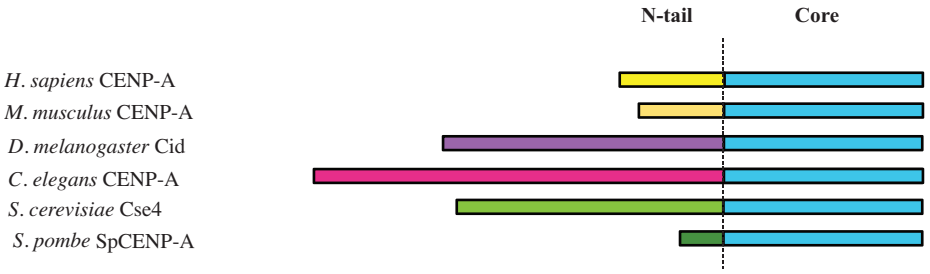


Figure 7. Schematic alignment of centromeric histones. Core domains (light blue) of centromeric histones (CENP-A and its homologs) are highly conserved among different species. On the contrary, N-terminal tails (indicated in different colors) are divergent even between related taxa. Adapted from Henikoff et al. 2001.

Regional centromeres contain blocks of CENP-A nucleosomes that are interspersed with blocks of canonical H3 nucleosomes (Allshire and Karpen 2008, Santaguida and Musacchio 2009, Fukagawa and Earnshaw 2014). The centromeric core domain and the pericentromere are characterized by a specific set of post-translational modifications. Histone H3-containing nucleosomes at the centromere core have marks that are specific for transcriptionally active chromatin, such as Lys4 and Lys36 methylation. In addition, CENP-A itself can be modified (Rošić and Erhardt 2016). On the contrary, the flanking pericentromeric domains are highly heterochromatic and characterized by trimethylation of Lys9 of histone H3 (H3K9me3) and are associated with HP1 protein (Rošić and Erhardt 2016).

CENP-A nucleosomes interact with a subset of the subunits of the constitutive centromere-associated network (CCAN). The CCAN is a group of proteins which associate with centromeric chromatin, providing a structural core to recruit outer kinetochore and inner centromere proteins (Hori et al. 2013) (Figure 8).

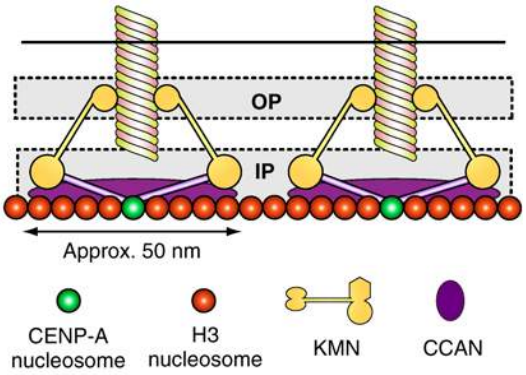


Figure 8. Schematic representation of recruitment of kinetochore components by centromeric chromatin.CENP-A nucleosomes anchor the kinetochore to the centromeric chromatin, forming a platform for sequential assembly of kinetochore components. Recruitment starts from the CCAN and continues with the KMN (Knl1–Mis12–Ndc80) network, which in turns contacts microtubules. Strong physical contacts between the inner plate (IP) and the outer plate (OP) of the kinetochore are required. Adapted from Santaguida and Musacchio 2009

Several models have been proposed for the geometric organization of centromeric chromatin, such as the looping model, the solenoid model and the boustrophedon model (Musacchio and Santaguida 2009, Ribeiro et al. 2010, Fukagawa and Earnshaw 2014) (Figure 9). According to the looping model and the solenoid model, the centromeric chromatin forms an amphipathic organization, with CENP-A nucleosomes on the exterior facing the kinetochore (Allshire and Karpen 2008, Musacchio and Santaguida 2009, Fukagawa and Earnshaw 2014). In the boustrophedon model, centromeric chromatin arranges in a sinusoidal wave in a series of layers, stacked on the top of each other (Ribeiro et al. 2010, Fukagawa and Earnshaw 2014).

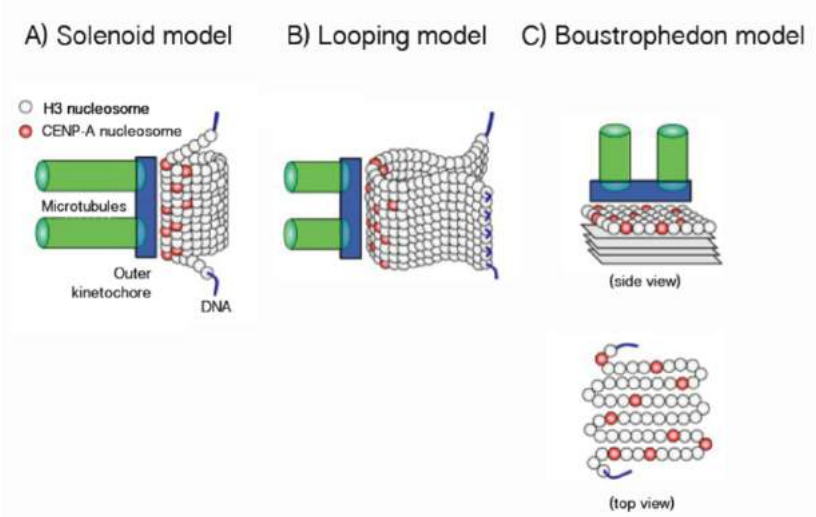


Figure 9. Current models for the spatial organization of centromeric chromatin. The three main models for describing the organization of centromeric chromatin are the solenoid model (A), the looping model (B) and the boustrophedon model (C). Adapted from Fukagawa and Earnshaw 2014.

CENP-C is a conserved essential inner kinetochore component and belongs to the CCAN (Saitoh et al. 1992). It works either downstream from CENP-A or in parallel with it in pathways of kinetochore assembly (Earnshaw 2015). CENP-C has been reported to have DNA binding activity, but appears to lack sequence specificity (Sugimoto et al. 1994). CENP-C binds to both CENP-A and to the factors that recruit CENP-A to chromatin (Earnshaw 2015). In addition, CENP-C interacts with CENP-B through two domains containing Mif2 homologous regions, which are also responsible for centromere localization (Suzuki et al. 2004).

CENP-C extends into the outer kinetochore, where it is responsible for tethering the KMN network, which in turns contact microtubules. Thus, CENP-C is an essential bridge between the inner and outer kinetochore (Earnshaw 2015).

4.2. CENP-B

CENP-B is a highly conserved centromeric protein which is primarily located inner chromatin region beneath the kinetochore plates (Cooke et al. 1990). CENP-B specifically binds to a 17 bp sequence known as the CENP-B box. Although other centromeric proteins are also DNA-binding proteins, CENP-B is the unique centromeric protein which exhibits unequivocal DNA sequence binding specificity (Fujita et al. 2015).

4.2.1. The CENP-B box

CENP-B was first demonstrated to specifically bind a 17 bp motif of the human alpha satellite, termed CENP-B box (Masumoto et al. 1989, Muro et al. 1992). Alphoid monomers with the CENP-B box were found in all the known alphoid subclasses except the one from the Y chromosome (Masumoto et al. 1989). Subsequent works demonstrated that only nine out of the 17 bp of the CENP-B box are essential for the recognition by CENP-B (Masumoto et al. 1993).

CENP-B boxes with those nine essential nucleotides were found in the centromeric satellite sequences of different species, such as the minor satellite in the house mouse *Mus musculus* (Masumoto et al. 1989), the 79 bp satellite in the Asian mouse *Mus caroli* (Kipling et al. 1995), a minor subtype of the alpha satellite in the African Green Monkey *Chlorocebus aethiops* (Yoda et al. 1996), in other several primate species (Haaf et al. 1995, Kugou et al. 2016) and several other mammalian species (Wu et al. 1995, Haaf and Ward 1995, Fantaccione et al. 1995). In addition, beyond the nine essential nucleotides, the CENP-B box of many species shows dyad symmetries consisting mainly of a palindromic sequence of 4 bp in which the two halves are separated by a 3 bp spacer (5’ cttCGTTggaAACGgga 3’; human CENP-B box with nucleotides of the palindromes in uppercase) (Stitou et al. 1999). The CENP-B box is the only common motif shared by these centromeric satellites, suggesting that CENP-B binding is a functionally important feature of mammalian centromeres (Kipling and Warburton 1997).

To our knowledge, the only exceptions reported to date have been found in the North African rodent *Leminscomys barbarus* and in the red-neck wallaby *Macropus rufogriseus* (Stitou et al. 1999, Bulazel et al. 2006). In the former case, the novel box is 19 bp long and conserves 12 of the 17 bp of the human one but only 5 of the 9 essential nucleotides for CENP-B binding and

the binding between this motif and the protein was not tested. Nonetheless, the dyad symmetries become extended from four to seven nucleotides (5' CtTAGTTTtggAAACTAtG 3'; nucleotides of the palindromes in uppercase). On the other hand, in the marsupial the novel box carries a GA dinucleotide instead of CG at positions 13-14, therefore affecting two essential nucleotides for CENP-B binding, although the protein was demonstrated to bind the box anyway (Bulazel et a. 2006). However, it has been reported that marsupials can carry amino acid substitutions in the CENP-B domain devoted to motif recognition (Master thesis by Demetrio Turati).

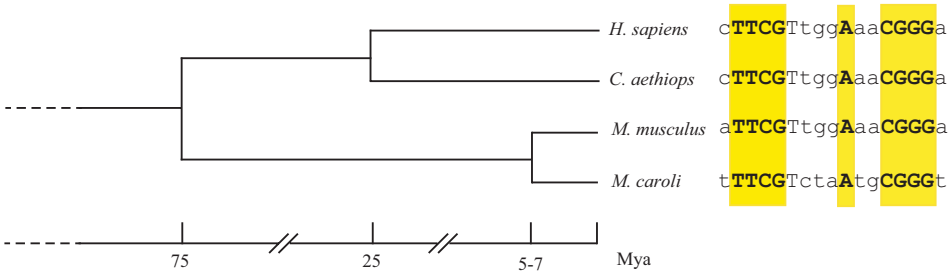


Figure 10. Alignment of identified CENP-B boxes in indicated mammalian species. The nucleotides that are essential for CENP-B binding are highlighted in yellow. Conserved bases are shown in upper-case letters. Adapted from Bulazel et al. 2006.

The CENP-B box contains two CpG dinucleotides and it is well known that, in eukaryotes, CpG methylation is an epigenetic DNA modification which is important for heterochromatin formation. Following the observation that demethylation of centromeric satellite DNA resulted in redistribution of CENP-B (Mitchell et al. 1996), it became clear that CpG methylation affects CENP-B binding. Actually, CpG methylation of the CENP-B box reduces the binding affinity between CENP-B and its binding site nearly to the level of nonspecific binding because of steric hindrance (Tanaka et al. 2005).

Recent data showed that CENP-B can be trimethylated at N-terminus and this α-N-trimethylation can enhance its binding to the CENP-B box (Dai et al. 2013). Since the methylation level increases after stress stimuli, such as high cell density, arsenite treatment and heat shock, it was proposed that cells may respond to these stresses by strengthening the interaction between CENP-B and centromeric DNA, which might play an important role in assembly, disassembly, and/or maintenance of centromere activity (Dai et al 2013).

4.2.2. CENP-B structure

CENP-B gene comprises a single exon and encodes a polypeptide of a molecular mass of about 65 kDa (Earnshaw et al. 1987, Sullivan and Glass 1991). CENP-B is a multidomain protein which is characterized by a DNA-binding region at the N-terminus, a dimerization domain at C-terminus, two acidic domains and, surprisingly, an endonuclease domain. This protein appears to be quite "fragile" since different works reported the formation of degradation products of CENP-B and mapped protease sensitive sites (Muro et al. 1992, Yoda et al. 1992, Tan et al. 2014).

Figure 11. Schematic representation of CENP-B domains. CENP-B domains are represented by boxes and lines. Lines represent flexible regions of the proteins. Major domains involved in CENP-B activity are the DNA binding domain (yellow), the endonuclease domain (grey), the acidic domains (red) and the dimerization domain (light blue). Adapted from Sullivan and Glass 1991 and Kitagawa et al. 1994.

The DNA binding region covers the first 129 amino acids of the N terminus and mediates the recognition of CENP-B box (Tanaka et al. 2001). The structure of the complex of the DNA binding region of CENP-B (CENP-$B_{1\text{-}129}$) and the CENP-B box was solved by X-rays crystallography by Tanaka and collaborators (Tanaka et al. 2001). CENP-$B_{1\text{-}129}$ is divided into four well-defined regions: the N-terminal arm, domain 1, the linker loop and domain 2, (Figure 11). Domains 1 and 2 have a helix-turn-helix motif and bind to adjacent major grooves of DNA. In the structure proposed by Tanaka and collaborators, the DNA binding region of CENP-B makes direct contacts with the nine essential nucleotides of the CENP-B box (Figure 12). Reflecting the conservation of the essential nucleotides of CENP-B box, the DNA binding regions is totally conserved between human, mouse and primates (Yoda et al. 1996). According to very recent data, the DNA binding domain of CENP-B specifically interacts with the CENP-A-H4 complex, but not with the H3-H4 complex, and CENP-B binding in the vicinity of CENP-A nucleosome substantially stabilizes the CENP-A nucleosome on alphoid DNA in human cells (Fuijta et al. 2015).

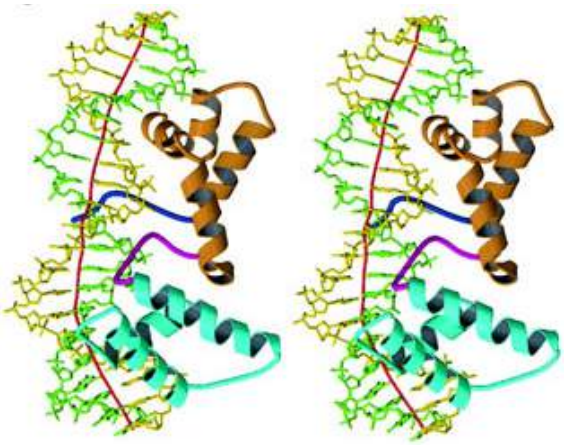


Figure 12. Crystal structure of CENP-B (1-129) N-terminal domain complexed with the CENP-B box. The N-terminal arm, domain 1, the linker loop and domain 2 are shown in blue, light brown, magenta and cyan, respectively. CENP-B binding causes DNA bending. Adapted from Tanaka et al. 2001.

Another extremely conserved region of CENP-B is the dimerization domain, which extends over the C-terminal region of 599 amino acid residues (Yoda et al. 1992, Kitagawa et al. 1995). Tawaramoto and collaborators determined the crystal structure of the dimerization domain (CENP-$B_{540\text{-}599}$) (Tawaramoto et al. 2003) (Figure 13). CENP-$B_{540\text{-}599}$ is composed of two amphipathic α helices, which are folded into an antiparallel configuration, and a flexible disordered C terminus, which is not involved in dimerization. The CENP-$B_{540\text{-}599}$ monomers dimerize to form an antiparallel, four helix bundle (Tawaramoto et al. 2003) (Figure 14). The CENP-B dimer was shown to be sufficiently stable to bundle together two CENP-B boxes distant up to 3.5 kb (Yoda et al. 1998). Since the CENP-B box sequence exists in every alpha satellite repeat (171 bp) of human centromeres, a model for DNA bundling by CENP-B dimer in the centromeric chromatin has been proposed: as shown in figure 14, CENP-B may accommodate a pair nucleosomes, between two CENP-B boxes tethered by the dimer (Yoda et al. 1998, Tawaramoto et al. 2003).

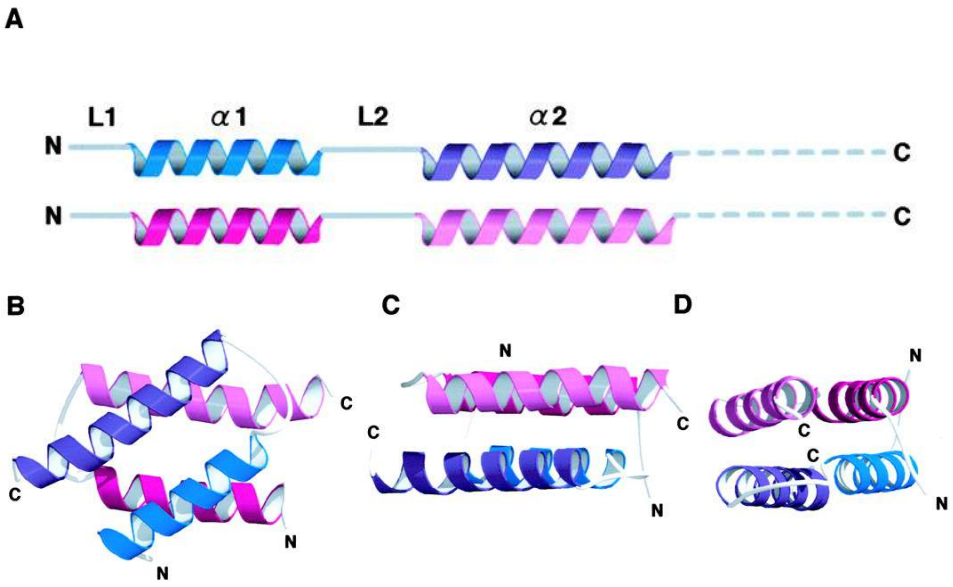


Figure 13. Crystal structure of CENP-B (540-599) dimerization domain. A) Secondary structure of CENP-B (540-599) monomers: each monomer is characterized by two α helices (α1 and α2), two loops (L1 and L2) and a disordered region (dashed line). B-D) Three views of CENP-B (540-599) structure during dimerization: two monomers dimerize to form an antiparallel four-helix bundle. Adapted from Tawaramoto et al. 2003.

The middle region of CENP-B is more variable among different species and comprise two extended clusters exceedingly rich in glutamic and aspartic acid residues (Earnshaw et al. 1987). These acidic domains are responsible for anomalous migration of CENP-B on SDS-PAGE. Indeed, the true molecular mass of CENP-B is about 65 kDa but migrates as a 80 kDa protein (Earnshaw et al. 1987). The first acidic cluster is responsible for interaction with CENP-C (Suzuki et al. 2004). This coding sequence of this domain is extremely rich in GAA and GAG codons, both coding for glutamate residues, which are often organized in short tandem repeats of the same triplet. It is worth reporting that a recent analysis on this first acidic domains in several tens of species belonging to seven mammalian orders revealed the existence of a great intra- and inter-order variability in the number and in the distribution of GAA and GAG stretches, resulting in variations in the overall number of glutamate residues (Master thesis by Demetrio Turati).

Interestingly, CENP-B is characterized also by an endonuclease domain which belongs to the DDE superfamily of endonucleases, so called because of the three conserved amino acids that are vital for functionality. These domains are found in transposable elements of the *pogo* superfamily, such as the human *Tigger* elements, which share an unexpected sequence similarity with CENP-B (Kipling and Warburton 1997). Notwithstanding, this domain in CENP-B is not likely to be still active and it is only a trace of evolutionary history (Marshall and Choo 2012).

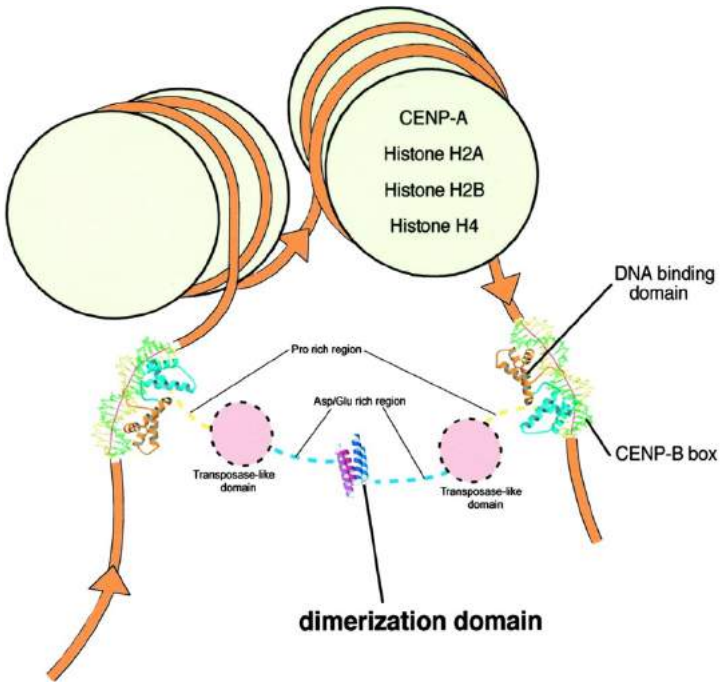


Figure 14. A model for DNA bundling by CENP-B dimer in centromeric chromatin. *Orange ribbons* with *arrowheads* indicate 171-base pair α-satellite repeats, which are wrapped into centromeric nucleosomes. A CENP-B dimer tethers together two CENP-B boxes, forming a loop containing two nucleosomes. Adapted from Tawaramoto et al. 2003.

4.2.3. Controversial role of CENP-B in the centromere

In spite of the fact that CENP-B was the first centromeric protein to be cloned (Earnshaw et al. 1987), little is known on CENP-B function and its exact role remains controversial (Earnshaw 2015). Actually, the CENP-B box

appears to be the only feature shared by centromeric satellites at the sequence level and both the CENP-B box and alphoid DNA sequence are required for *de novo* Mammalian Artificial Chromosome (MAC) formation and centromere assembly (Ohzeki et al. 2002, Okada et al. 2007), tempting speculation that CENP-B binding is functionally involved in centromere specification. On the other hand, active human neocentromeres and Y chromosomes from many species lack both CENP-B box and bound protein (Choo 2000, Amor and Choo 2002). Recently, lack of detectable levels of CENP-B binding have been reported also for autosomal centromeres of different primate species, confirming the lack of CENP-B boxes in subfamilies of alpha satellite DNA (Kugou et al. 2016, Suntronpong et al. 2016). Conversely, inactive centromeres of pseudodicentric chromosomes retain CENP-B, suggesting that its deposition is not sufficient for centromerization (Choo 2000). Moreover, CENP-B is not essential, since *CENP-B* knock-out mice are viable although they exhibit abnormal, lower body and testis weights for at least 10 weeks or uterine dysfunctions, suggesting an unknown possible role in the physiology of the reproductive tract (Hudson et al. 1998, Fowler et al. 2000).

These discrepancies led to the hypothesis that the protein could be functionally redundant or dispensable (Choo 2000). On one hand, it was suggested that *Tigger* elements could provide a functional redundancy for CENP-B and partially explain the lack of deleterious effects when CENP-B is absent (Kipling and Warburton 1997, Hudson et al. 1998, Casola et al. 2008). However, more recent studies state that CENP-B works alone without functionally redundant partners, since putative CENP-B paralogues are not present at mammalian centromeres (Marshall and Choo 2012).

A very recent study supported the theory of functional redundancy, suggesting that a key player of centromere specification might be the DNA secondary structure rather than its primary sequence. As shown in the work by Kasinathan and Henikoff (Kasinathan and Henikoff 2018), both centromeric satellites and satellite-less neocentromeres are predicted to adopt non-B-form conformations. In centromeric satellites harboring the CENP-B box, CENP-B mediates the DNA bending required for such conformation. On the other hand, centromeres lacking CENP-B binding sites are enriched in dyad symmetries which induce DNA to adopt the non-B-form. Therefore, the functional redundance of CENP-B is no more based on the presence of paralogues, but on sequence peculiarities of centromeric sequences which make up to the lacking of CENP-B binding sites.

Furthermore, Marshall and Choo proposed that, rather being dispensable for mitotic centromere function, the role of CENP-B at centromeres may be related to the theory of the CENP-A mediated meiotic drive (Marshall and Choo 2012). According to Marshall and Choo, accumulation of CENP-B box containing satellite sequences would increase CENP-A incorporation in the centromere through the presence of CENP-B and be selected via meiotic drive (Marshall and Choo 2012). This could explain why Y chromosomes are devoid of CENP-B boxes in all known mammal species (Marshall and Choo 2012). In this scenario, CENP-B is dispensable, because absence of CENP-B leads only to absence of meiotic drive, but at the same time it is conserved in mammals because of the same selection mechanism during meiosis (Marshall and Choo 2012). This hypothesis could also explain the accumulation of satellite sequences during the maturation of evolutionary new centromeres (Marshall and Choo 2012).

Recent data on human centromeres suggest that CENP-B stabilizes CENP-A and CENP-C maintenance at centromeres, increasing the centromere strength and fidelity of chromosome segregation (Fachinetti et al. 2015, Mohibi et al. 2015). In particular, according to the model proposed by Fachinetti and collaborators, CENP-C is recruited to centromeres by two parallel pathways: the CENP-A dependend pathway, based on the interaction between CENP-C and CENP-A carboxyl terminal tail, and the CENP-B dependent pathway, which relies on the binding between CENP-C and the first acidic domain of CENP-B. These two pathways are both required since the artificial depletion of CENP-B in human and mouse cells causes a 50% reduction of CENP-C at centromeres with subsequent increase of missegregation frequency, suggesting that CENP-B might be involved in the recruitment and in the retention of CENP-C at centromeres. (Fachinetti et al. 2015).

Interestingly, very recent data demonstrated that CENP-B mediates the SUMO-dependent recruitment of Daxx chaperon complex at centromeres (Morozov et al. 2017). This complex mediates the incorporation of H3.3, a histone variant which is involved in the maintenance of heterochromatin structure at telomeres, centromeres and pericentromeres (Jang et al. 2015). As direct consequence of its role in the recruitment of H3.3, CENP-B depletion causes a disruption of the H3K9me3 environment around centromeres, with subsequent erosion of pericentromeric heterochromatin and genome instability (Morozov et al. 2017). A contribution of CENP-B in heterochromatin formation has been highlighted also by the discovery that inactive pericentromeric arrays of alpha satellite from human centromeres are

actually transcribed and their long non coding transcripts still associated with CENP-B, participating in the formation of heterochromatin pericentromeric environment (McNulty et al. 2017).

5. THE TRIDIMENSIONAL NUCLEAR ARCHITECTURE OF CENTROMERES

A growing body of evidence suggests that the compartmentalization of the genome in the tridimensional nuclear architecture is essential for the modulation of genome expression and for the establishment and maintenance of cellular identity during differentiation (Solovei et al. 2016).

As reviewed in Solovei et al. 2016, in the tridimensional nuclear space, chromatin is subdivided into two main compartments, called A and B, which are spatially segregated in the nucleus. Within compartments chromatin is organized in spatial units, such as topologically associated domains (TADs), lamina-associated domains (LADs), nucleulus-associated domains (NADs) or pericentromere-associated domains (PADs). TADs are functional units which act as functional modules for physical interaction between regulatory elements, while LADs, NADs and PADs have structural roles in anchoring the genome within the nucleus (Solovei et al. 2016).

According to their functional role, A and B compartments closely correspond to active euchromatin (EC) and inactive heterochromatin (HC), respectively. The concepts of "euchromatin" and "heterochromatin" were introduced by Emil Heitz (Heitz 1928) upon the cytological observation that some chromosomal regions, termed "heterochromatin", remained condensed even during interphase while others, referred as "euchromatin" underwent postmitotic decondensation in the nucleus (Straub 2003). As proposed by Heitz, further studies revealed that this cytogenetic classification reflected functional features: euchromatin was demonstrated to be the active fraction of the genome, is gene-rich and replicate early in S-phase, while heterochromatin is transcriptionally inactive, gene-poor and replicate late in S-phase. In addition, euchromatin and heterochromatin are differentially marked by interspersed repetitive sequences, with the majority of SINEs residing in euchromatin while LINEs and LTR are mostly found in heterochromatin (Solovei et al. 2016). This functional separation is mirrored in the compartmentalization of chromatin in the nuclear architecture. Indeed, euchromatin occupies the nuclear interior, while heterochromatin is predominantly restricted to the periphery and nucleoli (Figure 15, reviewed

in Solovei et al. 2016). Since EC and HC domains are alternated along chromosomes, chromosomes are folded in order to weave between the two compartments. This nuclear organization is strongly conserved across taxa and cell types, with only few exceptions (Solovei et al. 2009, Solovei et al. 2016). During differentiation, a reshaping of the nuclear architecture occurs: the B inactive compartment expands, as result of progressive gene silencing, and the degree of segregation between euchromatin and heterochromatin increases.

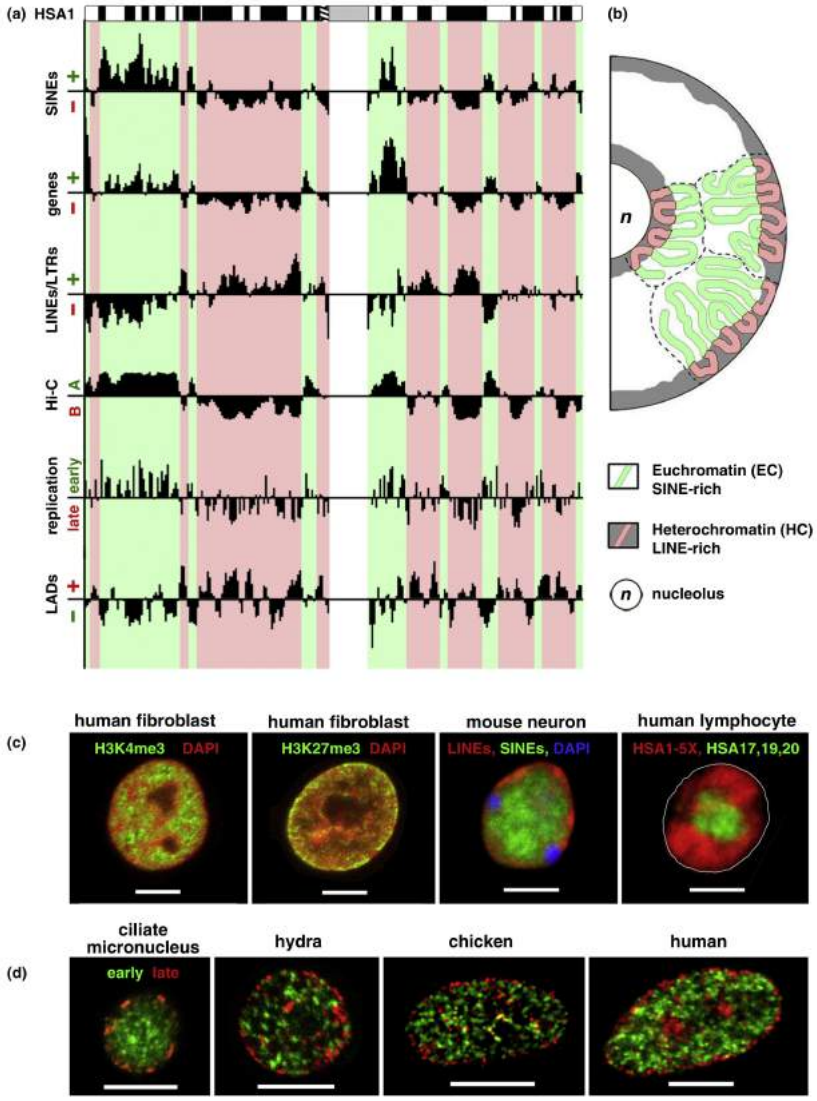


Figure 15. Euchromatic and heterochromatic chromosome regions and their spatial separation in the nucleus. a) Euchromatin (EC, green) and heterochromatin (HC, red) domains have complementary profiles regarding SINEs, genes, LINEs/LTR, A/B compartments, replication timing and LADs. The example of human chromosome 1 (HSA1) is reported. b) Compartmentalization of EC and HC in the nucleus. c) Examples of EC and HC localization in different mammalian nuclei revealed by active (H3K4me3) and inactive (H3K27me3) chromatin immunostaining, hybridization with probes for LINEs and SINEs, and visualization of gene-rich (HSA17, 19, 20) and less gene-rich (HSA1-5, X) chromosomes. d) Replication pattern of different compartments in different eukaryotes. (from Solovei et al. 2016)

In eukaryotes, centromeres and pericentromeric sequences usually aggregate in clusters in the tridimensional nuclear architecture and this phenomenon is called "centromere clustering". This clustering was observed in different cell types and taxa, ranging from lower eukaryotes, to plants, flies and mammals (Jones 1970, Nokkala and Puro 1976, Manuelidis 1984, Haaf and Schmid 1991, Funabiki et al. 1993, Jin et al. 1998, Weierich et al. 2003, Mayer et al. 2005, Fang and Spector 2005). Centromere clusters are cytologically visible as dense nuclear bodies, termed "chromocenters", which were firstly described at the beginning of the 20th century by Baccarini as foci strongly stained by nucleic acid dyes in plant nuclei (Baccarini 1908).

Chromocenters are found in the B compartment of heterochromatin, being characterized by DNA methylation and a panel of epigenetic modifications associated with chromatin compaction and transcriptional repression (Jagannathan and Yamashita 2017). In particular, in the mammalian nuclear architecture, centromere clusters are preferentially positioned at the nuclear periphery or around nucleoli (Weierich et al. 2003, Solovei et al. 2016). For example, in human and mouse lymphocyte nuclei, in human monocytes and human fibroblasts most centromere clusters are found at the nuclear periphery and centromeres were located at the periphery of the respective chromosome territory (Weierich et al. 2003, Solovei et al. 2004a). Nonetheless, differences in their tridimensional arrangements have been reported, depending on cell type and cell cycle stage (Figure 16). For instance, in mouse olfactory neurons, pericentromeric foci coalesce in centrally located foci (Clowney et al. 2012, Solovei et al. 2016). Similarly, following post-mitotic chromatin reorganization in mouse Purkinje cells in the cerebellum, nucleoli and chromocenters initially move inwards and fuse. In later stages, some centromeric clusters dissociate and return to the nuclear periphery (Manuelidis 1984, Solovei et al. 2004b, Solovei et al. 2016). During rod cell differentiation in nocturnal mammals, an inversion of euchromatin and heterochromatin nuclear positions occurs and thereby chromocenters coalesce in the nuclear interior (Solovei et al. 2009, Solovei et al. 2016). In addition, clustering of the kinetochore regions and their nuclear position change with the cell cycle stage in the same way in different cell types: a large fraction of centromeres is in the nuclear interior during early G1 without clustering, in late G1 and early S centromeres shift to the nuclear periphery and form clusters and the highest degree of centromere clustering is found in G0 cells (Solovei et al. 2004a).

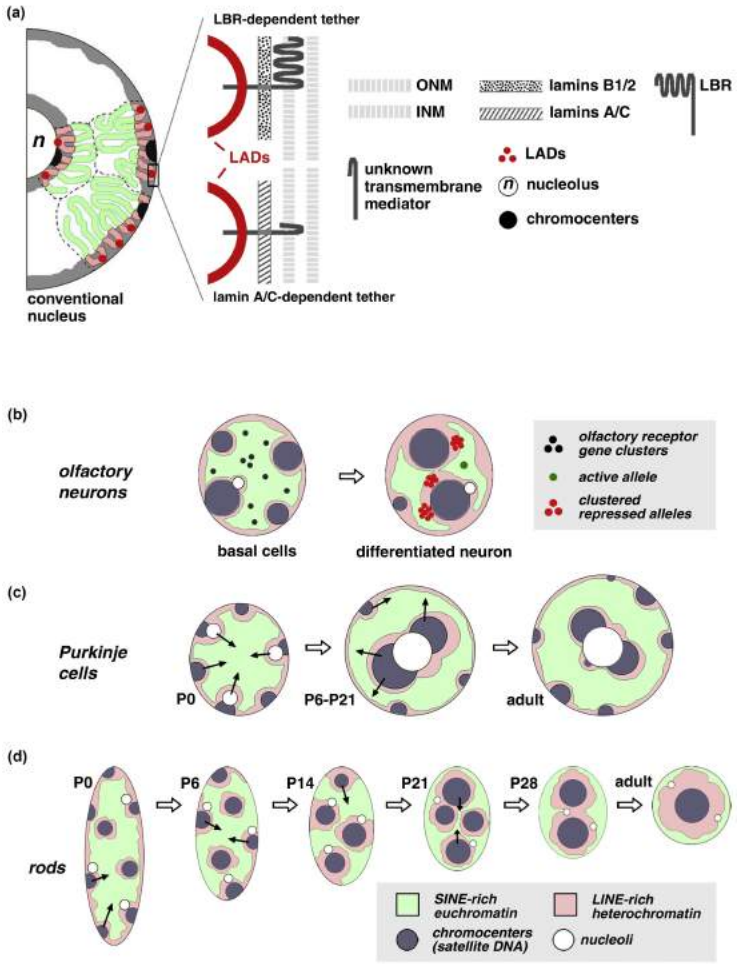


Figure 16. Position of chromocenters in different cell types. a) Conventional nucleus with peripheral HC and internal EC. Chromocenters localize at the periphery or around nucleulus. At least two major tethers of peripheral HC (LADs) binding are identified, the LBR (Lamin B receptor)-dependent and LA/C (lamina-costituent Lamin A/C)-dependent tethers. b) Repositioning and silencing of olfactory receptor in olfactory neurons during their differentiation. During this nuclear reshaping, chromocenters coalesce and repressed alleles relocate and cluster in large foci positioned in the HC zone around chromocenters. (c) Postmitotic reorganization of nucleoli and chromocenters in nuclei of Purkinje cells in the cerebellum (d) Nuclear inversion in rod cells of nocturnal mammals, as reported in Solovei et al. 2009. Rod photoreceptors stop proliferation at P5–P6 when they still have a conventional nuclear organization with HC adjacent to the nuclear periphery. In postmitotic rods, chromocenters and HC fuse in the nuclear interior, while EC relocates to a thin peripheral shell. (Adapted from Solovei et al. 2016)

In higher eukaryotes, the role of chromocenters, comprising both centromeric and the extendend blocks of "junk" pericentromeric satellite DNA, remains controversial (Jagannathan and Yamashita 2017). A possible explanation is given by the observation that the disruption of chromocenters resulted in a dramatic increase in micronuclei formation, thereby resulting in accumulation of DNA damages, chromosomal breaks and damages to the nuclear envelope integrity (Jagannathan and Yamashita 2017). Thus, chromocenters have been proposed to play a critical role in encapsulating the full genome in a single nucleus, acting as anchors in genome compartmentalization (Jagannathan and Yamashita 2017).

Despite the conservation of centromere clustering across eukaryotes and its proposed role in the tridimensional nuclear architecture, it is matter of debate whether this phenomenon relies on the presence of satellite DNA or on the epigenetically-defined centromeric function. It has been proposed that similarly typed sequences exhibit a high affinity to each other, thus driving the separation of compartments within the nucleus (Solovei et al. 2016). In particular, highly repetitive sequence might self-associate, acting as dominant seeds to promote segregation of heterochromatin and euchromatin within the nucleus (Krijger and de Laat 2013, Solovei et al. 2016). This mutual attraction could be tentatively attributed to chromocenter bundling proteins, such as HMGA1 identified in mouse and D1 in *Drosophila*, which can cross-link DNA molecules on multiple chromosomes and promote chromocenter formation (Jagannathan and Yamashita 2017).

On the other hand, recent studies in *Candida albicans* suggested that the basis of centromere clustering depends on epigenetically defined function and not on the primary DNA sequence (Burrack et al. 2016). *Candida albicans* is an organism with epigenetically-inherited centromeres which are made by about 3 kb central core sequences lacking any common motif, flanked by inverted repeats, short tandem repeats and transposon-associated repeats (Sanyal et al. 2004, Burrack et al. 2016). Despite the lack of satellite DNA characteristic of higher eukaryotes, these centromeres form clusters in the nucleus. Taking advantage of strains carrying 20 different neocentromeres, Burrack and collaborators demonstrated that, surprisingly, neocentromeres clustered with active native centromeres. On the contrary, in the wild type strain, or rather prior to neocentromere formation, region where neocentromeres could form did not exhibit strong interactions with other centromeres (Burrack et al. 2016), demonstrating that in this organism the basis of clustering is epigenetic.

In addition, the presence of mammalian species with centromeres lacking satellite DNA, such as the ones of the genus *Equus*, raises the question whether these satellite-free centromeres might participate in chromocenters and how satellite-free chromosomes are correctly partitioned in the nucleus.

6. THE "CENTROMERE EFFECT" ON MEIOTIC RECOMBINATION

The meiosis is a specialized form of cell division of a diploid cell in which a single round of DNA replication is followed by two nuclear divisions. During the first one, called Meiosis I, the homologous chromosomes pair, recombine and then segregate to opposite poles. In the second division, named Meiosis II, sister chromatids are separated and the overall process ends with the formation of four nuclei, each containing a complete haploid set of chromosomes.

A key step of meiosis is recombination between homologous chromosomes, which occurs in Prophase I and has the dual role of driving immediate chromosome segregation and generating novel variants for long term evolution. Prophase I is divided in four different phases, namely leptonema, zygonema, pachynema and diplotema (Figure 17, Baudat et al. 2013).

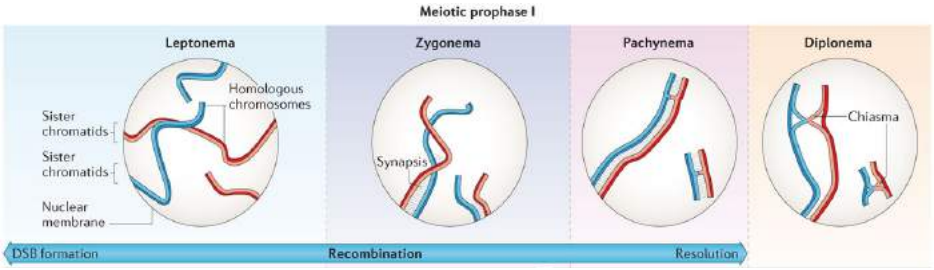


Figure 17. The four phases of meiotic Prophase I. The chromosomes begin to condense and meiotic recombination starts with the formation of double-strand breaks (DSBs) during leptonema. Synapsis between homologs starts at zygonema and the pairing is complete in pachynema. DSBs are progressively resolved by recombination and cross-overs between homologs could be visualized as chiasmata during diplonema. Adapted from Baudat et al. 2013.

During zygonema, homologous chromosome pair and form a bivalent structure and this process is coupled with the establishment of the synaptonemal complex (SC) (Baudat et al. 2013, Da Ines and White 2015).

The synaptonemal complex is a zipper-like protein structure which forms between pairs of homologous chromosomes (Figure 18). It is highly evolutionary conserved, being a key structure of meiosis from yeast to mammals (Zickler and Kleckner 1999, Gao and Colaiácovo 2018). The synaptonemal complex is composed by a central element and two rod-like lateral elements, joined together by transverse filaments (Syrjänen et al. 2014, Fraune et al. 2016).

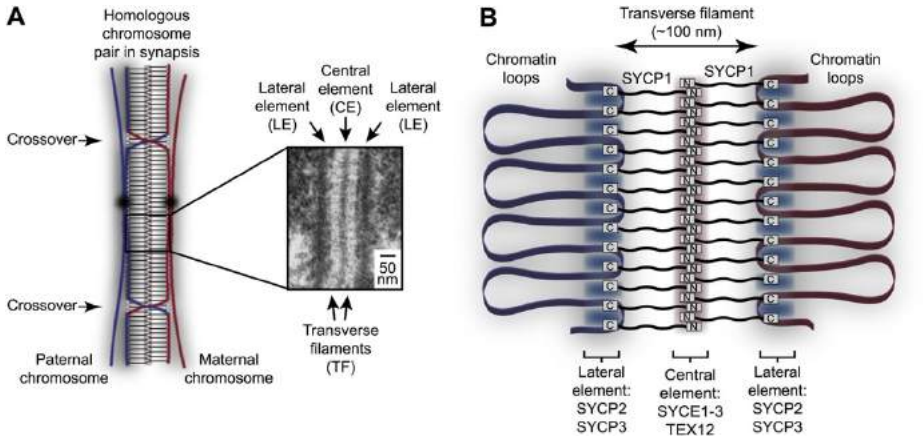


Figure 18. The mammalian synaptonemal complex. (A) Tripartite ultrastructural appearance in which transverse filaments bridge between a midline central element and lateral elements that coat the chromosome axes. Electron micrograph from Kouznetsova et al. 2011. (B) Model for assembly of the mammalian SC from its key components. SYCP1 forms the transverse filaments, with its N- and C-terminal regions located in the central and lateral elements respectively. The central element also contains SYCE1, SYCE2, SYCE3 and TEX12, while the lateral elements contain SYCP2 and SYCP3. Adapted from Syrjänen et al. 2014.

Homolog pairing is essential for meiotic recombination, which solves the double-strand breaks (DSBs) induced in leptonema. In particular, DSBs can be resolved by either crossovers (COs) or non-crossover mechanisms. Cross-overs consist of reciprocal exchanges between homologs, while non-crossovers involve only a unidirectional transfer of genetic information over short intervals and therefore have only a limited, local effect on genetic diversity (Baudat et al. 2013).

Since only a small fraction of DSBs are processed as COs, a highly regulated genetic control determines both CO homeostasis and their distribution along chromosomes (Capilla et al. 2016). In particular, it is well known that at least one CO is required per chromosome pair (Jones and Franklin 2006). Moreover, the presence of a CO influences the positioning of

other CO events on the same chromosome, in a phenomenon called CO interference (Jones and Franklin 2006).

A peculiar region for the regulation of CO positioning is the centromere. As a matter of fact, the centromere exerts a direct, negative effect on meiotic recombination, both within itself and on proximal regions (Beadle 1932, Mather 1938, Choo 1998). This effect is termed the "centromere effect" and it is conserved across eukaryotes (Choo 1998). Crossover suppression ranges from 5-fold to >200-fold in different organisms (Talbert and Henikoff 2010). The precise mechanisms responsible for this phenomenon are still controversial, though it has been hypothesized that selective pressure to reduce crossing over near the centromere would be strong. Indeed, recombination events too close to the centromere may disrupt pericentric sister chromatid cohesion, having dramatic impact on kinetochore functionality (Talbert and Henikoff 2010).

Aims of the work

Although dispensable for centromeric function, satellite DNA has been proposed to contribute to the stability of centromeres and their clustering in the tridimensional nuclear architecture. Despite their high divergence, centromeric satellites share a common motif, the CENP-B box, which is recognized by the centromeric protein CENP-B, the only known centromeric protein that exhibits DNA binding specificity. In spite of the high conservation of CENP-B protein and its binding motif across mammals, this protein appears dispensable for the centromeric function, giving rise to the so-called "CENP-B paradox". Thus, the interconnected roles of both satellite DNA and CENP-B are still an open issue.

Even though tandemly repeated DNA is a common feature of mammalian centromeres, in *Equus* species satellite DNA is uncoupled from centromeric function: beyond classical satellite-based centromeres, several centromeres are satellite-free and many satellite DNA loci are not centromeric. Thus, equid species represent a powerful model system for the study of the epigenetic establishment of the centromeric function.

Previous work from our laboratory demonstrated that also the rodent *Cricetulus griseus* carries peculiar centromeres, characterized by large clusters of telomeric-like repeats as well as chromosome-specific families of satellite DNA, suggesting a particular organization of centromeric satellite in this species.

The aims of this work were:

1. to characterize satellite-based centromeres of *Equus caballus*, identifying the major centromeric satellite DNA sequence;
2. to verify the presence of satellite-free centromeres in *Equus asinus*, to analyze their DNA sequence organization, positional stability and transmission;
3. to investigate the role of the CENP-B protein in the epigenetic establishment of centromeric chromatin in the genus *Equus*, focusing on the analysis of CENP-B gene and protein product, the study of CENP-B distribution with respect to other centromeric proteins and the identification of its binding sites in *Equus caballus*, *Equus asinus*, *Equus grevyi* and *Equus burchelli*;
4. to characterize the CENP-B binding pattern in *Cricetulus griseus*, focusing on the identification and the distribution of its binding sites;

5. to test whether the basis of centromere clustering depends on the primary DNA sequence (satellite DNA) or on the centromeric function, taking advantage of the presence of satellite-less and satellite-based centromeres in *Equus caballus* and *Equus asinus*;
6. to test whether the "centromere effect" on meiotic recombination is related to presence of satellite DNA or to the centromeric function. To investigate the phenomenon of "double-spotted" centromeres at meiosis.

Materials and Methods

1. CELL CULTURE

Horse (*Equus caballus*), donkey (*Equus asinus*), hinny (*Equus burdo*, hybrid of stallion and jenny), mule (*Equus mulus*, hybrid of jack and mare) and human (*Homo sapiens*) fibroblasts were isolated and established from skin biopsies under sterilized conditions. Grevy's zebra (*Equus grevyi*) primary fibroblasts were purchased from Coriell Repositories. Burchell's zebra (*Equus burchelli*) primary fibroblasts were kindly provided by Prof. Mariano Rocchi from the University of Bari (Bari, Italy). As far as Part 2 is concerned, primary fibroblast cell lines from DonkeyB, HorseA, HorseC, Horse D and Hinny were established from skin or testis biopsies of adult animals from Cornell University and kindly provided by Professor Douglas F. Antczak. HorseD fibroblasts were obtained from testicular tissue of a freshly castrated. MuleA, MuleB, and MuleC cell lines were derived from three mule conceptuses from normal pregnancies recovered on days 32–34 after ovulation via uterine lavage, as described (Adams and Antczak 2001). Immortalization of the MuleA fibroblast cell line was carried out as described in Vidale et al. (2012).

Primary fibroblasts were cultured in high-glucose DMEM (EuroClone) medium supplemented with: 20% fetal calf serum (EuroClone or Biowest), 2x NEAA (non-essential amino acids, EuroClone or Biowest), 2mM L-glutamine (SIGMA), 1x penicillin/streptomycin (SIGMA). Mule immortalized fibroblasts were grown in the same medium of primary fibroblasts supplemented with G418 sulphate (Invivogen) at the final concentration of 0.4 mg/ml.

HeLa cells were cultured high-glucose DMEM (EuroClone) medium supplemented with: 10% fetal calf serum (EuroClone), 1x NEAA (non-essential amino acids, EuroClone), 2mM L-glutamine (SIGMA), 1x penicillin/streptomycin (SIGMA).

Cells were maintained in a humidified atmosphere of 5% CO_2 at 37°C.

2. DNA EXTRACTION

Whole genomic DNA from horse, donkey, Grevy's zebra and Burchell's zebra fibroblasts was extracted according to standard procedures

(Sambrook and Maniatis, 1989). The equine BAC clones of interest, 37cen and CENPB-sat DNA probe were extracted from 10 ml of bacteria cultures with the Quantum Prep Plasmid miniprep kit (BioRad), according to supplier instructions.

3. PCR AND SEQUENCING

The details of the PCR amplification and sequencing of donkey satellite-less centromeres, reported in Part 2, can be found in the attached paper.

The primer pairs listed in Table 1 were used for amplification of the DNA segment of the coding sequence of CENP-B gene. The PCR was carried out in a 25 µl-final volume with 50-200 ng of genomic DNA, 20 pmol of each primer, 0.2 mM dNTPs (Fermentas), 1x Colourless Buffer (Promega) and 0.4 units of GoTaq® DNA polymerase (Promega). We used this thermal profile: 95°C for 2 min followed by 35 cycles at 95°C for 40 s, appropriate annealing temperature for 40 s and 72°C for appropriate extension time and a final extension cycle at 72°C for 10 min. The PCR products were electrophoresed through a 1% agarose gel.

PCR products were either treated with ExoI and FastAP (Thermo Scientific) or gel extracted using PCR Clean and Gel extraction kit (Promega) according to the manufacturer's protocol and then TA cloned using pGEM-T Easy vector (Promega). Sanger-sequencing was carried by the BMR Genomics company or by the GATC Biotech company.

F1	aagaattcgccaccATGGGCCCCAAGCGGCGGCAGCTGACGTTCC
F2	AGGATGGGCCCCAAGCGGCGGCAGCTGACG
F3	GTCAAGGGCATCATCCTCAAG
F4	TGCTTTCGTGAGGCTGGCTT
R1	aaggatccttGCTTTGATGTCCAAGACCCCGAACT
R2	CACGCCAGCCGGTCGTACTC
R3	GAGGGCAGTGGTGATAGTGG
R4	aaggatccttGCTTTGATGTCCAAGACCCCGAACT

Table 1. Sequence of the primers used to amplify and sequence CENP-B CDS. Tails containing restriction sites are written in lowercase.

4. ANTIBODIES

The ChIP-seq experiments described in Parts 1 and 2 were performed using a polyclonal antibody against human CENPA protein (Wade et al. 2009, kindly provided by Prof. Mariano Rocchi, Università di Bari) or a human CREST serum whose CENP-A specificity was previously demonstrated (Purgato et al. 2015).

The immunofluorescence experiments described in Parts 3, 4, 5 and 6 were performed with the following antibodies: anti-CENP-B ab84489 (Abcam), anti-CENP-B sc-22788 (Santa Cruz Biotechnology Inc.), anti-CENP-B H00001059-B01P (Abnova), anti-CENP-A sheep serum against the horse protein CENP-A (kindly provided by Professor Kevin F. Sullivan from NUI Galway, Galway, Ireland), the human CREST serum previously described, a rabbit anti-CENP-C polyclonal antibody (Wade et al. 2009), anti-B23 antibody (B0556, Sigma) anti-SCP3 antibody (Abcam, ab15093) and anti-MLH1 antibody (BD Pharmingen, 551091). ChIP-seq experiments described in Parts 3 and 4 were performed with anti-CENP-B sc-22788 (Santa Cruz Biotechnology Inc.), the anti-CENP-A sheep serum previously described, the polyclonal antibody against human CENPA protein (Wade et al. 2009) and the human CREST serum previously described (kindly provided by Dr. Claudi Alpini from "Fondazione I.R.C.S.S. - Policlinico San Matteo).

5. WHOLE PROTEIN EXTRACT PREPARATION AND WESTERN BLOTTING

To obtain total protein extracts for Western Blotting, about 3 million cells were washed twice with cold PBS and resuspended in a lysis buffer (50 mM Tris-HCl pH 6.8, 86 mM β-mercaptoethanol, 2% SDS) and boiled for 10 minutes. Proteins were then separated by SDS-PAGE on a 7.5% polyacrilamide gel and blotted to nitrocellulose membranes (Amersham™ Hybond™-ECL, GE-Healthcare) according to standard methods. The proteins on the filter were blocked incubating the membranes with 7.5% Skim Milk in PBST at 4°C for 8 hours on the rocket. The anti-CENP-B sc-22788 antibody, diluted 1:750 in 7.5% Skim Milk in PBST was incubated at 4°C for 15 hours on the rocket, followed by three 10 minutes washes with PBST at 4°C. The secondary antibody (HRP conjugated anti-rabbit for sc-22788), diluted 1:5000 in the blocking solution, was incubated for 1 hour at 4°C, followed by three 10 minutes washes with PBST at 4°C. To detect the protein

labeled to the antibody we used the BIO-RAD Clarity™ Western ECL Substrate kit following manufacturer's protocol. The exposition, from few seconds to 3 minutes, was performed using chemioluminescence films (GE Healthcare, Amersham Hyperfilm ECC), which were then developed and fixed. Subsequently, the membrane was washed in PBST at 4°C for 15 hours, on the rocket, for the immunodetection of α tubulin, used as loading control. The membrane was blocked with 7.5% Skim Milk in PBST at 4°C for 2,5 hours on the rocket. The anti- α tubulin antibody [DM1A] ab7291 (Abcam), diluted 1:5000 in 7.5% Skim Milk in PBST was incubated at 4°C for 1 hours on the rocket, followed by three 10 minutes washes with PBST at 4°C. Secondary antibody incubation and detection was performed as previously described for CENP-B immunodetection.

6. IMMUNOFLUORESCENCE

Primary fibroblasts were harvested, washed once with phosphate-buffered saline and re-suspended at a concentration of $4x10^4$ cells/mL in 0.075M KCl for 20 minutes at 37°C. The cell suspension was then supplemented with 25mM sucrose for 20 minutes at room temperature. 100 μl of cell suspension were cyto-spun (BHG Hermle Z380) onto slides at 1250 rpm for 8 minutes. Slides were fixed in cold methanol for 4 minutes on ice and then incubated in 1x PBS supplemented with 0.05% Tween-20 (PBST). Incubation with the anti-CENP-B ab84489 and sc-22788, diluted 1:80 in PBST, was performed for 2.5 hours at 37°C. Incubation with the anti-CENP-C polyclonal, diluted 1:100 in PBST, was performed for 1 hour at 37°C. Incubation with the anti-CENP-A polyclonal or human CREST serum, diluted 1:250 in PBST, was performed for 1 hour at 37°C. Slides were then washed three times for 5 minutes in PBST at room temperature. Secondary antibodies (FITC conjugated anti-rabbit, Alexa488 conjugated anti-mouse, Texas Red conjugated anti-mouse, Alexa488 conjugated anti-human, Alexa488 conjugated anti-sheep and rhodamine conjugated anti-sheep), diluted 1:100 in PBST, were added and incubated for 1 hour at 37°C. After two washes in PBST at room temperature, chromosomes were counterstained with DAPI (0.2 μg/ml) and mounted with Fluorescence Mounting Medium (Dako). Digital grey-scale images for fluorescence signals were acquired with a fluorescence microscope (Zeiss Axioplan) equipped with a cooled CCD camera (Photometrics). Pseudocoloring and merging of images were performed using the IpLab software.

7. CENPB-sat PLASMID VECTOR CONSTRUCTION

The portion of the CENPB-sat comprising the CENP-B box and lacking identity regions with the 37cen satellite was amplified from horse genomic DNA using the following primer oligonucleotides containing EcoRI and SalI adapters required for cloning purposes: CENPBsat-F 5'-ATTGAATTCCCTTTCTGACATAGGTGCTTTCTG-3' and CENPBsat-R 5'- ATTGTCGACGCTTTAGGACTTCTGCTTCTG-3'. PCR products were digested with EcoRI/SalI and cloned in the pSVal plasmid (Nergadze et al. 2009) using the same procedure described by Nergadze and collaborators to obtain an 8-copies-array of the cloned portion (Nergadze et al. 2014).

8. FLUORESCENCE *IN SITU* HYBRIDIZATION

Metaphase spreads preparation was performed as described in Piras et al. 2010. Briefly, mitotic cells were mechanically detached by blowing the medium on the dish surface. Then the cells were harvested, centrifuged and incubated with 10 ml KCl 0.075M at 37°C for 20 minutes and fixed in cold methanol: acetic acid (3:1) overnight. The fixative was changed two times and cells were spread onto glass slides.

Fluorescence *in situ* hybridization (FISH) was performed as described in Piras et al. 2010. Briefly, CENPB-sat satellite probe and whole genomic DNA were labeled by nick-translation with Cy3-dUTP or Alexa488-dUTP. For each slide, 250 ng of satellite and/or 25 ng of labelled whole genomic DNA in 50% hybridization solution were used. The probes were applied and both the probe and the metaphase spread preparation were simultaneously denatured on a hot block at 72°C for 3 minutes and 30 seconds. Hybridization was carried out overnight at 37°C. We then performed three post-hybridization washes of 5 minutes in 50% formamide, followed by three washes of 5 minutes in 2xSSC at 42°C. Chromosomes were counterstained with DAPI (0.2 μg/ml) and mounted with Fluorescence Mounting Medium (Dako). Digital grey-scale images for fluorescence signals were acquired with a fluorescence microscope (Zeiss Axioplan) equipped with a cooled CCD camera (Photometrics). Pseudocoloring and merging of images were performed using the IpLab software. Chromosomes were identified by computer-generated reverse DAPI banding according to the standard karyotypes.

9. IMMUNO-FISH ON METAPHASE SPREADS

After image acquisition, immunofluorescence slides were washed in 2xSSC for 10 minutes at room temperature and then fixed in cold methanol: acetic acid (3:1) for 15 minutes. The DNA probe was then applied and the slides were treated as for FISH experiments.

10. CHROMATIN IMMUNOPRECIPITATION

For each IP reaction at least 10 million cells were collected, centrifuged at 1700 rpm for 7 minutes and pooled. Formaldehyde, at the final concentration of 1%, was directly added to the pool of cells and left rocking 100 rpm at 26°C for 15 minutes. To quench formaldehyde, glycine was added to the final concentration of 0,125 M and left rocking at 26°C for 10 minutes. The pool was then centrifuged at 800 rcf for 5 minutes at 4°C to obtain a pellet, which was stored at -80°C for at least one night. The pellet was thawed gradually on ice and washed twice with PBS 1x supplied with Protease Inhibitor Cocktail (Roche).

The pellet was resuspended in ChIP lysis buffer (SDS 0,25%, 50 mM Tris-HCl pH 8, 10 mM EDTA pH 8) with PIC (Protease Inhibitor Complex), and divided into aliquots of 20 million cells per 650 µl. Resuspended cells were sonified with Branson Sonifier 250 to obtain fragments of 200-800 bp; the fragments size was checked on agarose gel.

Samples were centrifuged for 10 minutes at maximum speed at 4°C to collect the cross-linked sonicated chromatin. Each IP reaction was performed in 1250 µl of 10 million cells each. Therefore supernatant was brought to volume with Dilution buffer (0,5% Nonidet P40, 10 mM Tris-HCl pH 7,5, 2,5 mm MgCl2, 150 mM NaCl) supplied with PIC inhibitor.

Pre-clearing was performed with A/G beads (Protein A Sepharose™ 4 Fast Flow/Protein G Sepharose™ 4 Fast Flow, GE Healthcare), previously treated with a blocking buffer (phosphate-buffered saline containing sonicated *E. coli* genomic DNA 500 ng/µl and BSA 10 mg/ml) for 1 hour at 4°C on shacking. Then, after centrifugation at 4000 rpm for 5 minutes at 4°C, the supernatant was recovered and beads discarded. 240 µl of the supernatant were saved as Input (20% of the total chromatin used for each IP). The remaining was divided into aliquots and incubated first with the antibody of interest at 4°C overnight, followed by incubation with previously treated A/G beads for 3 hours at 4°C on the rocket. Samples were then centrifuged for 2

minutes at 1200 g at 4°C and the supernatant was removed. The beads were washed 5 times with cold ChIP wash buffer (0,25% SDS, 1% TritonX-100, 2 mM EDTA pH 8, 150 mM NaCl, 20 mM Tris-HCl pH 8) and the last wash with cold ChIP final wash buffer (0,25% SDS, 1% TritonX-100, 2 mM EDTA pH 8, 500 mM NaCl, 20 mM Tris-HCl pH 8). After discarding completely the last wash, the immunocomplexes were eluted adding ChIP elution buffer (1% SDS, 100 mM NaHCO3, 40 μg/ml RNase A). Samples were incubated at RT for 15 minutes, then at 37°C for 1 hour and finally reverse cross-linked at 65°C, over-night. The day after the DNA was purified and eluted using the kit Promega (Wizard SV Gel and PCR Clean-up System) according to the manufacturer's instructions.

After purification the DNA was quantified using the Quantus™ Fluorometer (Promega) with the QuantiFluor® ONE dsDNA System (Promega) according to manufacturer's instruction.

11. SLOT BLOT

Scalar amounts (0.5 ng, 1 ng, 2 ng) of immunoprecipitated and Input DNAs were transferred to nylon membranes (Amersham HybondTM-N, GE Healthcare) through a Minifold II apparatus (Schleicher and Schuell) and were denatured in NaOH 0.4 M/NaCl 0.6 M for 15 minutes at room temperature. The membranes were hybridized at 64°C for 18 hours in Church buffer containing the ^{32}P-α[dCTP]-labeled DNA probes (Megaprime DNA Labelling System, GE Healthcare kit), generated from the following DNA fragments: (i) 7 kb EcoRI/SacI 37cen fragment and (ii) a 7.2 kb EcoRI/SacI 2PI fragment, extracted from the plasmid clones described previously (Anglana et al. 1996; Piras et al. 2010); (iii) a 441 bp PCR-amplified fragment from horse genomic DNA, spanning a previously described ERE-1 insertion (Hill et al. 2010), were obtained using the 5'-CAAATGAATCAGCTCACCCTT-3' and 5'-ATAGGATCCTGAG AGACAACTTGCCACA-3' primers; (iiii) a telomeric probe, $(TTAGGG)_5$, previously prepared in our laboratory and described in Bertoni et al. 1994.

Post-hybridization washes were as following: twice in 2x SSC-0.5% SDS, 15 minutes and once in 0.2x SSC-0.5% SDS, 30 minutes at 64°C. The probe signal was detected exposing the filter over-night and the images were obtained using Cyclone Storaege phosphor system (Packard). The densitometric analysis was performed with the ImageJ 1.48v software.

Another exposition was performed for 5-10 days using traditional photographic films (Hyperfilm MP), which were then developed and fixed.

12. NEXT-GENERATION SEQUENCING OF ChIP EXPERIMENTS

An aliquot of DNA purified from immunoprecipitated or input chromatin was paired-end sequenced through an Illumina HiSeq2000 or HiSeq2500 platform by by IGA Technology Services (Udine, Italy). The length and the number of reads obtained in each dataset regarding Parts 3 and 4 are reported in Table 2. Details on the sequencing experiments of Parts 1 and 2 can be found in the attached papers.

Sample	**Read length (bp)**	**Total number of reads**	**Notes**
Horse CENP-B ChIP	100	27,682,380	
Horse CENP-B Input	100	55,359,526	
Donkey CENP-B ChIP	100	29,797,622	
Donkey CENP-B Input	100	126,605,282	
Grevy's zebra CENP-B	125	19,020,884	
Grevy's zebra CENP-A	125	32,468,528	
Grevy's zebra Input	125	24,488,430	
Burchell's zebra CENP-B	125	26,347,730	
Burchell's zebra CENP-A	125	24,326,822	
Burchell's zebra Input	125	19,102,434	
Horse CENP-A ChIP	100	42,683,528	*HorseC CENP-A ChIP in Part 2.*
Horse CENP-A Input	100	45,821,170	*HorseC CENP-A Input in Part 2.*
Donkey CENP-A ChIP	100	44,267,364	*DonkeyB CENP-A ChIP in Part 2.*
Donkey CENP-A Input	100	37,434,334	*DonkeyB CENP-A Input in Part 2.*
CHO CENP-B ChIP	125	27,621,076	
CHO CENP-B Input	125	33,440,796	
CHO CREST serum ChIP	125	21,182,738	
CHO CREST serum Input	125	37,809,056	

Table 2. Read length and total number of reads obtained from the sequencing of ChIP-seq experiments (Part 3 and Part 4).

13. RNA-seq

Total RNA was extracted from 6 million cells using the miRNeasy Mini Kit (QIAGEN) according to manufacturer's protocol. The quality of the extracted RNA was evaluated by electrophoresis through a 1% agarose gel. An aliquot of RNA was paired-end sequenced by IGA Technology Services through the Illumina HiSeq2500 platform. 40 million reads were requested for each sample. The length and the number of reads obtained in each dataset are reported in Table 3.

Sample	Read length (bp)	Total number of reads	Notes
Horse	125	59,090,294	*Dataset used in Parts 1 and 3.*
Donkey	125	79,164,582	

Table 3. Read length and total number of reads obtained from the sequencing of RNA-seq experiments.

14. BIOINFORMATIC ANALYSIS OF SEQUENCING DATA

Details of the bioinformatics analysis of sequencing data of Parts 1 and 2 are reported in the attached paper.

As far as Part 3 is concerned, for the identification of the CENP-B bound satellite from the horse reference genome, reads from the ChIP-seq experiment with the anti-CENP-B antibody on horse primary fibroblasts were aligned to the horse reference genome (EquCab 2.0, 2007 release) with Bowtie (version 1.1.2), using the single end mode and k = 10 correction in order to refine the mapping of reads from satellite repeats (Langmead et al. 2009). Peak calling on the sequencing data was then performed through MACS14 (version 1.4.1). Stringent criteria were arbitrarily applied: chrUn selection, fold enrichment > 8, -10Log_{10}(p-Value) >100 and FDR (%) < 1. The 57 top-ranked regions were analyzed through Tandem Repeat Finder (Benson et al. 1999) (Table 3). For each region, Tandem Repeat Finder reports one or more classes of tandem repeats, providing a consensus for each class. The 425 bp consensus sequence of CENPB-sat was obtain by Multalin alignment of all the 59 identified consensus sequences containing a canonical CENP-B box. Consensus sequences other than CENPB-sat identified by Tandem Repeat Finder were analyzed using RepeatMasker (http://www.repeatmasker.org/) (Table 4).

chr	start	end	Length (bp)	-10Log$_{10}$ (p-value)	Fold enrichment	FDR (%)	Satellite DNA families
chrUn	33874464	33887158	12695	1239.81	11.21	0	Functional CENPB-sat
chrUn	71614738	71622796	8059	3100	9.99	0	Functional CENPB-sat
chrUn	110961147	110964445	3299	3100	9.61	0	Functional CENPB-sat
chrUn	117211357	117214356	3000	3100	9.6	0	Functional CENPB-sat
chrUn	82122665	82128068	5404	3100	9.5	0	Functional CENPB-sat
chrUn	54812843	54845204	32362	1503.41	9.45	0	Functional CENPB-sat
chrUn	72051332	72054373	3042	3100	9.29	0	Functional CENPB-sat
chrUn	91274650	91279027	4378	3100	9.2	0	Functional CENPB-sat
chrUn	101795688	101799360	3673	3100	9.19	0	Functional CENPB-sat
chrUn	94942932	94947038	4107	3100	9.1	0	Functional CENPB-sat
chrUn	97854352	97858229	3878	3100	9	0	Functional CENPB-sat
chrUn	114321027	114324151	3125	2099.4	8.84	0	Functional CENPB-sat 2PI
chrUn	72542211	72549683	7473	3100	8.76	0	Functional CENPB-sat
chrUn	105093860	105097109	3250	3100	8.73	0	Functional CENPB-sat
chrUn	73982079	73989070	6992	3226.07	8.71	0	Functional CENPB-sat 2PI
chrUn	93732892	93736918	4027	3100	8.68	0	Functional CENPB-sat
chrUn	67249865	67260632	10768	3100	8.67	0	Functional CENPB-sat
chrUn	97242262	97246111	3850	3100	8.6	0	Functional CENPB-sat
chrUn	90074831	90079410	4580	3100	8.59	0	Functional CENPB-sat
chrUn	83871324	83873940	2617	3100	8.57	0	Functional CENPB-sat
chrUn	49208565	49216007	7443	1584.48	8.52	0	Functional CENPB-sat Degenerated CENPB-sat 2PI
chrUn	75011879	75018743	6865	3100	8.52	0	Functional CENPB-sat
chrUn	70779673	70787915	8243	3100	8.46	0	Functional CENPB-sat
chrUn	63171762	63175412	3651	623.55	8.43	0	Degenerated CENPB-sat 2PI
chrUn	66165502	66177256	11755	2558.77	8.4	0	Functional CENPB-sat
chrUn	67433207	67443916	10710	3100	8.37	0	Functional CENPB-sat
chrUn	112717344	112720649	3306	3100	8.36	0	Functional CENPB-sat
chrUn	74855626	74859946	4321	3100	8.34	0	Functional CENPB-sat
chrUn	105016617	105020096	3480	3100	8.32	0	Functional CENPB-sat
chrUn	96783313	96787107	3795	3100	8.29	0	Functional CENPB-sat
chrUn	109111816	109115169	3354	3100	8.28	0	Functional CENPB-sat
chrUn	114682903	114686026	3124	3100	8.28	0	Functional CENPB-sat 2PI
chrUn	65558641	65561764	3124	3100	8.26	0	Functional CENPB-sat
chrUn	74891764	74898613	6850	3100	8.24	0	Functional CENPB-sat
chrUn	100562408	100566192	3785	3100	8.24	0	Functional CENPB-sat
chrUn	96066658	96070674	4017	3100	8.23	0	Functional CENPB-sat
chrUn	84177172	84182335	5164	3100	8.22	0	Functional CENPB-sat
chrUn	80163195	80167137	3943	3100	8.21	0	Functional CENPB-sat
chrUn	92957610	92961928	4319	3100	8.21	0	Functional CENPB-sat
chrUn	92038299	92042736	4438	3100	8.19	0	Functional CENPB-sat
chrUn	115562128	115565201	3074	3100	8.17	0	Functional CENPB-sat
chrUn	116479316	116482452	3137	3100	8.17	0	Functional CENPB-sat
chrUn	103010170	103013844	3675	3100	8.16	0	Functional CENPB-sat
chrUn	93347995	93352279	4285	3100	8.15	0	Functional CENPB-sat
chrUn	97074723	97078518	3796	3100	8.15	0	Functional CENPB-sat

							Simple repeats
chrUn	73660504	73662578	2075	3138.48	8.14	0	Functional CENPB-sat
chrUn	113226530	113229779	3250	3100	8.12	0	Functional CENPB-sat
chrUn	97975453	97979320	3868	1440.33	8.1	0	Functional CENPB-sat 2PI
chrUn	106204234	106207770	3537	3100	8.1	0	Functional CENPB-sat
chrUn	77599200	77605346	6147	3100	8.09	0	Functional CENPB-sat
chrUn	92798774	92803115	4342	3100	8.09	0	Functional CENPB-sat
chrUn	70299321	70308126	8806	3100	8.07	0	Functional CENPB-sat
chrUn	95774383	95778362	3980	3100	8.07	0	Functional CENPB-sat
chrUn	111501421	111504497	3077	3100	8.06	0	Functional CENPB-sat
chrUn	112437155	112440412	3258	3100	8.06	0	Functional CENPB-sat
chrUn	84795813	84800812	5000	3100	8.04	0	Functional CENPB-sat
chrUn	102517076	102520774	3699	3100	8.02	0	Functional CENPB-sat

Table 4. Peak calling enriched genomic regions. For each peak, chromosome (chr), genomic coordinates on EquCab2.0 (start, end), length, statistical parameters ($-10Log_{10}$(p-value), fold enrichment and % FDR), satellite DNA families identified through Tandem Repeat Finder and RepeatMasker analysis.

To evaluate enrichment, genomic abundance or transcription of different satellite families, reads deriving from ChIP-seq or RNA-seq experiments of both Part 3 and Part 4 were mapped on a custom reference genome, made by the fasta sequences of the desired satellite monomer units, by Bowtie2.0 (Langmead and Salzberg 2012), using the single end mode and default parameter. In Part 3, ES22 (in RepBase) and AH010654.2 (NCBI Nucleotide) were used as reference sequences of 2PI satellite. SAT2PI in RepBase was not used since it is a wrong update of the 2PI sequence described in Piras et al. 2010: it contains both 22 bp units, corresponding to 2PI, and 419 bp units corresponding to degenerated CENPB-sat sequences. Count data from resulting BAM files were obtained using idxstats command from the Samtools package (Li et al. 2009).

De novo assembly of the CENP-B box environment was performed using the MEME-ChIP tool (Machanick and Bailey 2011) available in the MEME Suite web portal (http://meme-suite.org). We adopted a "consensus-walking" strategy starting from raw ChIP reads of the ChIP-seq experiments performed with the anti-CENP-B antibody. Using the canonical CENP-B box (5' NTTCGNNNNANNCGGGN 3') as a bait, we fished ChIP reads containing the box at three known positions in the read length: at the beginning, in the middle and at the end of the read. This selection procedure allowed us to have three groups of reads with a common pivot, the CENP-B box, at known positions. Each group of reads was analyzed with MEME-ChIP to detect common motifs beyond the always shared CENP-B box. MEME-ChIP identifies three short consensus motifs for each group of reads. All the

identified short motifs aligned to the CENPB-sat reference and we could derive the final species-specific consensus sequence by "consensus walking".

In Part 4, read alignment and peak calling was performed as described above for Part 3 (Table 5). Tandem Repeat Finder analysis provide us 20 consensus sequences of the repeated unit containing a G16>A box. By Multalin alignment of these consensus sequences we derived the consensus sequence of the repeated unit of this satellite. The KE379478:2296633-2297147 region comprises neither CENP-B boxes nor telomeric-like repeats. It consists of 9.2 tandem repetitions of a 42 bp repeat, sharing up to 51.6% of identity with the monomers of the SatCH5 satellite sequence deposited in NCBI Nucleotide (Accession numbers: AJ131828.1 and AJ131829.1).

Scaffold	Start	End	Length (bp)	$-10Log_{10}$ (p-value)	Fold enrichment	FDR (%)	Satellite families
KE376648	5515096	5516148	1053	736.82	4.7	0	Sau1a satellite Telomeric-like
KE379478	2296633	2297147	515	580.32	22.47	0	SatCH5
KE379717	17398	19257	1860	1330.59	17.79	0	Sau1a satellite Telomeric-like
KE381306	44	777	734	1004.49	6.34	0	Sau1a satellite
KE383256	3	6755	6753	3100	21.39	0	Sau1a satellite
KE383440	3	574	572	3100	119.09	0	Telomeric-like

Table 5. Peak calling enriched genomic regions. For each peak, scaffold, genomic coordinates on criGri1 (start, end), length, statistical parameters ($-10Log_{10}$(p-value), fold enrichment and % FDR), satellite DNA families identified through Tandem Repeat Finder analysis.

15. 3D-FISH, 3D-IMMUNOFISH AND FISH ON RETINA CRYOSECTIONS

3D-FISH experiments were performed using the procedure described by Solovei and Cremer (Solovei and Cremer 2010). Briefly, fibroblasts were grown on 20x20 mm coverslips till confluence, to obtain the majority of cells in G0 phase. Coverslips with cells were rinsed in 2 changes of 1x PBS at 37°C. Cells were fixed in 4% paraformaldehyde in PBS (pH 7.0) for 10 minutes at room temperature, adding a drop of 0.5% Triton X-100 in 1x PBS in the last minute of incubation. Coverslips were then washed three times with 1x PBS supplemented with 0.01% Tween-20 for 5 minutes at room temperature and subsequently permeabilized with 0.5% Triton X-100 in 1x PBS for 20 minutes at room temperature. Cells were equilibrated in 20% glycerol in 1x PBS for at least 1 hour at room temperature. Cells were then

treated with four cycles of freezing in liquid nitrogen (15-30 sec), thawing gradually at room temperature and soaking again with 20% glycerol. Coverslips were then washed three times with PBS/0.01% Tween-20 for 10 minutes at room temperature and incubated in 0.1M HCl for 5 minutes at room temperature, changing the solution one time. Cells were rinsed in 2xSSC and equilibrated in 50% formamide in 2xSSC for at least 30 min at room temperature or few days at 4°C.

The equine BAC clones of interest (Table 6) as well as the 37cen satellite probe and whole genomic DNA were labeled by nick translation with Cy3-dUTP, FITC-dUTP or Texas Red-dUTP using a slight modification of the procedure described by Piras et al. (Piras et al. 2010). Briefly, 50 ng of labelled whole genomic DNA or 37cen satellite and 125 ng of BAC probe in 50% formamide hybridization solution were used for each 24x24 mm coverslip. The probes were directly mounted on the coverslip, sealed with rubber cement and then both probe and cells were denatured on a hot block simultaneously, 3 minutes at 75°C. Hybridization was carried out in a water bath for two days at 37°C. Post-hybridization washes were 3 washes of 10 minutes in 2x SSC at 37°C followed by a 5 minutes wash in 0.1x SSC at 62°C. Nuclei were counterstained with DAPI 2 μg/ml.

In case of immuno-FISH for centromere immunostaining, the immunofluorescence part was performed before the glycerol equilibration. In particular, the primary antibody (CREST serum, diluted 1:200 in a 4% BSA, 0.01% Tween-20 in PBS blocking solution) was incubated for 1 hour at room temperature. Coverslips were then washed three times with PBS/0.01% Tween-20 for 5 minutes at room temperature. The secondary antibody (Alexa488 or rhodamine-conjugated anti-human) was incubated for 1 hour at room temperature and then washes were repeated.

Immunodetection of nucleoli was performed after the FISH protocol. Cells were incubated for 1 hour with the anti-B23 antibody (B0556, Sigma) diluted 1:100 in the blocking solution at room temperature. Secondary antibody (Alexa647-conjugated anti-mouse, diluted 1:500 in the blocking solution) was incubated 1 hour at room temperature.

As far as horse retina is concerned, cryosections were dried up for 30 minutes at room temperature and then re-hydrated in Na Citrate buffer for a minute and transferred into the same pre-warmed up to 80ºC buffer in water bath and incubated for 30 minutes. Crysections were then equilibrated in 2x SSC and incubated in 50% FA/SSC for 30-60 min. The section was covered with a chamber and the labeled probe was loaded. Cells and labeled probes were simultaneously denatured on hot block at 80°C for 3 minutes.

Hybridization and post-hybridization washes were performed as previously described.

Stacks of optical sections through whole nuclei and retina sections were collected using a Leica TCS SP5 confocal microscope equipped with Plan Apo 63×/1.4 NA oil immersion objective and lasers with excitation lines 405, 488, 561, 594, and 633 nm. Stacks were obtained with axial distance of 200 nm between optical sections. Dedicated ImageJ plugins were used to compensate for axial chromatic shift between channels in confocal stacks and to create RGB stacks (Walter et al. 2006, Ronneberger et al. 2008, Van der Werken et al. 2017).

BAC clone	EquCab2.0 coordinates	Localization	
		ECA	EAS
CH241-21D14	chr11:27,639,936 – 27,829,952	ECA11 cen	EAS13 q dist
CH241-232I17	chr11:46,749,358-46,973,150	ECA11 q dist	EAS13 cen
CH241-20K22	chr19:4,913,928-5,070,194	ECA19 q prox	EAS5 cen

Table 6. Equine BAC clones with horse genomic coordinates and cytogenetic localization on horse (ECA) and donkey (EAS).

16. IMMUNOFLUORESCE AND IMMUNO-FISH ON HORSE PACHYTENE SPREADS

Pachytene spreads were prepared starting from testes samples from three different horses, using a modified version of the protocol described in Peters et al. 1997. Testes were cut in small pieces (about 1 cm^3) and frozen at -80°C. Thin slices of frozen tissue were cut and immediately hydrated with ice-cold PBS 1x. The samples were then homogenized using a scalpel blade. All the procedure was carried out on ice. Cell suspension diluted 1:10 in a hypothonic KCl 75mM solution and incubated on ice for 13 minutes. About 50 µl of this suspension was applied on a glass slide that had been dipped just before in the fixative solution. Fixation with 1% formaldehyde, 0.015% TritonX-100 (pH 9.8) was used for the preparation of slides for immunofluorescence with the anti-CENP-A antibody. Fixation with 4% paraformaldehyde (pH 10) in 1x PBS, 0.015% TritonX-100 was used for the preparation of slides for immunofluorescence with the CREST serum. Slides were then kept in a humid chamber for 8 minutes at room temperature in the case of formaldehyde fixation and for 30 minutes at room temperature in the case of paraformaldehyde fixation.

Slides were permeabilized in 1x PBS supplemented with 0.005% Tween-20 for 25 minutes. Incubation with the anti-SCP3 antibody, diluted 1:200, anti-CENP-A sheep serum (diluted 1:100 in PBST) or CREST serum (diluted 1:250 in PBST) was performed for 1 hour at 37°C. Incubation with the anti-SCP3 antibody, diluted 1:200, CREST serum (diluted 1:250 in PBST) and anti-MLH1 antibody (diluted 1:50 in PBST) was performed for 14 hour at 4°C. The secondary antibodies (rhodamine conjugated anti-rabbit, Alexa488 conjugated anti-sheep, Alexa488 conjugated anti-human, Alexa647 conjugated anti-human, Alexa488 conjugated anti-mouse) were diluted 1:100 in PBST and incubated for 1 hour at 37°C.

For immuno-FISH, after immunofluorescence image acquisition, slides were denatured in 70% formamide in 2x SSC for 5 minutes at 74°C, treated with sodium thiocyanate 1M for 3 hours at 65°C and denatured again in 70% formamide in 2x SSC for 2 minutes at 74°C. Slides were then dehydrated for 2 minutes in cold 70% ethanol, 2 minutes in 70% ethanol at room temperature, 2 minutes in 90% ethanol at room temperature, 2 minutes in 100% ethanol at room temperature. Probe preparation was then performed as described above. Hybridization was performed for 48 hours at 37°C. Post-hybridization washes were: 5 minutes in 2x SSC supplemented with 0.05% Tween-20 at 42°C, three washes of 5 minutes in 4x SSC supplemented with 0.05% Tween-20

PART 1
THE MAJOR HORSE SATELLITE DNA FAMILY IS ASSOCIATED WITH CENTROMERE COMPETENCE

In *Equus caballus* all the centromeres, with the exception of the one of chromosome 11, are satellite-based. Previous work from our laboratory demonstrated that 37cen and 2PI are the most abundant satellite DNA families in this species and either one or both these satellites are present on all chromosomes, except chromosome 11, and only at primary constrictions (Piras et al. 2010).

To identify the satellite DNA repeats bearing the centromeric function, we performed a ChIP-seq experiment with an anti-CENP-A serum on horse primary skin fibroblasts. We proved that, in the horse, 37cen is the satellite DNA family bound by CENP-A and thus endowed with the centromeric function. The 37cen sequence bound by CENP-A is GC-rich with 221 bp units organized head-to-tail. The association between 37cen and CENP-A was confirmed through slot blot experiments, in which I was involved directly. In addition, we showed by RNA-seq that the centromeric satellite 37cen is transcriptionally active, adding new evidence to the hypothesis that centromeric transcripts may be required for centromere function. All these results were published in 2016 on Molecular Cytogenetics (see the attached publication).

PART 2
BIRTH, EVOLUTION, AND TRANSMISSION OF SATELLITE-FREE MAMMALIAN CENTROMERIC DOMAINS

In previous work, we described the first example of a natural satellite-free centromere on *Equus caballus* chromosome 11 (Wade et al. 2009). Cytogenetic data suggested that, in *Equus asinus*, several centromeres are satellite-free, while many satellite DNA loci are not centromeric (Piras et al. 2010). Thus, we investigated the satellite-free centromeres of *Equus asinus* by ChIP-seq with anti-CENPA antibodies.

We identified an extraordinarily high number of centromeres devoid of satellite DNA (16 of 31), namely the centromeres of EAS4, EAS5, EAS7, EAS8, EAS9, EAS10, EAS11, EAS12, EAS13, EAS14, EAS16, EAS18, EAS19, EAS27, EAS30 and EASX.

These satellite-less centromeres spanned 54–345 kb and contained one or two CENPA binding domains. Similar to what we described for the centromere of horse chromosome 11 (Purgato et al. 2015), the presence of two domains is due to different epialleles on the two homologs. The analysis of epiallele transmission in hybrids (three mules and one hinny) showed that centromeric domains are inherited as Mendelian traits, but their position can slide in one generation. Conversely, the position of the centromere is stable during mitotic propagation of cultured cells.

The sequences of the 16 donkey satellite-less centromeres were assembled by both NGS approach and Sanger method. Sequence analysis demonstrated that all of them lay in LINE- and AT-rich regions. In addition, five centromeres (EAS8, EAS9, EAS16, EAS18, and EAS19) were characterized by novel tandem repetitions of sequences that are single copy in the horse genome.

Our results demonstrate that the presence of more than half of centromeres void of satellite DNA is compatible with genome stability and species survival. The presence of amplified DNA at some centromeres suggests that these arrays may represent an intermediate stage toward satellite DNA formation during evolution. The fact that CENPA binding domains can move within relatively restricted regions (a few hundred kilobases) suggests that the centromeric function is physically limited by epigenetic boundaries.

These results were published in 2018 on Genome Research (see the attached publication). In particular, I was involved in sequence assembly and

contributed to immunoprecipitation experiments and RT-PCR experiments to prove the presence of tandem repetitions in the subset of centromeres described above.

PART 3
CENP-B IN THE GENUS *Equus*

Results

CENP-B is the only known centromeric protein that exhibits unequivocal DNA binding specificity, recognizing the so-called CENP-B box. In spite of the high conservation of CENP-B protein and its binding motif across mammals, this protein appears dispensable for the centromeric function, giving rise to the so-called "CENP-B paradox". The aim of this work was to shed light on the role of the CENP-B protein in the epigenetic establishment of centromeric chromatin in mammals, taking advantage of the model system of the genus *Equus*, given the extraordinary plasticity of their centromeres. Indeed, the genus *Equus* provides us the opportunity of evaluating the association between CENP-B, satellites and centromeres in a system characterized by the coexistence of both satellite-associated and satellite-free centromeres.

1. CENP-B GENE AND PROTEIN

In order to characterize CENP-B in the genus *Equus*, we examined its gene structure and protein product in *Equus caballus* (horse, ECA), *Equus asinus* (domestic donkey, EAS), *Equus grevyi* (Grevy's zebra, EGR) and *Equus burchelli* (Burchell's zebra, EBU).

Among these equid species, a high-quality genome assembly at the chromosomal level is available only for the horse (Wade et al. 2009), while only draft sequences of the donkey genome comprising unassembled scaffolds are available (Orlando et al. 2013, Huang et al. 2015, Renaud et al. 2018). In the EquCab3 assembly of the horse, the locus of CENP-B is entirely present, maps at chr22:19,582,683-19,584,500 and the coding sequence of CENP-B is annotated (XM_023626169). On the other hand, the coding sequence of CENP-B gene was partial in the donkey genome scaffolds. Indeed, the predicted donkey CENP-B protein (XP_014722562.1) is partial and lacks the first three amino acids. No CENP-B gene record was available for the other species. Thus, we assembled the CENP-B coding sequence of these species, using both Sanger sequencing and NGS data obtained in our laboratory (see Materials and Methods).

As for all studied mammals, all the equid CENP-B genes were found to contain no introns. A comparative analysis of the CENP-B gene in the four *Equus* species revealed that at position 1288 an in-frame GAG insertion can be observed in the donkey, in the Grevy's zebra and in the Burchell's zebra compared to the horse, while, at position 1408 an in-frame GAG deletion is found in the donkey compared to the other three species (Figure 1A). Indeed, the CENP-B gene codes for a 605 amino-acid long protein in the horse and the donkey while in the Grevy's zebra and in the Burchell's zebra the CENP-B protein contains 606 aa. The in-frame insertions and deletions occur in the 1210-1407 region encoding the first one of the two acidic domains of CENP-B (Sullivan and Glass 1991, Kitagawa et al. 1994), which are exceedingly rich in glutamic acid and coded by clusters of GAG and GAA codons repeated from two to six times (Figure 1A).

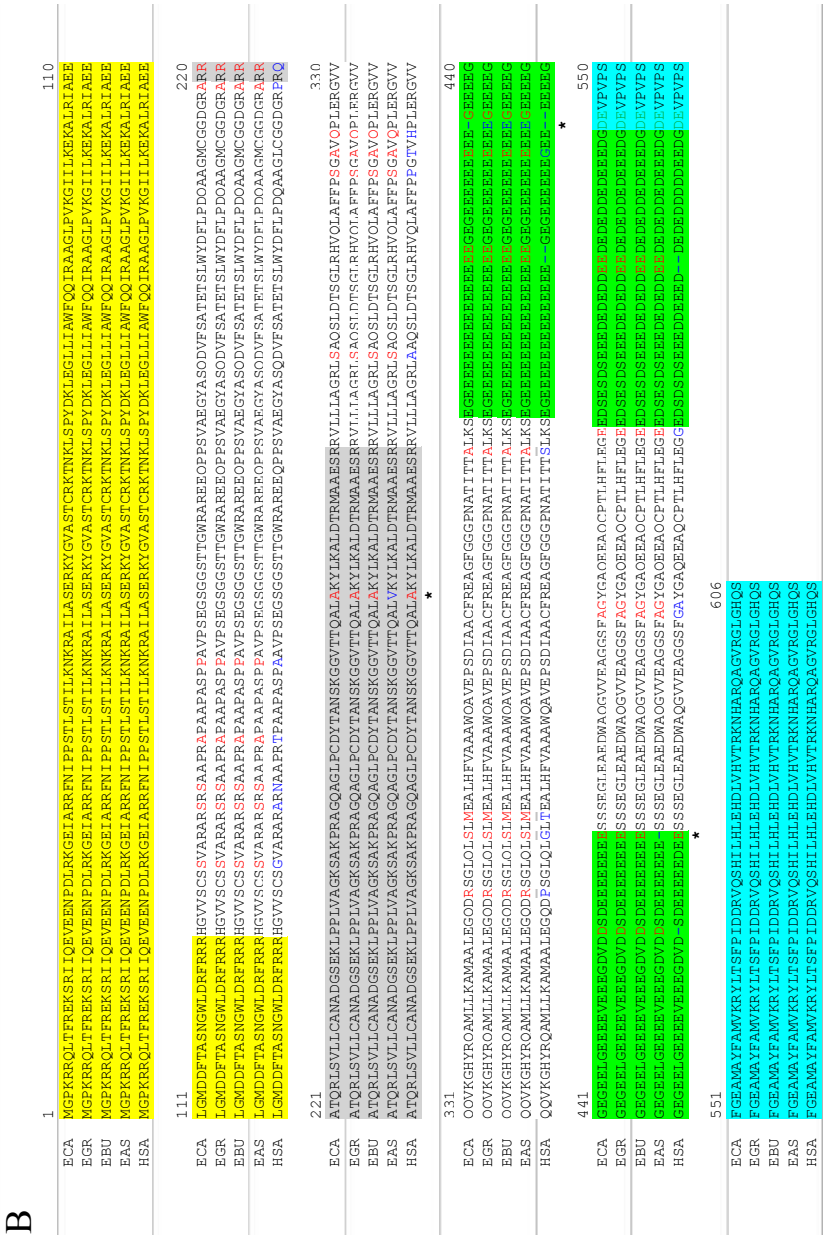
Moreover, we detected 9 single nucleotide differences in the CENP-B coding sequences of these species: eight are silent while one determines an alanine-to-valine substitution at position 275 in the donkey protein with respect to the other three species (Figure 1B).

It must be underlined that the DNA binding domain and the dimerization domain of all the equid CENP-B proteins are identical to those of human and of nearly all the other mammalian species studied so far (see Introduction), suggesting that in equids CENP-B is functional and able to recognize a canonical CENP-B box.

A

Figure 1. Coding sequence of CENP-B in the four equid species. A) Alignment of the coding sequence of the first acidic domain of horse (ECA), Grevy's zebra (EGR), Burchell's zebra (EBU) and donkey (EAS) CENP-B proteins. GAG triplets are highlighted in cyan and GAA triplets are highlighted in yellow. Missing or additional GAG codons are written in red and marked with an asterisk. B) Alignment between horse (ECA), Grevy's zebra (EGR), Burchell's zebra (EBU), donkey (EAS) and human (HSA, NP_031708.2) CENP-B proteins. The DNA binding domain, the endonuclease domain, the acidic domains and the dimerization domain are highlighted in yellow, grey, green and turquoise blue, respectively. Amino acid differences are shown using blue (high consensus color) or red (low consensus color). The three amino acid differences among the equid CENP-B proteins are marked with an asterisk.

B



CENP-B expression was detected in all the four species by western blotting. However, its amount is different in the four species. As shown in Figure 2, the amount of CENP-B is similar in the horse and in the donkey. The Grevy's zebra shows a higher expression level of CENP-B, compared to horse and donkey. On the contrary, the amount CENP-B in the Burchell's zebra is lower than that of horse and donkey.

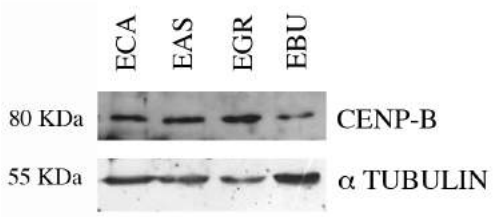


Figure 2. Expression of CENP-B protein in the four equid species. Western blotting analysis on whole nuclear extract from horse (ECA), donkey (EAS), Grevy's zebra (EGR) and Burchell's zebra (EBU). The band corresponding to CENP-B was detected at 80 kDa and the one of the α tubulin loading control was detected at about 55 kDa.

2. ABSENCE OF CANONICAL CENP-B BOXES IN THE HORSE MAJOR CENTROMERIC SATELLITE AND IN THE SATELLITE-LESS CENTROMERES

It is well described in literature that CENP-B boxes are contained in centromeric satellites. Surprisingly, sequence analysis of the horse major centromeric satellite 37cen (Cerutti et al. 2016) revealed that no CENP-B recognition motifs are contained in this satellite family. No CENP-B binding sites were detected also in 2PI satellite, the other highly represented satellite DNA family of equid species (Piras et al. 2010).

The extraordinarily high number of satellite-less centromeres in these equid species (Wade et al. 2009, Piras et al. 2010, Nergadze et al. 2018) raises the question whether CENP-B boxes might be present at such centromeres. We searched for the presence of CENP-B recognition sites in the sequences of the satellite-less centromeres that we previously assembled, namely the unique horse satellite-less centromere (Wade et al. 2009) and the 16 satellite-free centromeric domains of *E. asinus* (Nergadze et al. 2018). No CENP-B boxes were detected within these centromeric domains.

3. LOCALIZATION OF THE CENP-B PROTEIN

Recent data suggested that CENP-B contributed to the fidelity of segregation by interacting with both CENP-A and CENP-C (Fachinetti et al. 2015). In particular, CENP-B was shown to mediate a pathway for CENP-C recruitment and maintenance at centromeres. Artificial depletion of CENP-B in human and mouse cell lines resulted in dramatic reduction of CENP-C levels, thus increasing chromosome missegregation frequency (Fachinetti et al. 2015).

In order to test this hypothesis in our model system, we initially analyzed the localization of CENP-B with respect to that of CENP-A and CENP-C by immunofluorescence on metaphase spreads from primary fibroblasts of *E. caballus*, *E. asinus*, *E. grevyi* and *E. burchelli*. As in the horse (Wade et al. 2009), in the three other equid species, CENP-A and CENP-C marks all primary constrictions (Figure 3 and 4).

In *E. caballus* (*2n*=64), CENP-B signals could be detected at the primary constriction of nine pairs only: three metacentric (2, 6 and 10) and six acrocentric chromosomes (17, 18, 21, 23, 24 and 29) (Figure 5). The signal intensity varies significantly among different chromosomes. By double immunofluorescence experiments, we demonstrated that all CENP-B signals co-localized with CENP-A and CENP-C signals, confirming their centromeric position at cytogenetic level (Figure 3A, Figure 4A). However, the signals do not show the typical specked pattern of the ones of CENP-A and CENP-C but are broad and surround the dots of CENP-A and CENP-C, suggesting an extended pericentromeric localization for CENP-B. Indeed, CENP-B signals cover the whole primary constriction, while CENP-A and CENP-C, as expected (Blower et al. 2002, Allshire and Karpen 2008, Fukagawa and Earnshaw 2014), localize on the outer centromere region (Figure 6). Differently from CENP-B, the intensity of CENP-A and CENP-C signals did not vary between different chromosomes and did not correlate with CENP-B level (Figure 3A and 4A).

In *E. asinus* (*2n*=62), no CENP-B signal could be detected, while all centromeres were labeled by CENP-A and CENP-C (Figure 3B and 4B). In order to rule out that technical problems may be responsible for the lack of signals on donkey chromosomes, we examined CENP-B localization in a fibroblast cell line from *E. burdo* (hinny), a domestic equid hybrid that is the offspring of a horse stallion and a jenny donkey. It is important to remind that in the hinny hybrid half of the chromosomes derive from the horse and half from the donkey. In hinny metaphase spreads, we could detect only nine

CENP-B labeled chromosomes, which correspond to the chromosomes identified in the horse (Figure 7). Thus, we confirmed that only nine horse chromosomes are labeled while no donkey chromosome is cytogenetically marked by CENP-B.

In *E. grevyi* (*2n*=46), the majority of CENP-B signals are uncoupled from CENP-A and CENP-C (Figure 3C and 4C). In particular, CENP-B localizes at the p arm terminus of eight metacentric chromosomes (1, 2, 5, 10, 13, 14, 15 and 16), at both the centromere and the p arm terminus of two metacentric chromosomes (6 and 12) and at the q arm terminus of three acrocentric chromosomes (20, 21 and 22) (Figure 8). Polymorphism regarding signal presence can be observed in chromosomes 2, 13 and 16, where only one homolog displayed CENP-B signal (Figure 8). As for the horse, the signals have a spot-like appearance, differently from the speckled pattern of CENP-A and CENP-C (Figure 3C and 4C).

Finally, in *E. burchelli* (*2n*=44), as in the donkey, no CENP-B signals could be detected, whereas CENP-A and CENP-C signals are homogeneous among all the centromeres (Figure 3D and 4D).

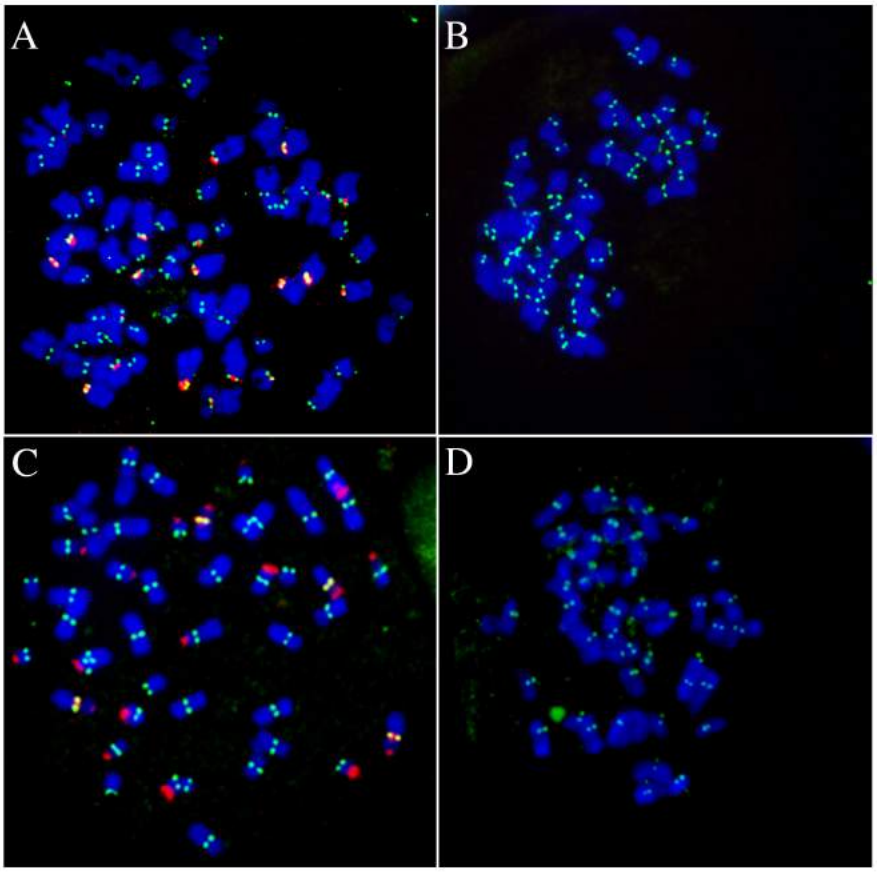


Figure 3. Localization of CENP-B and CENP-A in the four equid species. Double immunofluorescence with an anti-CENP-B antibody (red) and an anti-CENP-A serum (green) on DAPI-stained metaphase chromosomes from horse (A), donkey (B), Grevy's zebra (C) and Burchell's zebra (D).

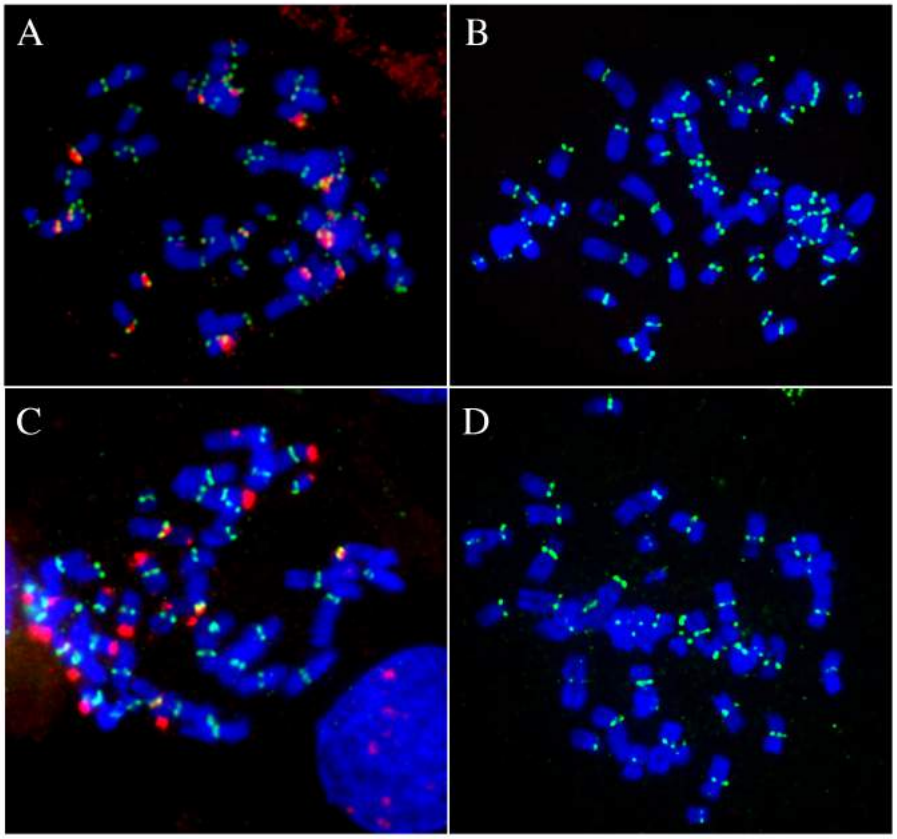


Figure 4. Localization of CENP-B and CENP-C in the four equid species. Double immunofluorescence with an anti-CENP-B antibody (red) and an anti-CENP-C serum (green) on DAPI-stained metaphase chromosomes from horse (A), donkey (B), Grevy's zebra (C) and Burchell's zebra (D).

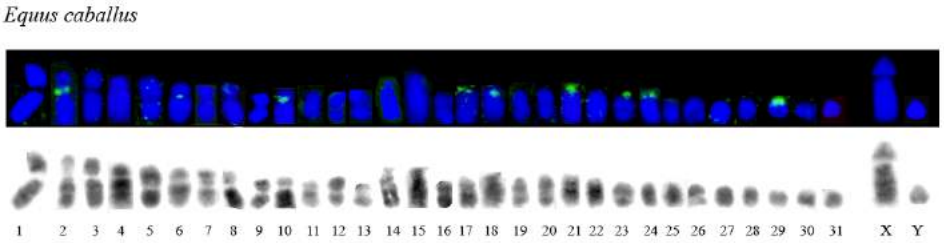


Figure 5. Identification of chromosomes with CENP-B signal in the horse. In the panel above, CENP-B signals (green) on pseudocolored chromosomes (blue). In the panel below, computer-generated reverse DAPI banding.

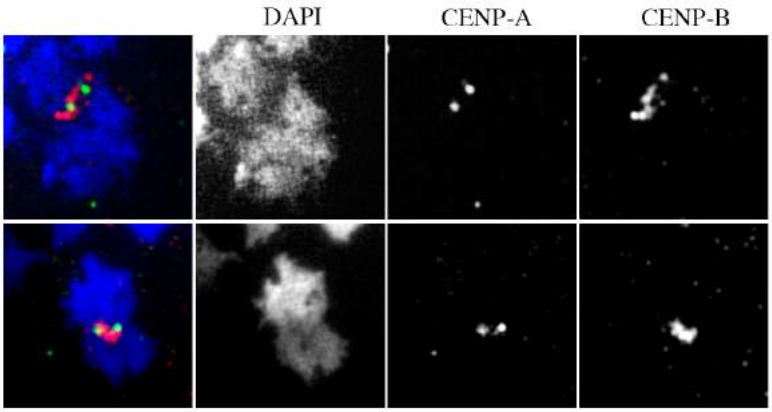


Figure 6. Pattern of CENP-B signals in the horse. Immunofluorescence with an anti-CENP-B antibody (red) and an anti-CENP-A serum (green) on DAPI-stained metaphase chromosome from horse.

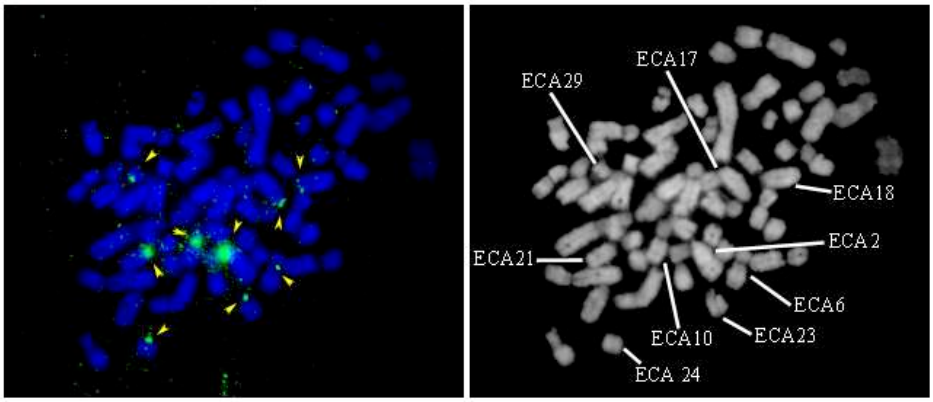


Figure 7. Localization of CENP-B in the hinny. On the left, immunodetection of CENP-B signals (green, arrows) on DAPI-stained metaphase chromosomes from hinny. On the right, DAPI image converted to black and white is shown. Chromosomes with CENP-B signals are identified.

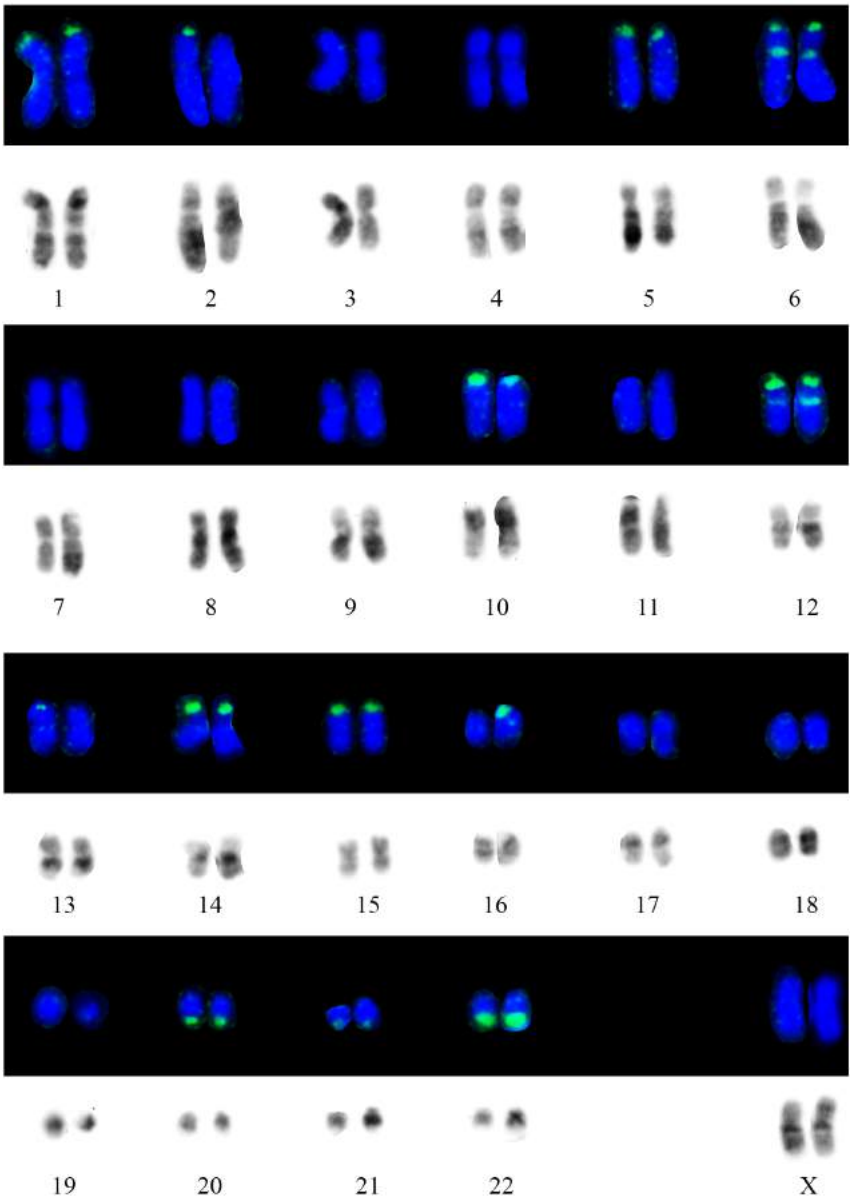


Figure 8. Identification of chromosomes with CENP-B signal in the Grevy's zebra. For each line: above, CENP-B signals (green) on pseudocolored chromosomes (blue); below, computer-generated reverse DAPI banding

4. CHARACTERIZATION OF THE CENP-B BOUND SATELLITE

4.1. ChIP-seq identification of the CENPB-sat satellite in the horse genome

In order to characterize the CENP-B binding sites in the horse genome, we performed a ChIP-seq experiment with an antibody against the centromeric protein CENP-B to immunoprecipitate the chromatin extracted from *E. caballus* skin primary fibroblasts, using EquCab2 assembly as reference sequence. It must be underlined that the majority of satellite DNA sequences cannot be placed in the genome assembly and, thus, highly repetitive DNA sequences lacking chromosome assignment are included in a "virtual" chromosome, named "unplaced" (chrUn); since CENP-B boxes are known to be contained in satellite sequences, we focused our analysis on "unplaced" contigs. We chose the EquCab2 assembly rather the EquCab3.0 one since, in this last assembly, many contigs derived from satellite-based centromeric domains were placed in the genome and removed from the chrUn. We preferred to focus our analysis to the overall unplaced scaffolds of EquCab2.0 to avoid the assembly bias present in EquCab3.

Applying specific stringent criteria for peak calling (see Materials and Methods), 57 highly enriched regions, spanning 2-32 kb, of chrUn were identified (see Materials and Methods, Table 4). Sequence analysis demonstrated that these regions are composed by tandem repeats of up to 76 repetitions of an about 425 bp unit, comprising a canonical CENP-B box (5' TTTCGTCTGAGCCGGGT 3'). The consensus sequence of the repeated unit of this novel satellite family, from now called the CENPB-sat satellite, is shown in Figure 9 as a logo. Its GC content is 50.5%.

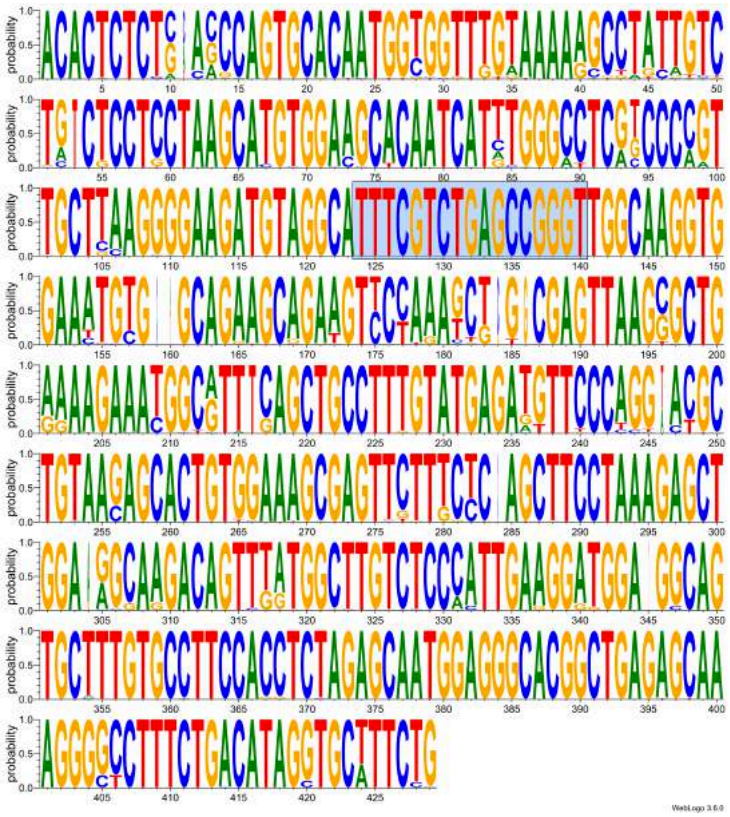


Figure 9. Consensus sequence of the unit of the CENPB-sat satellite. Logo of the consensus sequence of CENPB-sat. The CENP-B box is highlighted with a cyan background.

It should be pointed out that seven of these 57 enriched regions contain also CENPB-sat monomers with a degenerated CENP-B box as well as CENPB-sat arrays interrupted by stretches of 2PI satellite. The 2PI satellite, previously described in our laboratory (Piras et al. 2010), is composed by 22 bp tandem repetitions and known to be pericentromeric in the horse (Piras et al. 2010, Cerutti et al. 2016). Thus, the intermingling between CENPB-sat and 2PI arrays suggests that these two satellite DNA families may be spatially related and confined to the pericentromeric regions of the horse. This hypothesis is further supported by the results presented in the following paragraphs.

We previously showed that the major centromeric satellite DNA family of the horse, 37cen, does not contain any CENP-B box. However, a 224 bp fragment of CENPB-sat which does not contain the CENP-B box shares 72% identity with the 221 bp unit of 37cen (Figure 10). It is therefore tempting to hypothesize a common evolutionary origin between these two different satellite DNA families. On the contrary, no relationship with other known equid satellite DNA families was detected.

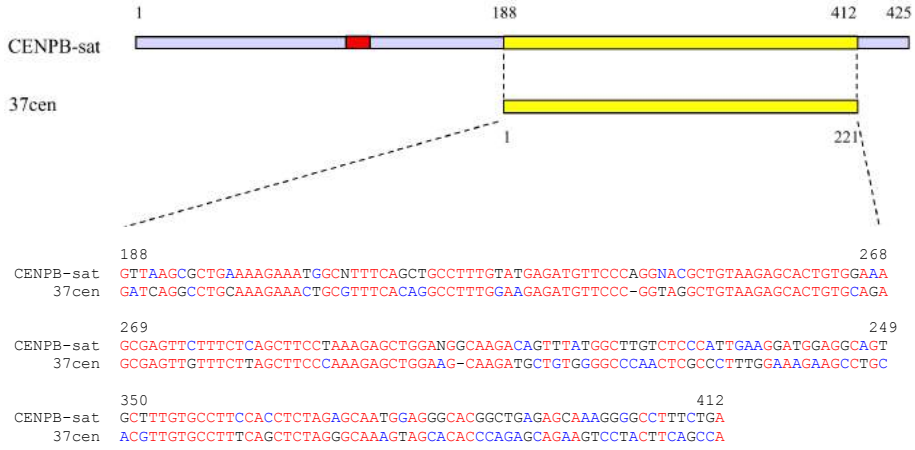


Figure 10. Identity region between CENPB-sat and 37cen. In the upper part of the figure, schematic representation of CENPB-sat with the CENP-B box (red) and 37cen; the identity region is the yellow one. In the low part of the figure, alignment between the identity regions of CENPB-sat and 37cen. Coordinates of CENPB-sat are reported. High consensus nucleotides and low consensus nucleotides are shown in red and blue, respectively.

4.2. Genomic abundance of CENPB-sat

ChIP-seq experiments with the same anti-CENP-B antibody were performed on the chromatin extracted from skin primary fibroblasts of *E. asinus*, *E. grevyi* and *E. burchelli*. We could evaluate the genomic abundance of CENPB-sat with respect to the other satellite DNA families among the four species by comparing normalized read counts in the input DNA (Figure 11). Input reads from each species were thus aligned on the horse CENPB-sat, 2PI and 37cen sequences. As control, we used the ERE-1 retrotransposon, which

is interspersed throughout the equid genomes and is expected not to be enriched (Cerutti et al. 2016, Nergadze et al. 2018).

The results of this analysis show that the CENPB-sat satellite is present in all the species but with very different genomic abundance. The Grevy's zebra is the species with the highest genomic representation of this satellite, followed by the horse. In the donkey and Burchell's zebra genome, CENPB-sat is very poorly represented (Figure 11).

Furthermore, we detected differences in the genomic abundance of both 37cen and 2PI, confirming our previous cytogenetic results (Piras et al. 2010). Indeed, 37cen is highly represented in the genomes of the horse, where it was shown to be associated with centromere competence (Cerutti et al. 2016), and well represented in the donkey genome. On the contrary, it is nearly absent in both Grevy's and Burchell's zebra. On the other hand, 2PI is the most abundant satellite DNA family in all the four species, but its levels progressively decrease from the horse, to the donkey, to the Grevy's zebra and finally to the Burchell's zebra (Figure 11). As expected, the genomic abundance of ERE-1 retrotransposon was similar in the four species (Figure 11).

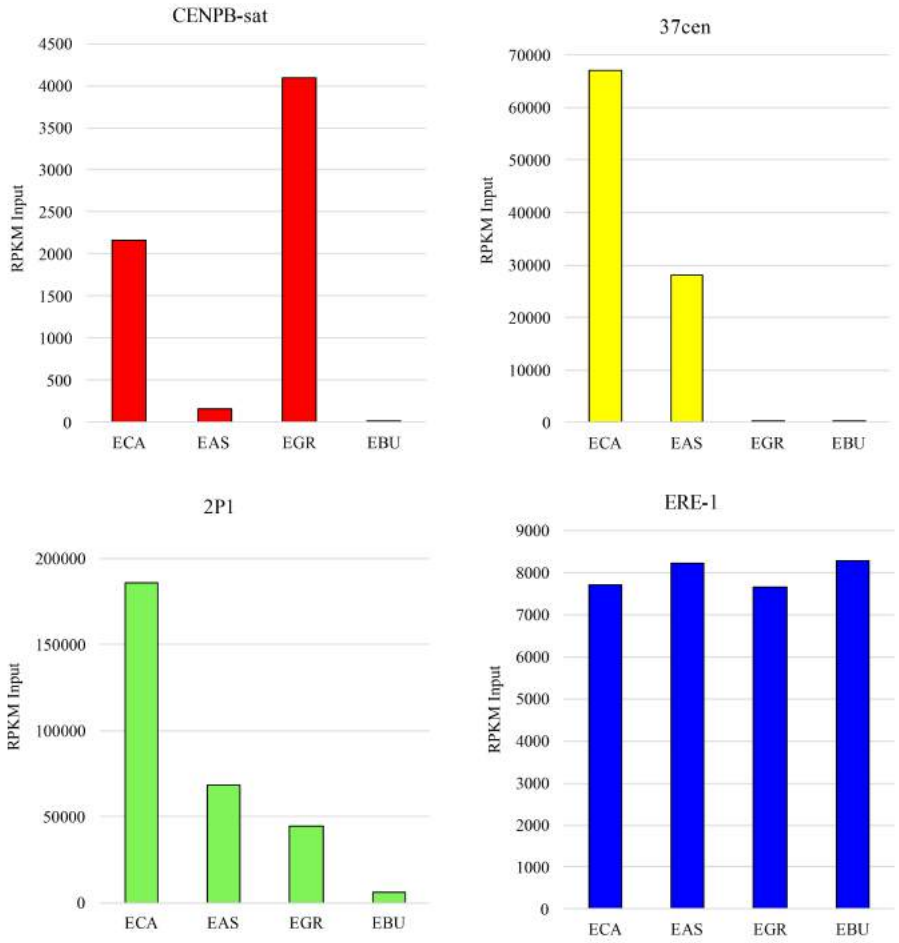


Figure 11. Genomic abundance of the different satellite DNA families in in horse (ECA), donkey (EAS), Grevy's zebra (EGR) and Burchell's zebra (EBU). ERE-1 retrotransposon is used as control.

4.3. CENPB-sat is the CENP-B bound satellite

To test whether CENPB-sat, which contains the CENP-B box, is actually bound by CENP-B, we aligned the reads from input DNA and immunoprecipitated DNA from each species to the horse CENPB-sat sequence. As control, we used the ERE-1 retrotransposon, which is

interspersed throughout the equid genomes and is expected not to be enriched (Cerutti et al. 2016, Nergadze et al. 2018).

The CENPB-sat satellite is enriched in all the immunoprecipitated samples, demonstrating that it is bound by CENP-B in all these species (Table 1). However, it should be noted that the enrichment observed in the donkey and the Burchell's zebra is based on a small number of reads (Figure 11). Conversely, ERE-1 was equally represented in the immunoprecipitated and in the input DNA, as expected (Table 1).

	CENP-B bound chromatin			
	E. caballus	***E. asinus***	***E. grevyi***	***E. burchelli***
CENPB-sat	6.3	1.8	6.8	4.1
ERE-1	0.9	1.1	1.1	1.2

Table 1. Fold enrichments of CENPB-sat in CENP-B bound chromatin of horse (*E. caballus*), donkey (*E. asinus*), Grevy's zebra (*E. grevyi*) and Burchell's zebra (*E. burchelli*). Enrichment values were measured as the ratio between normalized read counts (RPKM) in immunoprecipitated and input DNA. ERE-1 retrotransposon is used as control.

The low enrichment of CENPB-sat in the donkey immunoprecipitated chromatin could be due to the fact that only a fraction of the small number of the CENPB-sat copies is bound by CENP-B, presumably because of sequence degeneration that impaired the recognition by CENP-B.

To test this hypothesis, we evaluated the conservation of the CENP-B box in the genomes of the four species. We deduced a consensus of the CENP-B box sequence starting from the Input reads mapping on the horse CENPB-sat. As shown in Figure 12A, in the horse the CENP-B box is highly conserved, meaning that the majority of boxes found in the horse genome are functional. In the Grevy's zebra and in the Burchell's zebra, the box is mainly conserved and a few mutations are observed in essential nucleotides. On the other hand, in the donkey the box is mainly mutated in two essential nucleotides (C4>T and C13>T). However, as expected, the canonical box is the only one enriched in the CENP-B bound chromatin (Figure 12B), suggesting that sequence degeneration could be the cause of the poor binding of CENP-B to CENPB-sat.

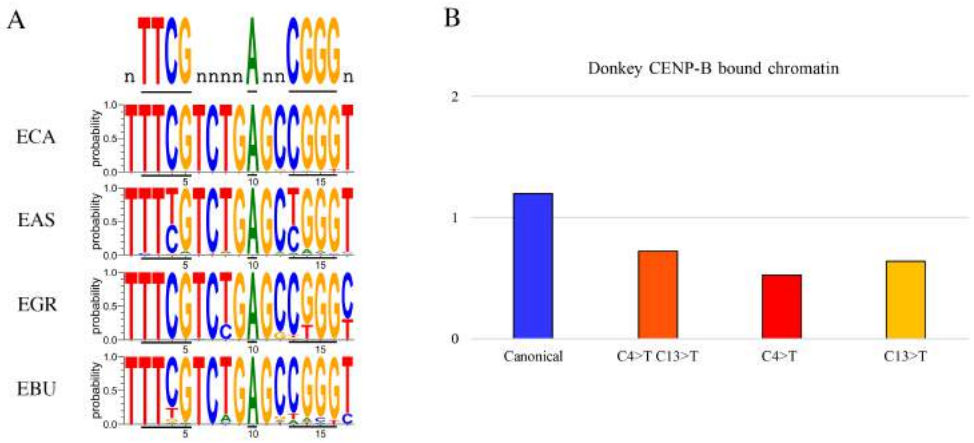


Figure 12. Conservation of the CENP-B box in the genomes of the four species. A) Logos of the consensus sequence of the CENP-B box in the genomes of horse (ECA), donkey (EAS), Grevy's zebra (EGR) and Burchell's zebra (EBU). In the upper part of the panel, a logo showing the 9 essential nucleotides for CENP-B binding is shown. B) Enrichment of the four variants of the CENP-B box in the CENP-B bound chromatin in donkey. The enrichment was measured as ratio between normalized read counts in immunoprecipitated and input DNA.

It is important to remind that CENPB-sat was identified from the horse genome, which is the only well assembled genome among equids. However, it is well known that satellite sequences are extremely divergent, even among closely related species. Although CENPB-sat was demonstrated to be bound by CENP-B in all the four equids, we could not exclude that the non-caballine species could have another satellite containing the CENP-B box which is not present in the horse.

To this end, for each non-caballine species, we *de novo* assembled the sequence environment around the conserved CENP-B box using raw ChIP reads (see Materials and Methods). We could derive an unbiased species-specific consensus of the sequence environment around the CENP-B box. In particular, we could assemble a 153 bp consensus for the Grevy's zebra, a 154 bp consensus for the Burchell's zebra and a 118 bp consensus for the donkey. All the derived species-specific consensus sequences match to the horse CENPB-sat (86%, 86% and 88% identity for *E. asinus*, *E. burchelli* and *E. grevyi*, respectively). This strategy allowed us to exclude that these equid species carry a different species-specific CENPB-sat.

In conclusion, CENPB-sat is the satellite bound by CENP-B in all the four species. However, it is important to underline that, since the copy number

and conservation of this sequence is highly variable, the amount of DNA bound CENP-B protein varies from very high (horse and Grevy's zebra) to undetectable (donkey and Burchell's zebra) in immunofluorescence experiments.

4.4. Functional annotation of CENPB-sat

Previous results from other species demonstrated that the presence of the satellite embedding the CENP-B box was coupled with the centromeric function (Masumoto et al. 1989, Kipling et al. 1995, Kipling and Warburton 1997). To test whether this is true for equid CENPB-sat, we took advantage of ChIP-seq experiments with an antibody raised against the centromeric protein CENP-A that we previously performed on the same cell lines (Nergadze et al. 2018, PhD thesis by Francesco Gozzo). In horse, Grevy's zebra and Burchell's zebra, the enrichment of CENPB-sat is very low (Table 2). Conversely, CENPB-sat is highly enriched in the CENP-A bound chromatin of the donkey. It is important to remember that the genomic amount of CENPB-sat is very low in this species. The high enrichment of CENPB-sat in the CENP-A bound chromatin means that a high fraction of the few copies of CENPB-sat resides in centromeric cores.

37cen is enriched in the horse, confirming our previous results (Cerutti et al. 2016), and in non-caballine species (Table 2). It is important to remind that the genomic amount of 37cen is extremely low in the two zebras, therefore the high enrichment of 37cen means that the few copies of 37cen mainly localize in the centromeric core. On the other hand, the enrichment of 2PI in the CENP-A bound chromatin is very low in the four species. Regarding the horse, this result is in agreement with the notion that 2PI resides mainly at pericentromeric positions (Cerutti et al. 2016). As far as the non-caballine species are concerned, 2PI is observed both at non centromeric and centromeric positions, in accordance with the low enrichment values of 2PI in these species. As expected, ERE-1 retrotransposon is not enriched in all the immunoprecipitated samples.

	CENP-A bound chromatin			
	E. caballus	***E. asinus***	***E. grevyi***	***E. burchelli***
CENPB-sat	1.4	17.4	1.7	1.4
2PI	1.5	1.2	1.3	1.1
37cen	1.7	2.1	16.0	16.9
ERE-1	1.1	1.0	1.0	1.0

Table 2. Fold enrichments of CENPB-sat in CENP-A bound chromatin of horse (*E. caballus*), donkey (*E. asinus*), Grevy's zebra (*E. grevyi*) and Burchell's zebra (*E. burchelli*). Enrichment values were measured as the ratio between normalized read counts (RPKM) in immunoprecipitated and input DNA. ERE-1 retrotransposon is used as control.

Centromeric and pericentromeric satellites are transcribed in a number of species from yeast to mammals (Rošić and Erhardt 2016). In agreement with this notion, we demonstrated that the horse major centromeric satellite 37cen is transcribed as well (Cerutti et al. 2016). We evaluated the transcription of CENPB-sat in comparison with the other satellite DNA families by analyzing the RNA-seq data obtained from total RNA of horse and donkey primary fibroblast cell lines and described in the attached paper (Cerutti et al. 2016). As shown in Figure 13, differently from 37cen and 2PI, very few reads corresponding to CENPB-sat transcripts were observed.

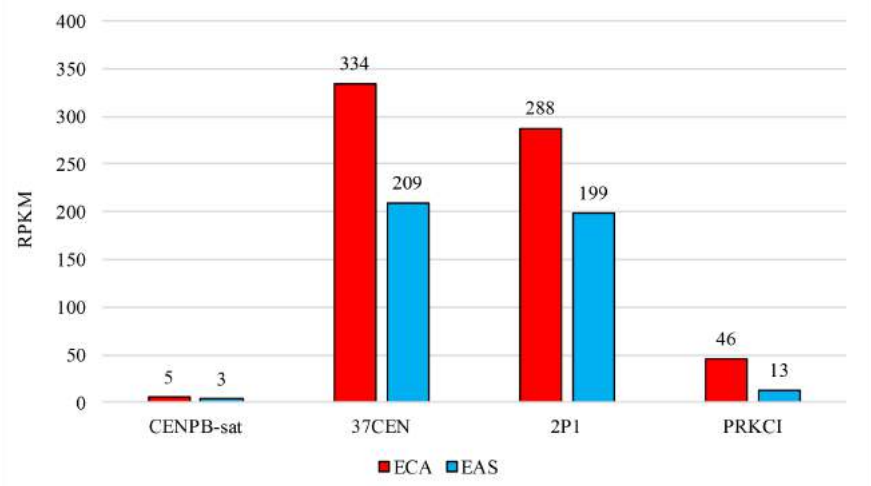


Figure 13. Transcription of CENPB-sat, 37cen and 2PI. Expression values of CENPB-sat, 37cen and 2PI and protein kinase C iota (PRKCI) are reported as RPKM.

It is well known that the centromeric and pericentromeric chromatin is characterized by specific histone modifications that were suggested to mediate centromere specification and kinetochore assembly (Fukagawa 2017). We characterized the epigenetic profile of CENPB-sat in comparison with 37cen and 2PI satellite, taking advantage of a panel of ChIP-seq experiments performed in our laboratory with antibodies against different chromatin markers (H3K9me3, H3K4me3, H3K4me2, H3K36me2, H4K20me1 and RNA polymerase II) on a horse fibroblast cell line (PhD thesis by Riccardo Gamba, PhD thesis by Marco Corbo). As reported in Table 3, all the satellite DNA families are characterized by the H3K9me3 heterochromatic signature and by low, if any, enrichment of markers associated to transcriptionally permissive and active chromatin (H3K4me3, H3K4me2, H3K36me2, H4K20me1, RNA polymerase II).

	H3K9me3	H3K4me3	RNApolII	H3K4me2	H3K36me2	H4K20me1
CENPB-sat	4.3	1.3	0.9	0.2	0.1	0.9
37cen	3.8	1.3	0.9	0.2	0.1	1.0
2PI	5.2	1.6	1.1	0.2	0.1	1.1
ERE-1	1.0	1.2	1.3	0.6	1.3	1.2

Table 3. Enrichment values of different satellite families in different epigenetic modifications of chromatin in horse. Enrichment values were measured as the ratio between normalized read counts (RPKM) in immunoprecipitated and input DNA. ERE-1 is used as control.

4.5. Chromosomal localization of CENPB-sat

As mentioned above, satellite DNA sequences are not completely placed in the genome assemblies or frequently misassembled. Thus, we could not derive any information on the chromosomal localization of CENPB-sat from ChIP-seq data. To this end, we cloned the 246 nucleotides of the repeat unit, containing the CENP-B box and lacking the identity region with 37cen (Figure 10), in a plasmid vector. The plasmid was used as probe in FISH experiments on metaphase chromosomes from *E. caballus*, *E. asinus, E. grevyi* and *E. burchelli* (Figure 14).

In *E. caballus*, CENPB-sat was localized at the primary constriction of five meta- or submeta-centric chromosomes (2, 6, 8, 10 and X) and sixteen acrocentric chromosomes (14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 27, 28,

29, 30, 31) (Figure 14A and 15A). In addition, we performed two-color FISH experiment on metaphase chromosomes with CENPB-sat and total horse genomic DNA as probes. As demonstrated in Piras et al. 2010, total genomic DNA labels all the satellite DNA loci, in particular all the centromeres with the exception of the one of ECA11. The intensity of these signals is always high with the exception of ECA2. On the contrary, on this chromosome, the CENP-B signal is very strong (Figure 16A and 17).

Since the number of chromosomes carrying CENPB-sat exceeds the number of CENP-B positive chromosomes identified by immunofluorescence, we performed immuno-FISH experiments with the anti-CENP-B antibody and CENPB-sat as probe. All the above described eighteen CENP-B signals colocalized with CENPB-sat FISH signals. However, on some centromeres, we could detect CENP-B sat signals only, indicating the lack of detectable CENP-B binding at these loci (Figure 18A and B). It is important to note that in this immuno-FISH experiment some of the signals previously observed in FISH experiments (Figure 14A and 15A) were not detected due to the differences in sensitivity and resolution between FISH and immuno-FISH.

In *E. asinus*, hybridization signals of CENPB-sat were detected only at the primary constriction of chromosome 3 (Figure 14B and 17). CENPB-sat FISH signals are very weak and thus images were collected with a higher exposition compared to the horse. The two-color FISH experiment with CENPB-sat and total donkey genomic DNA showed that, similarly to the horse, the EAS3 centromere carries a well-defined signal of CENPB-sat whereas the genomic DNA signal at this locus is very faint compared to the other ones. This observation confirms that the genomic abundance of this satellite family is very low in this species. It is worth remembering that in the donkey we could not detect any CENP-B signal by immunofluorescence (Paragraph 3), suggesting that at the EAS3 CENPB-sat locus many CENP-B box are mutated and thus not functional in recruiting CENP-B at sufficient levels to be detected by immunofluorescence (Figure 16B and 17).

In *E. grevyi*, CENPB-sat was located at one non-centromeric end of twelve meta- or submeta-centric chromosomes (1p, 2p, 5p, 6p, 7p, 8p, 10p, 12p, 13p, 14p, 15p and 16p) and four acrocentric chromosomes (19q, 20q, 21q and 22q) and at the centromeric region of metacentric chromosomes 6 and 12. In addition, polymorphism regarding the intensity of CENPB-sat signals was detectable between the two homologs in several chromosome pairs (EGR2, EGR13, EGR16 and EGR19). In particular, in chromosomes 13 and 19, we could detect the hybridization signal on only one homolog of the

chromosome pair (Figure 15B). These results fit with immunofluorescence data, where we could observe polymorphism for the presence of CENP-B signals at EGR2, EGR13 and EGR16 (Figure 8). Differently from the horse and the donkey, the two-color FISH experiment showed perfect colocalization between the zebra genomic DNA signals and the ones of CENPB-sat. In other words, differently from the other species, there were no genomic DNA signals without underlying CENPB-sat signals, thus confirming the high genomic abundance of this satellite DNA family in this species (Figure 16). In addition, it is worth noticing that in EGR14 two extended blocks of satellite DNA (genomic DNA hybridization) are present at the p terminus and CENPB-sat signal overlaps only with the most telomeric one (Figure 17).

As for *E. caballus*, the number of CENPB-sat carrying chromosomes exceeds the number of chromosomes identified by immunofluorescence (Paragraph 3). Immuno-FISH experiments showed a few hybridization signals without underlying detectable CENP-B binding (Figure 18C and D). As for *E. caballus*, the number of CENPB-sat hybridization signals is lower than that previously detected by FISH.

In *E. burchelli*, no hybridization signal of CENPB-sat was detected at all, while signals following hybridization with genomic DNA were detected as already described in Piras et al. 2010, confirming the extreme paucity of this satellite family within the genome of the Burchell's zebra (Figure 14D and 16D).

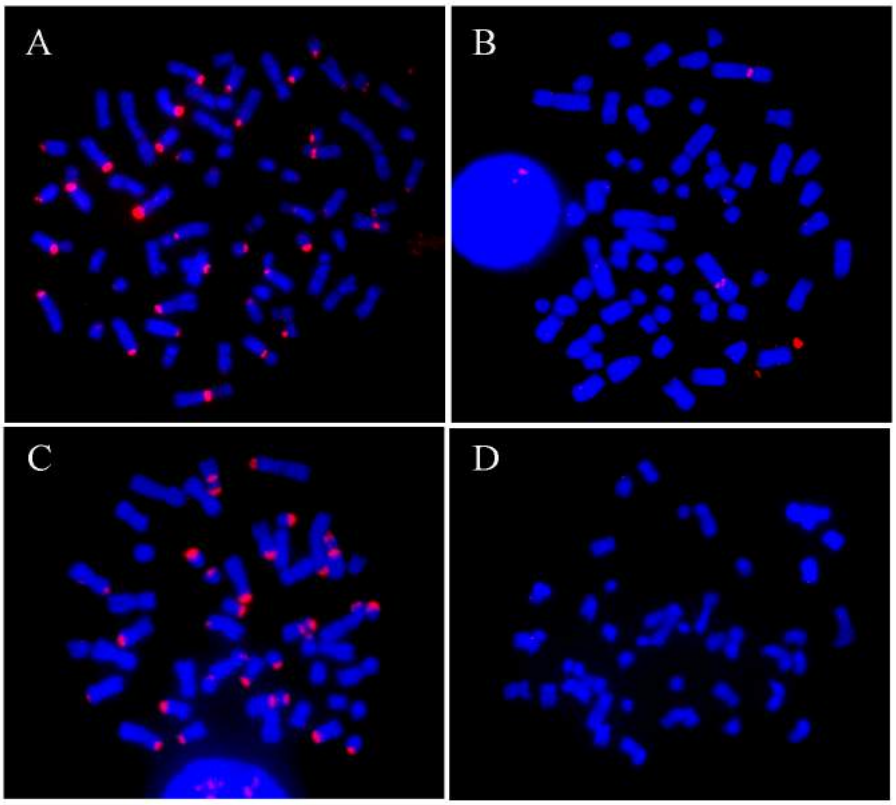


Figure 14. FISH localization of CENPB-sat (red) in metaphase chromosomes from horse (A), donkey (B), Grevy's zebra (C) and Burchell's zebra (D).

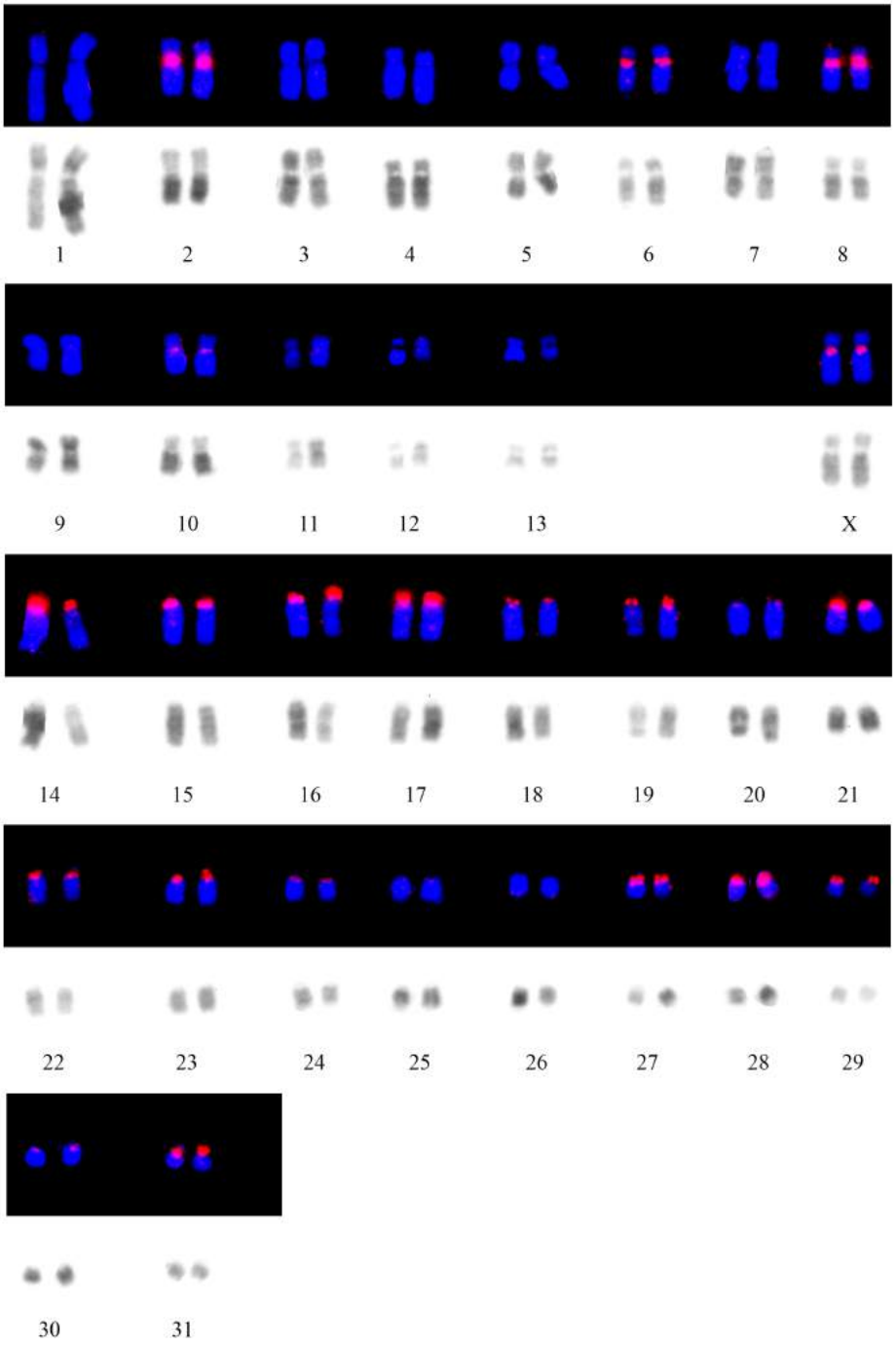


Figure 15A. Localization of CENPB-sat on horse chromosomes.

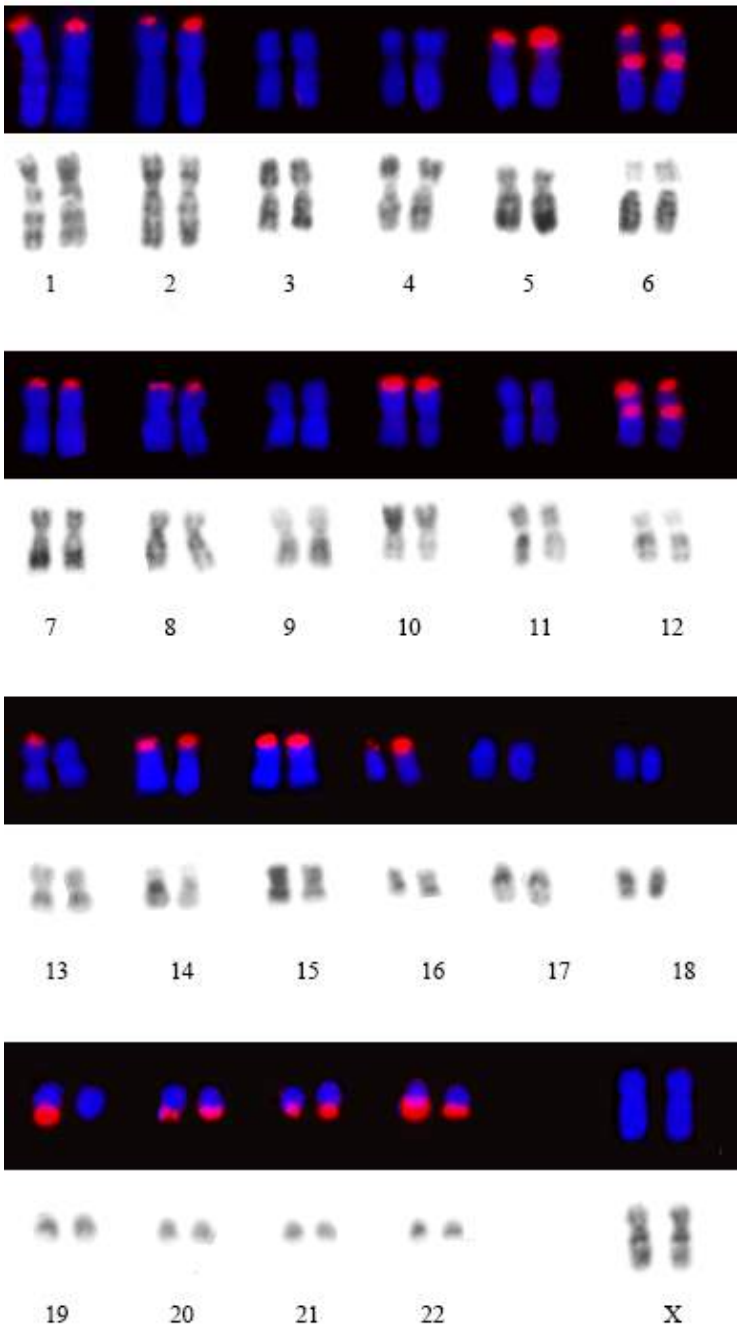


Figure 15B. Localization of CENPB-sat on Grevy's zebra chromosomes

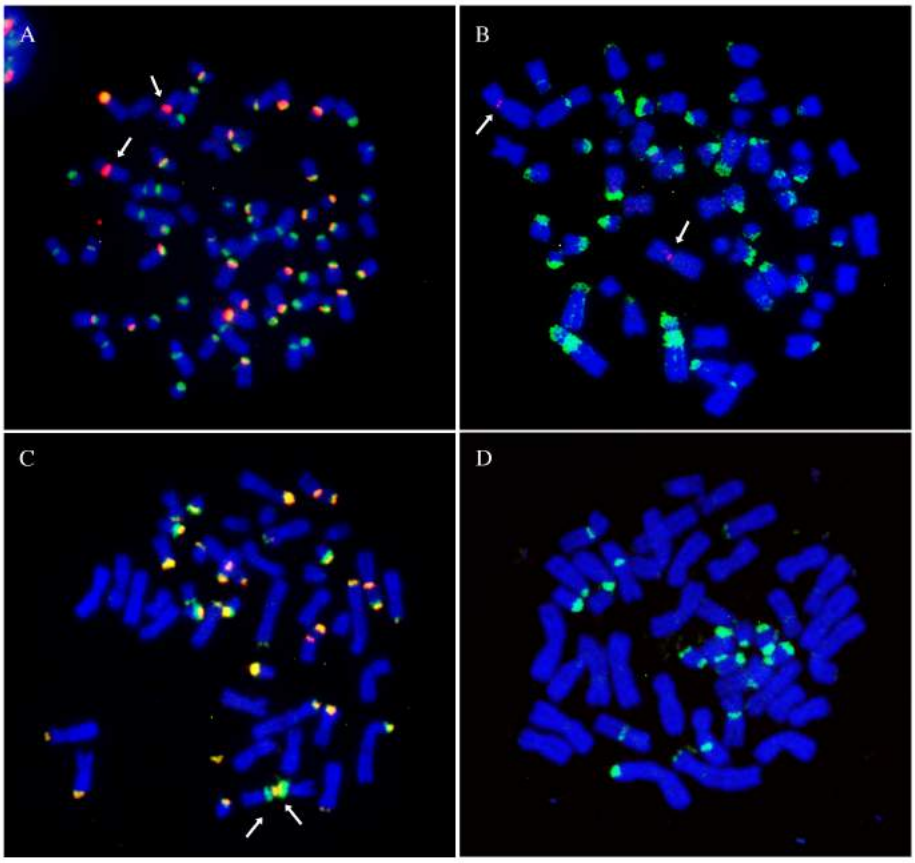


Figure 16. FISH localization of total genomic DNA (green) and CENPB-sat (red) on metaphase chromosomes from horse (A), donkey (B), Grevy's zebra (C) and Burchell's zebra (D). White arrows point ECA2 (A), EAS 3 (B) and EGR14 (C) chromosomal pairs.

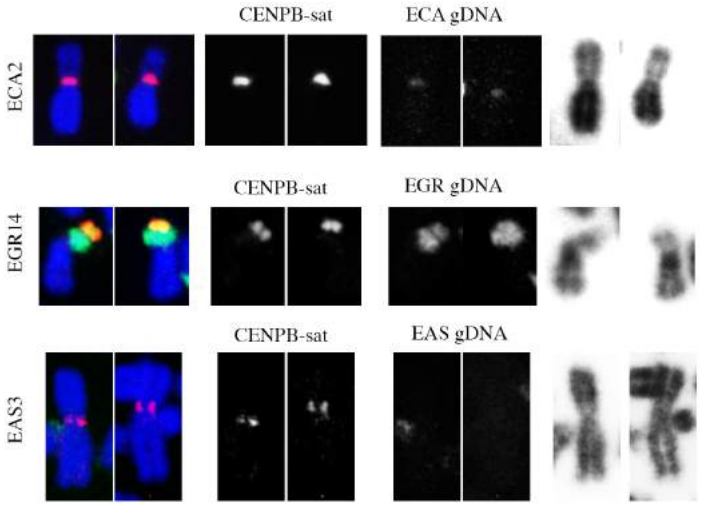


Figure 17. Total genomic DNA signals (green) and CENPB-sat signals (red) on ECA2, EGR14 and EAS3.

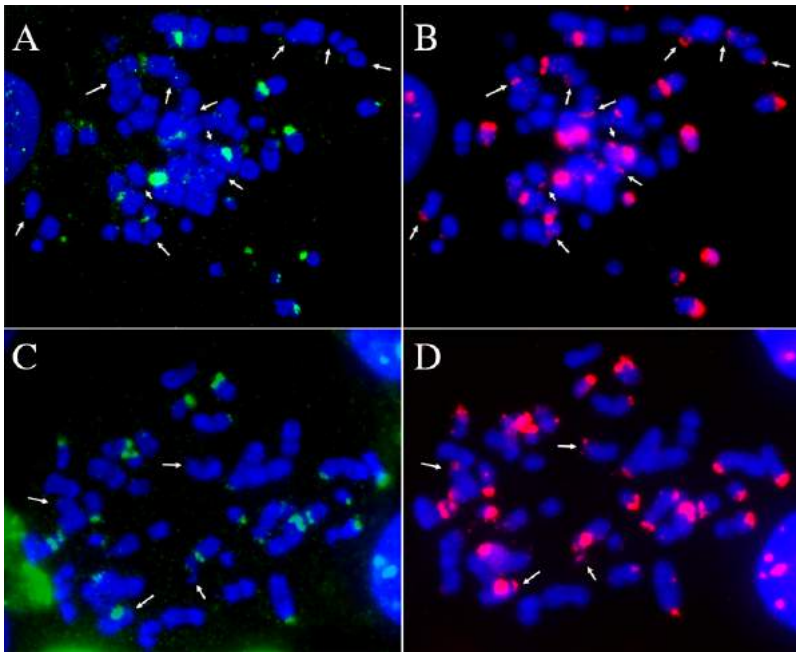


Figure 18. Localization of CENP-B protein and CENPB-sat on horse (A-B) and Grevy's zebra (C-D) metaphase chromosomes. On the left (A-C), CENP-B signals (green) on pseudocolored chromosomes of horse (A) and Grevy's zebra (C). On the right (B-D), the same metaphase after hybridization with CENPB-sat probe (red). Immunofluorescence signals could not be detected after FISH. White arrows point to chromosomes with CENPB-sat signal but without CENP-B signals.

Discussion

CENP-B is the only known centromeric protein that exhibits unequivocal DNA binding specificity, mediating the recognition of a target site, named the CENP-B box, which comprises nine essential nucleotides correctly spaced within a 17 bp motif (5' nTTCGnnnnAnnCGGGn 3'). The CENP-B binding site represents the only common motif shared by the highly divergent centromeric satellites of mammals. This extreme conservation reflects the total conservation of the functional domains of CENP-B, namely the N-terminal DNA binding domain and the C-terminal dimerization domain. In spite of the high conservation of CENP-B and its binding site, this protein appears dispensable for centromeric function. This "CENP-B paradox" is still an open issue and the role of CENP-B in the epigenetic establishment of centromeric chromatin remains controversial.

In order to shed light on the role of this elusive protein, we investigated the binding pattern of CENP-B in the genus *Equus*, given the extraordinary plasticity of equid centromeres. Indeed, exceptionally frequent centromere repositioning events mark the rapid karyotypic radiation during the phylogeny of equid. The result is that equids are characterized by centromeres at different maturation stages, ranging from classical satellite-based centromeres to "immature" satellite-less centromeres. In addition, blocks of satellite DNA uncoupled from centromeric function are frequently present at chromosomal termini, representing relics of ancestral inactivated centromeres or traces of satellite loci exchange during karyotype rearrangements. Since it is well described in the literature that CENP-B binding sites are comprised in centromeric satellites, the genus *Equus* provide the opportunity to evaluate the association between CENP-B, centromeres and satellites.

This work focuses on *E. caballus* (horse), *E. asinus* (domestic donkey), *E. grevyi* (Grevy's zebra) and *E. burchelli* (Burchell's zebra). These species were chosen as representatives of different scenarios of uncoupling between satellite DNA and centromeric function and karyotypic rearrangements. In the horse (*2n*=64), all the centromeres, with the exception of the one of chromosome 11, are satellite-based and the major centromeric satellite families is 37cen (Wade et al. 2009, Piras et al. 2010, Cerutti et al. 2016). In the donkey (*2n*=62), an extraordinary high number of satellite-less centromeres is present (16 out of 32), while satellite DNA loci are either centromeric or non centromeric (Piras et al. 2010, Nergadze et al. 2018). A high number of chromosomal fusion events led to the highly similar

karyotypes of Grevy's zebra (*2n*=46) and Burchell's zebra (*2n*=44) and, according to the available comparative karyotypes, the two species share 13 autosomal chromosomal pairs. They are characterized by several satellite-less centromeres as well as non centromeric satellite DNA loci (Musilova et al. 2007, Piras et al. 2010, Musilova et al. 2013). However, the distribution of satellite DNA loci strongly differs between the two species: in the Grevy's zebra, the majority of satellite DNA loci are found at non centromeric chromosomal termini, while in the Burchell's zebra satellite DNA is mainly present at satellite-based centromeres or at subcentromeric regions with only few non centromeric loci at terminal positions (Piras et al. 2010). Thus, these species represent four different scenarios for unraveling the role of CENP-B and the evolution of its binding sites.

1. Equid CENP-B proteins are functional and can recognize a canonical CENP-B box

The analysis of the coding sequences of CENP-B in four equid species, namely *E. caballus* (horse), *E. asinus* (domestic donkey), *E. grevyi* (Grevy's zebra) and *E. burchelli* (Burchell's zebra), confirmed the extreme conservation of this protein among mammals. In particular, the two functional domains of CENP-B – the N-terminal DNA binding domain and the C-terminal dimerization domain – are totally identical to those of the other mammalian species studied so far, suggesting that in the genus *Equus* CENP-B is functional and able to dimerize and bind a canonical CENP-B box.

Besides the functional domains, the entire protein appeared highly conserved among the four species, since only three amino acid differences were detected. The first one was an alanine-to-valine substitution in the endonuclease domain of the donkey with respect to the other equid species as well as human and mouse protein (Earnshaw et al. 1987, Sullivan and Glass 1991). Previous data from our laboratory showed that this mutation is present also in the two Asiatic asses *Equus hemionus onager* and *Equus kiang* (Master thesis by Demetrio Turati), suggesting that this mutation probably arose during the divergence of the lineage of asses. Although alanine and valine have different structural properties, reciprocal substitutions of these amino acid are tolerated and buffered by tertiary interactions in the overall structure of alpha helices (Gregoret and Sauer 1998), suggesting that this change does not impact the functionality of the protein. In addition, this substitution occurs in the endonuclease domain, which is likely to be inactive,

being considered only a trace of the evolutionary history of this protein (Marshall and Choo 2012). Thus, this domain is supposed to be prone to accumulate mutations because amino acid changes should not alter the overall function of the protein.

The other two differences concern the first acidic domain. This gene region is rich in GAG and GAA triplets coding for glutamate residues that are organized in short tandem repeats. In particular, *E. caballus* and *E. asinus* miss a glutamate residue at different positions with respect to the two zebras, because of the deletion of a codon in a GAG stretch. Previous data from our laboratory demonstrated that different species of mammals are characterized by a surprising high variability in the length of GAG and GAA clusters, leading to the hypothesis that this region might be subjected to copy number variation due to DNA polymerase slippage during DNA replication (Master thesis by Demetrio Turati). Moreover, the human protein, which consists of 599 amino acids, is 7-8 amino acid shorter than the equid CENP-B proteins. These additional residues are aspartate or glutamate residues of the two acidic domains. Interestingly, the first acidic domain is involved in the interaction with CENP-C and we might hypothesize that the length variability of the acidic domain among mammals could be an evolutionary driving force to modulate the interaction network of CENP-B. Structural studies will be required to test whether this variability affects protein functionality.

We can conclude, from sequence data, that horse and the donkey CENP-B are likely to be functional and should recognize a canonical CENP-B box, according to literature data.

In addition, CENP-B is expressed in all the four species but, whereas the horse and donkey showed similar amount of expressed protein, a reduction of CENP-B expression is observed in the Burchell's zebra and an increase of CENP-B expression is found in the Grevy's zebra.

2. Peculiarities of CENP-B binding in the genus *Equus*

It is well described in literature that CENP-B boxes are contained in centromeric satellites and represent the only common motif shared by these highly divergent sequences (Kipling et al. 1995, Kipling and Warburton 1997). Conversely, these motifs have never been detected in "single-copy" clinical neocentromeres, confirming the association between CENP-B binding sites and centromeric satellite families (Choo 2000, Saffery et al. 2000, Amor and Choo 2002).

Sequence analysis of the extraordinarily high number of equid satellite-less centromeres (one in the horse and 16 in the donkey) demonstrated that these centromeric domains lack any CENP-B recognition motif. Thus, we proved the absence of CENP-B on mammalian evolutionary satellite-free centromeres, confirming what had been previously described for clinical neocentromeres (Choo 2000, Saffery et al. 2000, Amor and Choo 2002).

However, we uncovered a surprising deficiency of CENP-B recognition motifs also in the main satellite DNA families of the genus *Equus*, namely 37cen and 2PI (Piras et al. 2010). Furthermore, 37cen is the major centromeric satellite of the horse, building the centromeric core of satellite-based centromeres in this species (Cerutti et al. 2016). To our knowledge, this is the first reported case of lack of the CENP-B box in a centromeric satellite DNA family.

On the other hand, in the horse genome assembly EquCab2 we identified a functional novel equid satellite DNA family which contains a functional CENP-B box in which the previously identified nine essential nucleotides for CENP-B binding are conserved (5' tTTCGtctgAgcCGGGt 3'). This satellite, termed CENPB-sat, is made by tandemly repeated 425 bp monomers, arranged in a head-to-tail fashion. Differently from the majority of centromeric satellites, it is not AT rich, since the content of AT is nearly equal to that of GC nucleotides. Sequence analysis suggested that this novel family is evolutionary related to 37cen, since a 224 bp fragment of CENPB-sat which does not contain the CENP-B box shares 72% identity with the 221 bp unit of 37cen. Thus, we might speculate a common evolutionary origin for the two satellites. Otherwise, CENPB-sat is not related at the sequence level with 2PI satellite family. Nonetheless, CENP-B and 2PI seem to be intermingled in the horse genome. Indeed, there are arrays or CENPB-sat, with either functional or degenerated CENP-B boxes, interrupted by stretches of 2PI satellite.

It should be noticed that all these satellite sequences share the same epigenetic landscape, in particular the H3K9me3 signature, confirming the heterochromatic environment of satellite DNA in mammals. However, RNA-seq analysis demonstrated that, differently from 37cen and 2PI (Cerutti et al. 2016), CENPB-sat is not transcribed in horse and donkey fibroblasts.

We demonstrated that the CENPB-sat is the one bound by CENP-B in *E. caballus*, *E. asinus*, *E. grevyi* and *E. burchelli*. However, its genomic abundance widely varies in these species, roughly recapitulating the differential expression of the protein. Indeed, the Grevy's zebra is the species

with the highest genomic abundance of this satellite, followed by the horse. The genomic amount is dramatically reduced in the donkey and in the Burchell's zebra, which is the species with the lowest amount of this satellite. In addition, although equid CENP-B proteins are predicted to be functional as in the other mammals, these species exhibit four different scenarios according to the binding pattern of this elusive protein and the distribution of its binding sites. Surprisingly, all these binding patterns differ from the typical mammalian one, where CENP-B localizes at all primary constrictions with the exception of Y chromosome, suggesting an unconventional role of CENP-B in the genus *Equus*.

1) In *E. caballus* (*2n*=64), CENPB-sat is not present at all centromeres but localizes, at FISH resolution, only at the primary constrictions of five meta- or submeta-centric chromosomes (ECA2, ECA6, ECA8, ECA10 and ECAX) and sixteen acrocentric chromosomes (ECA14, ECA15, ECA16, ECA17, ECA18, ECA19, ECA20, ECA21, ECA22, ECA23, ECA24, ECA25, ECA26, ECA29, ECA30 and ECA31). This satellite is not highly abundant in the horse genome. CENPB-sat is the one embedding the conserved CENP-B box and recognized by CENP-B, as demonstrated through ChIP-seq. Surprisingly, not all satellite loci are bound by CENP-B at such levels to be detected by immunofluorescence, suggesting that this satellite is undergoing sequence degeneration, losing the ability to be recognized by the protein. Indeed, CENP-B can be detected by immunofluorescence at the primary constriction of only 9 out of 32 chromosome pairs, suggesting that CENP-A and CENP-B are uncoupled in this species.

The wide and heterogeneous appearance of CENP-B signals, in contrast with the speckled pattern of CENP-A, suggests a broad binding domain comprising both the centromeric core and the pericentromere. In particular, CENP-B covers the whole primary constriction including the inner region, differently from CENP-A and CENP-C. The pericentromeric localization of CENP-B is demonstrated by the fact that the CENPB-sat is enriched at very low levels in the CENP-A bound chromatin. In addition, our previous data demonstrated that the centromeric core of satellite-based centromeres is made by arrays of 37cen, which is not bound by CENP-B because of the absence of binding sites. Thus, in the horse, the domains of CENP-A and CENP-B are separated and satellite-based centromeres are made by a core of 37cen, flanked by pericentromeric arrays of CENPB-sat intermingled with 2PI satellite. In addition, in this system, despite the lack of

detectable levels of CENP-B at the majority of centromeres, all the primary constrictions are marked by CENP-C and, at the immunofluorescence resolution, the intensity of CENP-C signals does not correlate with the intensity of CENP-B signals. Thus, our data contradict the hypothesis that CENP-B is directly involved in the targeting and maintenance of CENP-C at centromeres (Fachinetti et al. 2015), suggesting that its interaction with CENP-B is dispensable and not universal.

2) In *E. asinus* (*2n*=62), the situation completely changes. The majority of centromeres are devoid of satellite DNA and thus lack binding sites for CENP-B. Indeed, CENPB-sat is poorly represented in the donkey genome, while 2PI and 37cen are well represented both at satellite-based centromeres and at the remnants of ancestral centromeres. The CENPB-sat satellite sequence could be detected, at FISH resolution, only at the primary constriction of EAS3, where a faint signal could be distinguished. In the donkey lineage, this satellite probably underwent sequence degeneration: the CENP-B box is highly divergent and accumulated mutations in essential nucleotides (C4>T and C13>T substitutions). These variant boxes are likely to prevent CENP-B binding, since only the canonical non-mutated boxes are functional. The low enrichment of CENPB-sat in the CENP-B bound chromatin is a further evidence of the degeneration of this satellite and its progressive failure to recruit CENP-B: the majority of CENPB-sat loci are no more functional and only a minor fraction of these is still able to recruit the protein. These very few sites appear to be mainly present at satellite-based centromeres, given the enrichment of CENPB-sat in the CENP-A bound chromatin.

A loss of functional CENP-B binding sites is supported also by the absence of detectable CENP-B signals on donkey chromosome by immunofluorescence. We can reasonably hypothesize that the amount of bound protein is too low to be detected by immunofluorescence and a considerable fraction of protein remains unbound. Moreover, in the hinny, CENP-B signals are detected only on the chromosomes corresponding to the ones identified in the horse, while no chromosome deriving from the donkey parent was labeled. This observation confirmed that the DNA binding sites rather than the protein itself are responsible for lack of binding. Nonetheless, it should be remarked that the amount of CENP-B protein is similar in horse and donkey. Thus, it is still an open issue whether the unbound donkey protein could exert an additional function that is unrelated to the centromeric one.

3) In *E. grevyi* (*2n*=46), 2PI and CENPB-sat are the major satellite DNA families, while 37cen is very poorly represented. Surprisingly, CENPB-sat mainly resides at non centromeric positions, labeling a non-centromeric end of nine submeta- or meta-centric chromosomes (EGR1p, EGR2p, EGR5p, EGR7p, EGR8p, EGR10p, EGR14p, EGR15p and EGR16p) and of three acrocentric chromosomes (EGR20q, EGR21q and EGR22q). EGR6 and EGR12 show both the pter and the primary constriction labeled by CENPB-sat. EGR13p and EGR19q display a polymorphism, being only one homolog of the chromosome pair labeled by the satellite probe. Despite this non-centromeric localization, the Grevy's zebra is the non-caballine species with the highest conservation of the CENP-B box. Accordingly, the majority of CENPB-sat loci are bound by CENP-B, which is highly expressed compared to the other equid species. To our knowledge, this is the first report of such extreme spatial uncoupling between CENP-B and centromeric function, suggesting a different role for this elusive protein.

4) Although *E. burchelli* (*2n*=44) shares many karyotype rearrangements with *E. grevyi*, the scenario of CENP-B distribution is totally different, with interesting similarities with the donkey. In particular, the poor representation of the CENPB-sat is extreme in this zebra. On the other hand, the CENP-B box is highly conserved in the essential nucleotides for CENP-B binding while the sequence environment in which it is embedded is dramatically divergent, suggesting that this satellite is becoming more and more degenerated and being lost from its genome. In fact, CENPB-sat cannot be detected by FISH, suggesting that only degenerated remnants of this satellite still exist. In this species, the major satellite DNA family is indeed 2PI, although the total satellite DNA content of this genome is quite low, confirming previous cytogenetic data (Piras et al. 2010). As for the donkey, the paucity of CENP-B binding sites is reflected by the absence of detectable protein binding by immunofluorescence. It is worth noticing that this species is the one with the lowest detected protein expression, suggesting that the losing of CENP-B binding sites and the reduction of expression are two interconnected phenomena, and thus confirming CENP-B dispensability.

3. Satellite DNA and karyotype evolution in the genus *Equus*

According to the current model, the ancestral karyotype of Perissodactyla mainly comprises acrocentric chromosomes and, during the radiation of equids, many fusion and centromere repositioning events reshaped the karyotypes of these species (Trifonov et al. 2008, Musilova et al. 2013). *E. caballus* is considered the closest species to the equid ancestor, being characterized by the highest diploid number and the highest number of acrocentric chromosomes. However, fusion and centromere repositioning events occurred also in the horse, leading to the appearance of its metacentric chromosomes (Trifonov et al. 2008, Wade et al. 2009, Trifonov et al. 2012, Musilova et al. 2013). ECA11 is the most striking example, deriving from a centromere repositioning event which led to an immature centromere lacking satellite DNA (Carbone et al. 2006, Wade et al. 2009).

In *E. caballus*, CENPB-sat is mainly found at pericentromeric position of acrocentric chromosomes. Indeed 16 out of 18 acrocentric chromosome pairs carry this satellite DNA family, while only few meta- or submeta-centric chromosomes are labeled by CENP-B sat (5 out of 14 chromosomal pairs). It is important to remind that the horse acrocentric chromosomes are supposed to correspond to unaltered equid ancestral chromosomes. In *E. asinus*, EAS3, corresponding to ECA 2q+3q, is the only chromosome with CENPB-sat, at centromeric position. Interestingly, the syntenic group ECA 2q+3q is maintained in all the equid species with the exception of the horse, suggesting that in this case EAS3 reflects the ancestral configuration (Myka et al. 2003, Trifonov et al. 2008, Trifonov et al. 2012). In *E. grevyi*, the majority of chromosomes (16 out of 23 pairs) exhibits CENPB-sat at a non-centromeric chromosomal terminus. Comparing the position of the "ancestral" horse centromere to the CENPB-sat localization on the orthologous EGR chromosome, four different situations appear:

A) maintenance of CENPB-sat at the locus orthologous to the "ancestral" ECA centromere (EGR6cen, EGR12cen, EGR16pter and EGR8pter), supporting the knowledge that the blocks of satellite DNA at non centromeric position are the remnants of an ancient inactivated centromere (Piras et al. 2010). Figure 19A shows the EGR16 case;

B) presence of CENPB-sat at the opposite terminus with respect to the "ancestral" ECA centromere (EGR6pter, EGR20qter, EGR21qter and EGR22qter). Figure 19B shows the EGR22 example;

C) loss of CENPB-sat (ECA15, ECA16, ECA18, ECA19, ECA21, ECA22 and ECA25). According to Musilova et al. 2013, the corresponding

ancestral chromosomes underwent fusion events in the zebra lineage. We might hypothesize that CENPB-sat was lost during fusion. CENPB-sat is absent also in EGR3, which is co-linear with the above mentioned EAS3 in which CENPB-sat is localized at the primary constriction. Figure 19C shows the EGR11 case;

D) presence of CENPB-sat on the p terminus of some metacentric EGR chromosomes and absence on the ECA orthologs (EGR2, EGR5, EGR13, EGR14 and EGR15). Surprisingly, also the ECA orthologous chromosomes are metacentric. All the horse metacentric chromosomes, which derived from centromere repositioning events relative to the perissodactyl ancestral karyotype (Trifonov et al. 2012) (ECA1, ECA4, ECA7, ECA9, ECA11, ECA12, ECA13), lack CENPB-sat. This observation leads to the speculation that the EGR CENPB-sat loci might correspond to ancestral centromeres and suggests that these horse repositioned centromeres accumulated satellite DNA families other than CENPB-sat (Piras et al. 2010). Figure 19D shows the example of EGR15.

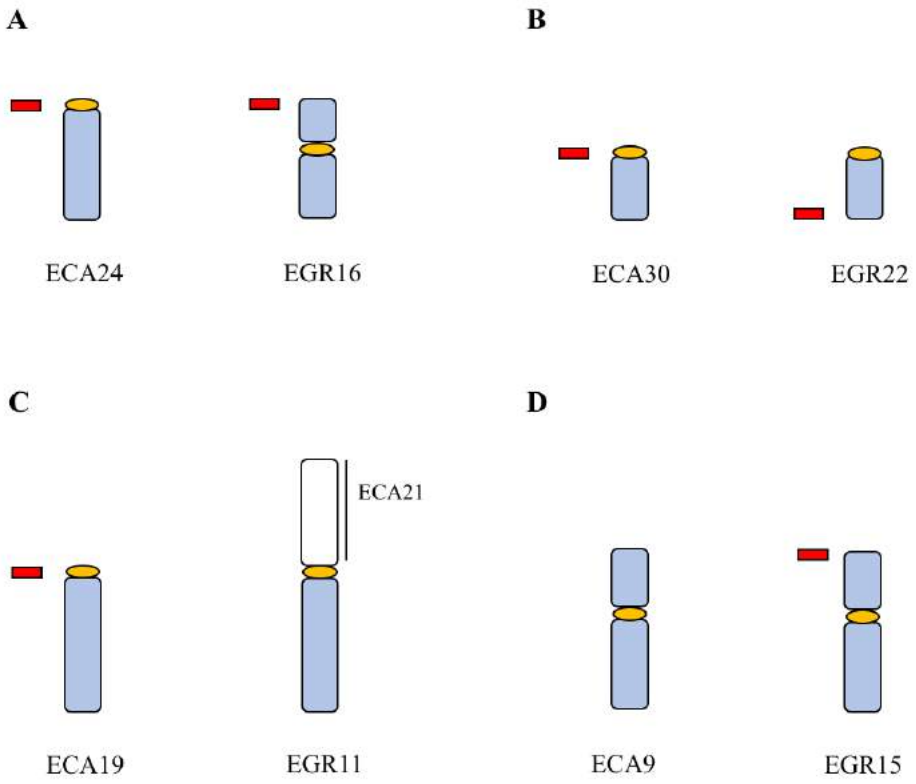


Figure 19. Position of CENPB-sat in ECA-EGR orthologous chromosomes. CENPB-sat is represented with a red rectangle. A) Example of maintenance of CENPB-sat at ancestral position in horse (ECA24) and Grevy's zebra (EGR16). B) Example of presence of CENP-B sat at discordant positions in horse (ECA30) and Grevy's zebra (EGR22). C) Example of loss of CENPB-sat from the ancestral position (ECA19) in Grevy's zebra (EGR11). D) Example of maintenance of CENPB-sat at a terminus of Grevy's zebra (EGR15) and loss in horse (ECA9).

It is tempting to speculate that a progressive uncoupling between CENP-B and centromeric function marked equid phylogeny. In accordance with the library hypothesis for satellite DNA evolution (Salser et al. 1976, Fry and Salser 1977), these equid species share a common set of satellite DNA families which underwent expansion or shrinkage in the different species. In Figure 20 a possible model is proposed. We might hypothesize that ancestral centromeric DNA was composed by arrays of CENPB-sat intermingled with 2PI stretches (Figure 20A). In this context CENP-B and CENP-A domains would have been overlapped, as for the other mammals. A trace of this arrangement is still found in ECA2, where CENPB-sat (Figure 14B and 17) and 2PI (Piras et al. 2010) reside at the primary constriction. During evolution, 37cen could have arisen through amplification of the portion of CENPB-sat which lacks the CENP-B box and also nowadays shares a high identity with 37cen. In the horse, these 37cen arrays would have been expanded in the centromeric core becoming able to recruit CENP-A, as previously demonstrated at the molecular level (Cerutti et al. 2016). Subsequently, CENPB-sat and 2PI would have been pushed out towards the pericentromere, leading to the uncoupling between CENP-A and CENP-B domains described above (Figure 20B). CENPB-sat is still present at the primary constriction of a subset of chromosomes but is becoming more and more divergent and thus losing the ability to recruit CENP-B (Figure 20C). Progressive sequence degeneration is demonstrated by the presence of CENPB-sat loci no more able to recruit CENP-B at the immunofluorescence level.

In the donkey and the two zebras, the uncoupling between CENP-A and CENP-B is extreme both for the reduction of CENP-B binding sites (*E. asinus* and *E. burchelli*) (Figure 20D) and for the spatial separation of CENPB-sat and centromeric function (*E. grevyi*) (Figure 20E and 20F). The extended and conserved CENPB-sat arrays of the Grevy's zebra at non centromeric sites could represent the relics of ancestral inactivated centromeres. In addition, exchange of satellite DNA between opposite chromosomal termini could have played a role in reshaping of the Grevy's zebra karyotype, since CENPB-sat loci are frequently found at opposite chromosomal terminus with respect to the ancestral centromere. This hypothesis is in agreement with the observation that in primates terminal centromeres can move from one chromosome end to the other and the extended heterochromatic blocks of satellite DNA could play a role in this rearrangement (Bailey et al. 2002, Ventura et al. 2004).

In the donkey and in the Burchell's zebra, the ancestral blocks of CENPB-sat could have been progressively eroded and only 2PI or 37cen remnants are still visible because of differential expansion in these species (Piras et al. 2010).

CENP-B has been proposed to be involved in centromere strength and stability (Fachinetti et al. 2015, Mohibi et al. 2015) and maintenance of pericentric heterochromatin, acting as a barrier against genome instability (Morozov et al. 2017). In addition, it was proposed that CENP-B plays a role in the maintenance of CENP-C at centromeres (Fachinetti et al. 2015). Our results are not in agreement with this hypothesis because CENP-B defective centromeres do not show any visible impairment in recruiting and maintaining CENP-A and CENP-C. Taking together our results, we might wonder whether the uncoupling between CENP-A and CENP-B could be a driver of the exceptionally frequent centromere repositioning events and chromosome rearrangements that occurred in this genus. Actually, we proved that the uncoupling between CENP-B and CENP-A does not affect the viability of the species but surprisingly all these species are characterized by an extraordinary centromere and karyotype plasticity, suggesting that their centromeres are in a dynamic state and thus prone to evolve rapidly.

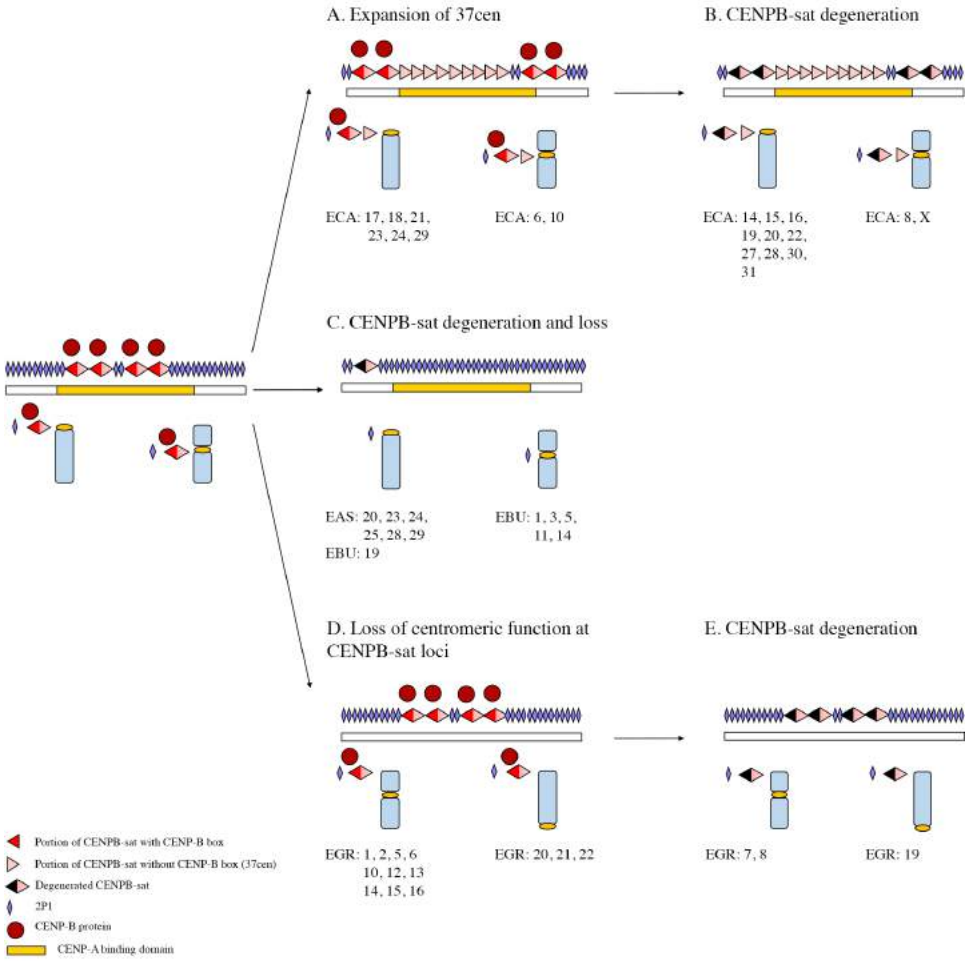


Figure 20. Model of CENPB-sat evolution and CENP-A/CENP-B uncoupling in the genus *Equus*. The different mechanisms that may have led to uncoupling between CENP-A and CENP-B are reported. The chromosomes which belong to the different categories are listed for each species.

Conclusions

We demonstrated that all the equid species express a functional CENP-B protein which recognizes a canonical CENP-B box. Surprisingly, the CENP-B binding site is not comprised in the equid major satellite DNA families, namely 37cen and 2PI, but resides in a novel satellite DNA family, the CENPB-sat satellite. Differently from the mammalian species studied so far, the CENPB-sat is not the major centromeric satellite and the genus *Equus* is characterized by marked uncoupling between CENP-B and CENP-A. In the horse, CENP-B domains are restricted to a subset of pericentromeres and are excluded from the centromeric core, becoming more and more degenerated and losing the ability to recruit CENP-B. In the donkey and the Burchell's zebra, the progressive reduction and degeneration of binding sites have led to the disappearance of detectable levels of CENP-B binding at all chromosomes. On the other hand, in the Grevy's zebra CENP-B is present only at non centromeric chromosomal termini, interpreted as the relics of ancestral inactivated centromeres, while CENP-B is undetectable at all active centromeres with the exception of two chromosomal pairs.

In conclusion, during the rapid karyotype evolution that marked equid phylogeny, different mechanisms of centromere maturation and satellite DNA evolution emerged in the different lineages, all resulting in the uncoupling between the centromeric function and CENP-B. We propose that this separation between CENP-A and CENP-B domains could be the reason for the exceptional plasticity of equid centromeres: CENP-B defective centromeres of the equid species could be in a more fragile equilibrium, compared to classical CENP-B coupled centromeres, leading to the evolutionary instability of the centromeric domains that characterized these species.

PART 4
BRIDGING TELOMERES, CENTROMERES AND CENP-B IN CHINESE HAMSTER

As mentioned in the Introduction, previous work from our laboratory proved that the Chinese hamster (*Cricetulus griseus)* genome is characterized by large clusters of the telomeric-like TTAGGG repeat which are localized at all the centromeres, with the exception of Y chromosome (Bertoni et al. 1996). Since TTAGGG repeats do not contain any CENP-B box, we speculated that CENP-B binding might be uncoupled from the centromeric function. It was also shown that a different satellite repeat composed of 33 bp units is localized at the primary constriction of chromosome 5 (Faravelli et al. 1998). Sequence analysis of this satellite revealed that it does not contain any CENP-B box. The aim of this work was to study CENP-B and its binding sites in Chinese hamster.

Results

1. CENP-B GENE AND PROTEIN SEQUENCE IN *Cricetulus griseus*

Sequence comparison between the human (P07199) and the Chinese hamster (P48988) CENP-B protein revealed 93% of identity. The CENP-B protein of Chinese hamster is 606 amino acid long, 7 amino acids longer than the human protein because of the different length of the two acidic domains (Figure 1). The CENP-B DNA binding domain (1-129) is totally conserved, like all mammalian species studied so far, suggesting that Chinese hamster CENP-B could recognize a canonical CENP-B box. However, an arginine-to-tryptophan substitution in the Chinese hamster compared to the human is observed at position 596 of the dimerization domain (Figure 1).

Figure 1. Alignment between human (HSA, P07199) and Chinese hamster (CGR, P48988) CENP-B proteins. The DNA binding domain, the endonuclease domain, the acidic domains and the dimerization domain are highlighted in yellow, grey, green and turquoise blue, respectively. Amino acid differences are shown using blue (high consensus color) or red (low consensus color). The amino acid difference in the dimerization domain is marked with an asterisk.

2. CHROMOSOMAL LOCALIZATION OF CENP-B IN A CHO CELL LINE

We investigated the localization of CENP-B in a Chinese Hamster Ovary cell line (CHO-PV, Bertoni et al. 1996) by immunofluorescence. To label all centromeres we used a human CREST serum. While CREST signals, as expected, were detected at the primary constriction of all chromosomes (figure 2A), CENP-B signals could be detected only at the primary constriction of 14 chromosomes out of 19 (figure 2B). It is important to underline that this CHO-PV cell line is characterized by a hemizygous karyotype, marked by dramatic rearrangements with respect to the normal diploid karyotype. In particular, the karyotype of CHO-PV comprises eight unrearranged chromosomes (one pair of 1, one pair of 5, one 2, one 4, one 8 and one 9), two chromosomes with simple rearrangements (3p- and 6q), five rearranged chromosomes of the Z set (Z2, Z4, Z7, Z8 and Z13) and four chromosomes of the R set (R1, R2, R3 and R4) (Bertoni et al. 1994, Bertoni et al. 1996). Chromosome identification by computer-generated reverse DAPI-banding showed that the pair of chromosomes 1 and chromosomes 2, Z2 and Z7 are not bound by CENP-B at levels detectable by immunofluorescence.

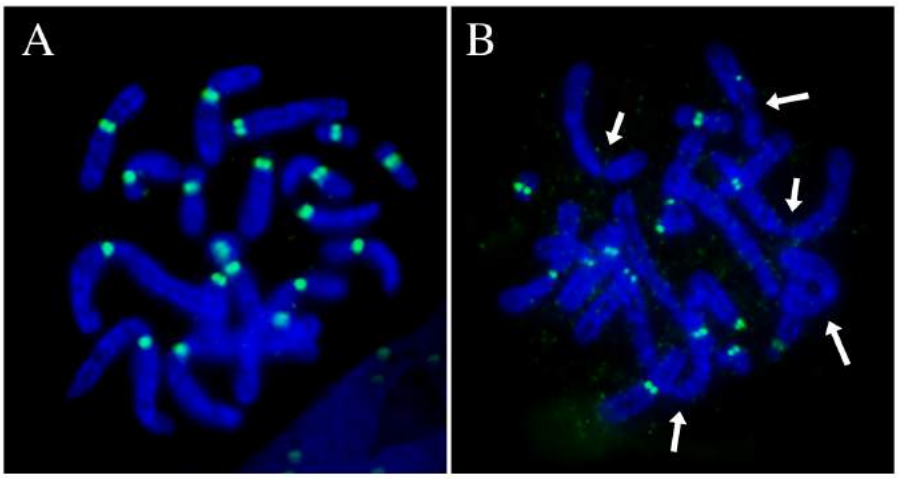


Figure 2. Immunoflurescence experiments with a CREST serum (A) and an anti-CENP-B (B) antibody and on CHO-PV metaphase chromosomes. A) CREST (green) signals on pseudocolored chromosomes (blue). CREST signals are detected at all primary constrictions. B) CENP-B (green) signals on pseudocolored chromosomes (blue). CENP-B signals can be detected only at the primary constriction of 14 chromosomes. 5 chromosomes (white arrows) lack CENP-B signals.

3. ChIP-seq IDENTIFICATION OF THE CENP-B BOUND SATELLITE

In order to characterize CENP-B binding sites, we adopted the same strategy used in *Equus caballus*. We performed a ChIP-seq experiment with an antibody against CENP-B on cross-linked chromatin extracted from CHO-PV cells, using the criGri1 assembly (C_griseus_v1.0) as reference genome. Differently from the EquCab2 assembly, criGri1 is a draft sequence comprising only unassembled scaffolds. Applying specific criteria for peak calling (see Materials and Methods), six highly enriched genomic regions, spanning 0.5-6 kb, were identified (Table 5 in Materials and Methods).

In these regions, we identified tandem repeats of an about 80 bp unit which comprises a mutated CENP-B box in which G16 is replaced by A (5' tTTCGttgtAtcCGGAc 3'). Surprisingly, no canonical box (5' nTTCGnnnnAnnCGGGn 3') was detected in these regions. The consensus sequence of the 80 bp repeated unit is shown in figure 3 as a logo. The identified CENP-B bound satellite repeat shares 94% identity with the previously described "9TK clone Sau1a" satellite (Accession number: AH005391.3 in NCBI database). This satellite family, isolated from chromosome 5 (Shampay et al. 1995), is organized head-to-tail and contains a highly conserved portion of 35 bp (Shampay et al. 1995). We observed that this is the portion containing the mutated CENP-B box that was not previously identified. From now, the consensus sequence of the 9TK clone Sau1a satellite we derived from the analyzed enriched genomic regions will be termed CENPB-Sau1a satellite. It is worth noticing that some genomic regions enriched in CENP-B bound chromatin contain also telomeric-like repeats indicating that the CENPB-Sau1satellite is intermingled with telomeric-like repeats, typically localized at Chinese hamster primary constrictions (Table 5 in Materials and Methods).

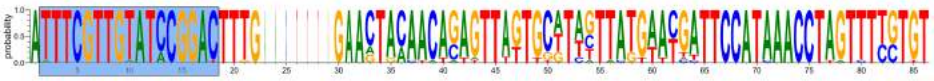


Figure 3. Consensus sequence of the CENP-B bound satellite of CHO-PV. The G16>A CENP-B box is highlighted with a cyan background. The width of each stack is proportional to its representation among the sequences used for generating the consensus sequence.

To confirm that CENP-B binds the CENPB-Sau1a satellite, we aligned the reads from input and immunoprecipitated DNA to the CENPB-Sau1a sequence. Reads were also aligned to the TTAGGG repeat and SatCH5

satellite (Faravelli et al. 1998) sequences, which lack the CENP-B box and therefore are not expected to be enriched in the immunoprecipitated sample. As reported in Table 1, CENPB-Sau1a is the only sequence enriched in the immunoprecipitated chromatin.

	CENP-B bound chromatin
CENPB-Sau1a	4.2
SatCH5	1.0
Telomeric-like repeat	1.0

Table 1. Fold enrichment of different satellite DNA families in the CENP-B bound chromatin of CHO-PV cells. Enrichment values were measured as the ratio between normalized read counts (RPKM) in immunoprecipitated and input DNA.

In order to definitely prove that the Chinese hamster CENP-B protein binds a mutated CENP-B box, we deduced a consensus of the CENP-B box sequence starting from the ChIP and Input reads mapping on the CENP-B Sau1a. As shown in Figure 4, the G16>A substitution of the box is totally conserved in both immunoprecipitated and input DNA. Moreover, other mutated boxes were detected in both the immunoprecipitated and input DNA but they were less represented than the G16>A variant (Figure 4). On the contrary, no canonical boxes were detected in the raw reads.

Taken together, these results prove that this CENP-B box variant is the functional CENP-B box in Chinese hamster. To our knowledge, this newly described CENP-B box was not described before. In addition, telomeric-like sequences are not bound by CENP-B despite their well described localization at the majority of Chinese hamster primary constrictions (Bertoni et al. 1996).

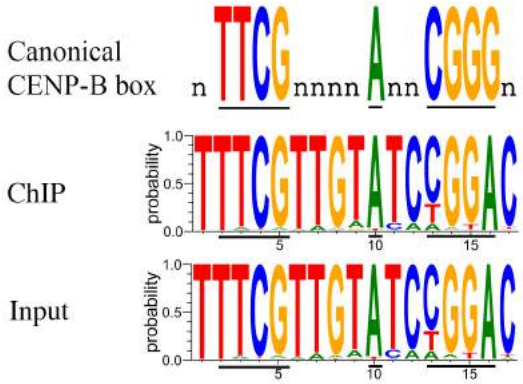


Figure 4. Chinese hamster CENP-B box. Logos of the consensus sequence of the CENP-B box in immunoprecipitated (ChIP) and input DNA of CHO-PV cells. In the upper part of the panel, a logo showing the nine canonical nucleotides for CENP-B binding is shown.

Thanks to Input data, we could estimate the genomic abundance of the different satellite families in the Chinese hamster genome. In accordance with previous work (Bertoni et al. 1996, Shampay et al. 1995, Faravelli et al. 1998), the telomeric-like arrays are extremely abundant in the Chinese hamster genome, while the other families, including the CENPB-Sau1a are less represented (Table 2).

	Genomic abundance
CENPB-Sau1a	5027.5
SatCH5	43750
Telomeric-like repeat	394004.8

Table 2. Genomic abundance of different satellite DNA families in CHO-PV cells. The value of the genomic abundance was measured as the normalized read count (RPKM) of input DNA.

4. TELOMERIC-LIKE REPEATS AT CHO CENTROMERES

To confirm the coupling between telomeric repeats and centromeric function, we performed a ChIP-seq experiment in the CHO-PV cell line with the same CREST serum used in immunofluorescence experiments (Figure 2A). ChIP and Input reads were mapped on the custom reference genome made by the sequences of the different satellite DNA families described before. As reported in Table 3, both the telomeric-like repeats and CENPB-Sau1a are enriched in the immunoprecipitated sample. Thus, we

demonstrated, at the molecular level, that telomeric-like repeats are associated with centromeric function. Regarding the enrichment of the CENPB-Sau1a satellite, it must be underlined that the CREST serum is not specific against CENP-A and, in particular, the one we used is highly enriched in antibodies against CENP-B (data not shown). It is important to remind that the genomic amount of CENPB-Sau1a is very low compared to the one of the telomeric repeats. The high enrichment of CENPB-Sau1a means that almost all the few copies of CENPB-sat reside in immunoprecipitated chromatin. On the contrary, the enrichment of the the telomeric-like repetitions is lower because they are expected to be present in large amounts at pericentromeric, interstitial and telomeric positions (Bertoni et al. 1996).

	Fold enrichment in CEN chromatin
CENPB-Sau1a	11.7
SatCH5	1.1
Telomeric-like repeat	2.7

Table 3. Fold enrichment of different satellite DNA families in the centromeric chromatin (CREST) of CHO-PV cells. Enrichment values were measured as the ratio between normalized read counts (RPKM) in immunoprecipitated and input DNA.

Discussion

1. Telomeric-like repeats bear the centromeric function in Chinese hamster

The Chinese hamster, as with several other vertebrate species, is characterized by the presence of very extended blocks of telomeric-like TTAGGG repetitions at centromeric positions, as result of chromosomal fusion and fission events during karyotype evolution. In particular, previous work from our laboratory proved that large clusters of TTAGGG repeats can be detected by FISH at nearly all primary constrictions both in a normal diploid cell line and in a CHO-K1 derived cell line (CHO-PV), demonstrating that in this species telomeric-like repetitions are spatially related to centromeres (Bertoni et al. 1996). We proved, at the molecular level, that telomeric-like repeats actually bear the centromeric function, since they are enriched in the centromeric DNA immunoprecipitated with a CREST serum. Our results are in accordance with recent works assessing that intrachromosomal telomeric repeats could act as seeds for centromerization in fission yeast and *Drosophila* (Olszak et al. 2011, Castillo et al. 2013).

However, despite its functional association with the centromeric core, the TTAGGG repeat does not contain any CENP-B binding site, in a scenario similar to that of the genus *Equus*.

2. The peculiar binding pattern of Chinese hamster CENP-B protein

Sequence analysis of the coding sequence of the CENP-B gene confirmed the extreme conservation of the N-terminal DNA binding domain among mammals. On the other hand, we detected an arginine-to-tryptophan substitution in the Chinese hamster compared to the human protein. Although conformational studies have not been performed, the highly different chemical properties of these amino acids raise the question whether this mutation could be well tolerated or whether conformational changes may arise leading to impairment in the dimerization. However, to our knowledge, this is the first report of a mutation in the C-terminal dimerization domain among mammals. In addition, we can hypothesize that this mutation specifically arose in the Cricetidae lineage, since in the mouse this amino acid substitution is not present. As with the equid CENP-B proteins, the Chinese hamster protein is 7 amino acid longer than the human proteins and these

additional residues are glutamate or aspartate residues of the two acidic domains. As mentioned in Part 3, the first acidic domain is involved in the interaction with CENP-C. Structural studies will be required to test whether this variability affects protein functionality.

Despite the peculiarity of the Chinese hamster protein, CENP-B appears functional, being detectable by immunofluorescence in the CHO-PV cell line. However, in a scenario similar to that of the horse, CENP-B can be detected only at the primary constriction of 14 out of 19 chromosomes, being undetectable on the two chromosomes 1 and on chromosomes 2, Z2 and Z7, while all the centromeres were labeled by the CREST serum, as expected. Differently from the equid species, CENP-B signals are homogeneous on the different chromosomes and display the typical "specked pattern" of CENP-A, suggesting a positioning in the centromeric core. The absence of signals on a few chromosomes is likely the result of a paucity of binding sites, suggesting a heterogeneous distribution of the satellite bound by CENP-B among chromosomes.

The satellite bound by CENP-B was identified by ChIP-seq with an anti-CENP-B antibody and consists of an 80 bp monomer containing a non-canonical CENP-B box. Surprisingly, this motif contains only eight of the nine essential nucleotides (5' tTTCGttgtAtcCGGac 3'), being characterized by a G16>A transition. Despite the mutation of this essential nucleotide, this box is actually bound by CENP-B, being enriched in the immunoprecipitated DNA, while no canonical CENP-B box can be found in the CHO-PV genome. According to structural studies, N7 of G16 is recognized by the side chain NH_2 group of Arg125 through hydrogen bonding (Tanaka et al. 2001). Our results prove that a G16>A transition could be tolerated by the protein without abolishing binding.

Nucleotides substitutions in the essential nucleotides of the CENP-B box were reported for the North African rodent *Leminscomys barbarus* (Stitou et al. 1999) and the red-neck wallaby *Macropus rufogriseus* (Stitou et al. 1999, Bulazel et al. 2006). In the former case, the novel box is 19 bp long and conserves only five of the nine essential nucleotides. However, its functionality in recruiting CENP-B was never tested. On the other hand, in the marsupial two essential nucleotides of the CENP-B box are mutated, but the protein was demonstrated to bind this box anyway (Bulazel et a. 2006). Although data on the DNA binding domain of *Macropus rufogriseus* are not available, it has been reported that a few marsupial species carry amino acid substitutions in the CENP-B domain devoted to motif recognition (Master thesis by Demetrio Turati). Thus, we can hypothesize that this variation in the

DNA binding domain might have led to a change in binding specificity of the protein. The situation is different in the Chinese hamster, where the DNA binding domain of CENP-B is totally conserved with the human one. Thus, to our knowledge this is the first reported case of a functional CENP-B box with a mutation in one essential nucleotide, still recognized by the totally conserved N-terminal domain of CENP-B. Although we identified an amino acid change in the C-terminal dimerization domain, it is unlikely that this mutation could affect the binding specificity of the protein.

Interestingly, the novel defined consensus of the satellite bound by CENP-B corresponds to a previously discovered minisatellite, namely 9TK clone Sau1a satellite (Accession number: AH005391.3 in NCBI database), which was isolated from chromosome 5 and found to be poorly represented in the Chinese hamster genome and organized in a head-to-tail fashion (Shampay et al. 1995). At the time of isolation, Shampay and collaborators noted that this satellite contains a highly conserved 35 bp portion, suggesting that this region could be involved in protein recruitment. In fact, this part is the one endowed with the mutated CENP-B box. In addition, we confirmed the fact that this satellite represents only a minor fraction of the satellite DNA library of Chinese hamster, which is mainly made of telomeric-like repetitions.

Furthermore, we detected telomeric-like repeats, more or less degenerated, within arrays of the CENP-B bound Sau1a satellite, suggesting that this satellite family is intermingled with the highly abundant TTAGGG arrays.

Conclusions

The investigation of CENP-B and its binding sites in Chinese hamster cells revealed that also this species is also a valuable model for studying CENP-B function. As in the genus *Equus*, centromeres without detectable levels of CENP-B are present also in this species, demonstrating that the typical binding pattern of CENP-B – in which CENP-B localizes at all primary constrictions with the exception of the Y chromosome - is not universal across mammals.

We demonstrated that the satellite bound by CENP-B matches a previously identified satellite, organized in tandem repeats of 80 bp monomers and poorly represented in the Chinese hamster genome. Surprisingly, the CENP-B box contained in this satellite family carries a G>A transition in one of the eight essential nucleotides. Intriguingly, the DNA binding domain of the Chinese hamster CENP-B is totally conserved with the typical mammalian one. To our knowledge, this is the first report of a change of CENP-B binding specificity without mutations in the DNA binding domain of the protein, suggesting a sort of CENP-B tolerance to mutations in the binding motif which was excluded so far.

As for the genus *Equus*, the CENP-B binding motif is not contained in the major centromeric satellite, which is the TTAGGG repeat. Indeed, we proved at the molecular level the coupling between telomeric-like repeats and centromere function, previously observed at the cytogenetic level (Bertoni et al. 1996). The bridge between telomeres and centromeres is attributed to chromosomal fusions and fission events during karyotype evolution and TTAGGG repeats at centromeres are supposed to be derived from ancient telomeres. In addition, it is well known that Chinese hamster is characterized by chromosome-specific families of satellite DNA (Faravelli et al. 1998), revealing a varied landscape of satellite DNA families at these peculiar centromeres. Drawing a parallelism with the results obtained in the genus *Equus*, we might wonder whether the peculiar rearrangements of Chinese hamster karyotype that characterize our CHO-PV cell line could be facilitated by the uncoupling between CENP-B and the centromeric function. To this end, we plan to characterize the binding pattern of CENP-B in normal diploid Chinese hamster cells and to extend the analysis to other species of the family Cricetidae. The understanding of the chromosomal distribution of CENP-B binding sites could help us to shed light on the controversial role of CENP-B.

PART 5
TRIDIMENSIONAL NUCLEAR ORGANIZATION OF CENTROMERES AND SATELLITE DNA IN THE GENUS *Equus*

Results

In the tridimensional architecture of mammalian nuclei, centromeres cluster at the nuclear and nucleoli periphery. It is a matter of debate whether centromere clustering depends on the presence of satellite repeats or on the centromeric function. In the genus *Equus* several centromeres are satellite-free, whereas many satellite DNA loci are not centromeric. Thus, equids represent a unique model to test whether the basis of centromere clustering depends on the primary DNA sequence (satellite DNA) or on the centromeric function. To answer this question, we analyzed centromere clustering in *E. caballus* and *E. asinus*.

1. NUCLEAR ORGANIZATION OF 37cen SATELLITE IN THE HORSE

In the horse (*2n*=64), the 37cen satellite is the major centromeric satellite and localizes, at the cytogenetic level, at all primary constrictions with the exception of the satellite-less centromere of ECA11 and the satellite-associated centromere of ECA2, where only 2PI (Piras et al. 2010, Cerutti et al. 2016) and CENPB-sat (Part 3 of this thesis) sequences are localized at the primary constriction. It is well known that pericentromeric and centromeric satellites form clusters in the 3D-organization of the nucleus (Jin et al. 2000, Weierich et al. 2003, Padeken et al. 2013, Burrack et al. 2016).

To visualize the spatial organization of the 37cen satellite in the horse we analyzed 35 nuclei by 3D FISH. The number of 37cen signals ranged from 15 to 28 with a mean of 21.1 ± 3.7 signals per cell. Since a total of 60 centromeres is labeled by 37cen in the horse (Piras et al. 2010), our results suggest a good degree of clustering for this satellite DNA sequence. Focusing on their intranuclear localization, 12.3 ± 2.9 signals were peripheral, 3.3 ± 1.5 localized around nucleoli and the remaining 5.5 ± 2.8 were internal (Figure 1). It must be underlined that all the nucleoli were always marked by the presence of 37cen clusters with a peculiar horseshoe shape (Figure 2). These

structures were visible also in cell types other than fibroblasts, such as the cell types found in retina sections (Figure 3). It should be noted, however, that this classification of 37cen signals has the limitation of depending on hybridization efficiency, nucleus size and relative position of satellite DNA signals. Actually, two separate but close signals might appear as single signals when the nucleus is highly compact or their intensity is high.

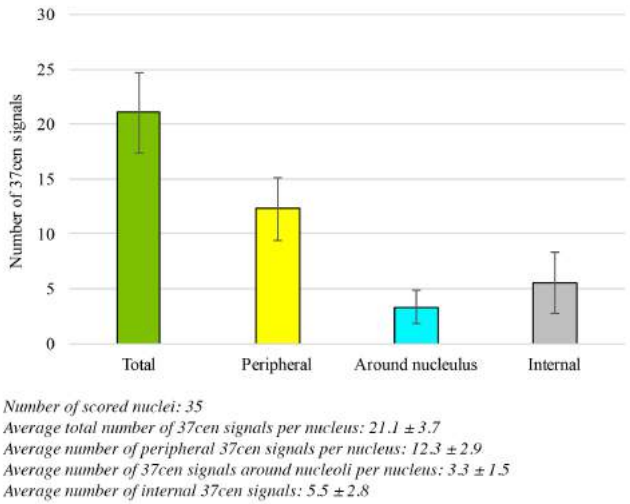


Figure 1. Clustering and intranuclear position of the 37cen satellite in the horse. A) On the left, average number of 37cen signals per nucleus classified according to intranuclear position. Error bars represent standard deviations. Number of nuclei = 35.

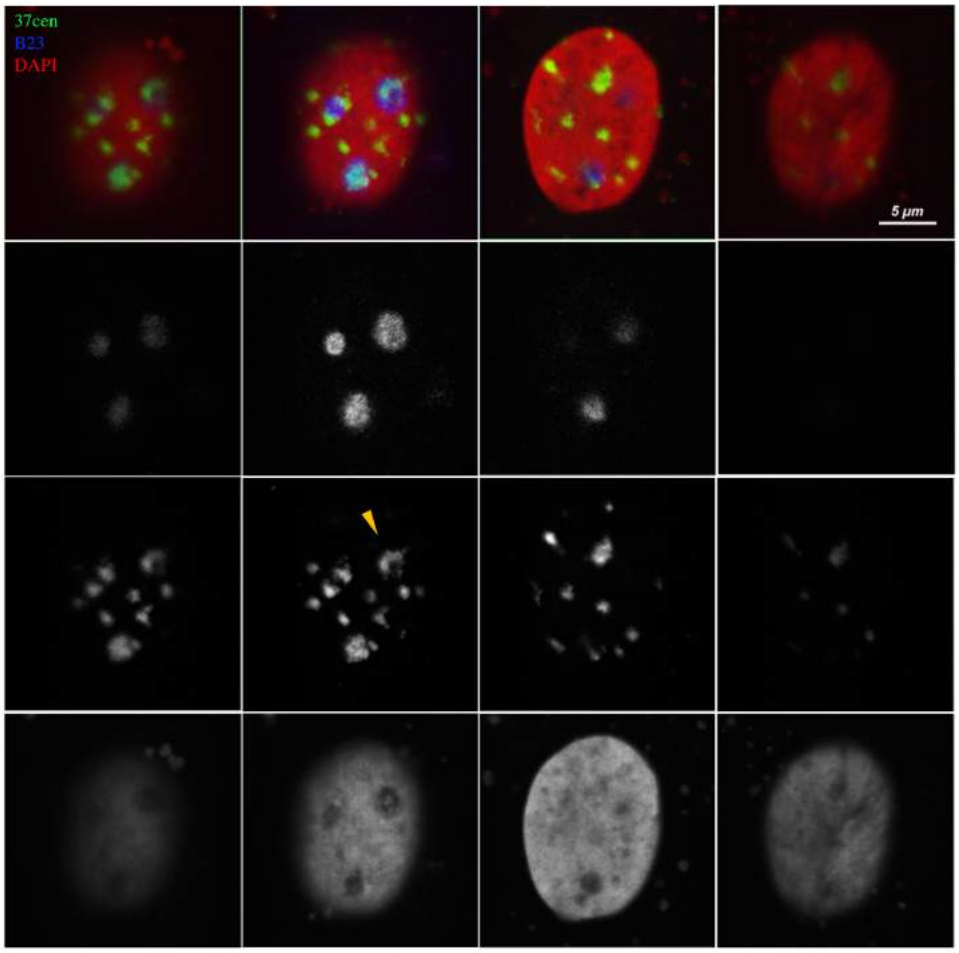


Figure 2. Organization of the 37cen satellite in horse primary fibroblasts. Partial series of optical sections of a horse fibroblast nucleus (red) hybridized with 37cen (green). Nucleoli (blue) were immuno-stained with an anti-B23 antibody. Yellow arrows point to examples of "horseshoe" clusters of 37cen.

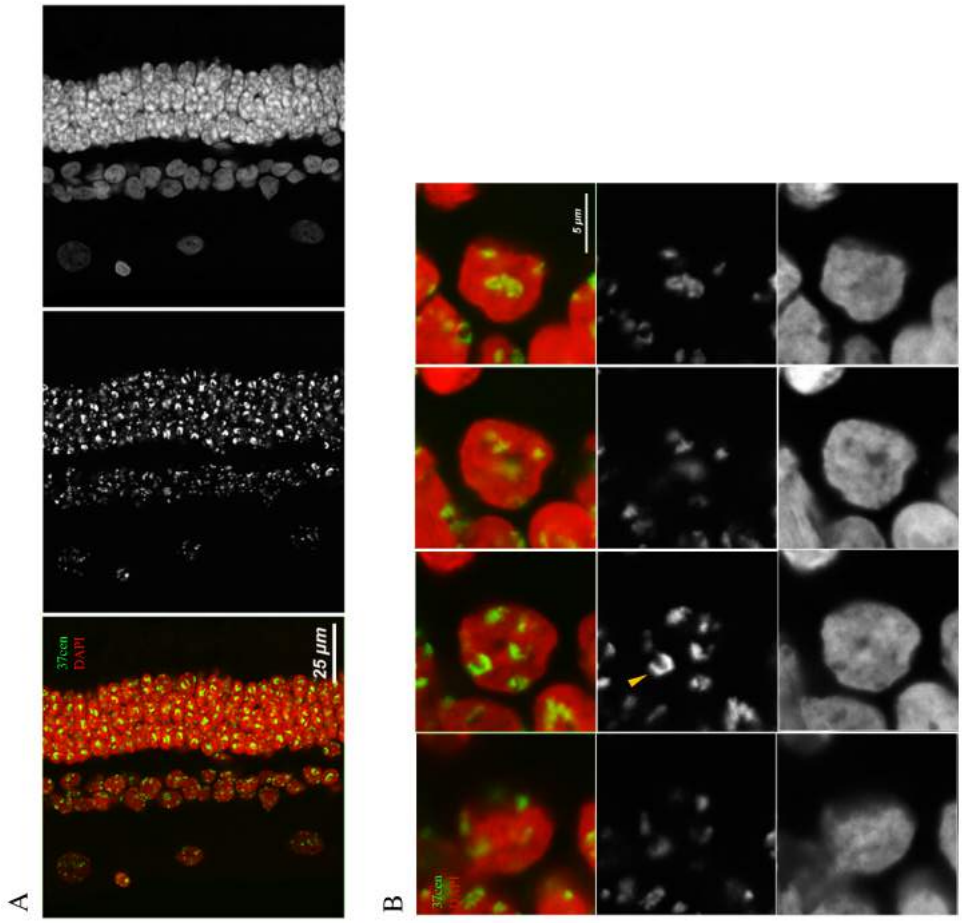


Figure 3. Organization of the 37cen satellite in the horse retina. A) Overview of a horse retina section hybridized with 37cen (green). B) Partial series of optical sections of a cell from the inner nuclear layer from horse retina hybridized with 37cen (green). Yellow arrows point to examples of "horseshoe" clusters of 37cen.

2. NUCLEAR ORGANIZATION OF SATELLITE DNA IN THE DONKEY

In the donkey (*2n*=62), satellite DNA is uncoupled from centromere function. Indeed, 16 centromeres out of 31 are satellite-less and satellite DNA frequently resides at non-centromeric termini as the remnant of an inactivated ancestral centromere (Piras et al. 2010, Nergadze et al. 2018). In particular, at cytogenetic level, satellite DNA localizes at one non centromeric terminus

of 13 meta- or submeta-centric donkey chromosomes (1p, 4p, 6p, 7p, 8p, 9p, 11p, 12p, 13p, 17p, 14q, 15q and 30q) and on the centromeric region of 13 chromosomes (1, 2, 3, 20, 21, 23, 24, 25, 28, 29, 30, X and Y). In addition, satellite DNA was detected at one interstitial site of X chromosome (Piras et al. 2010).

To investigate the clustering of satellite DNA in the donkey, a 3D FISH experiment was performed on donkey nuclei using donkey genomic DNA as probe for total satellite DNA (Piras et al. 2010). Given the fact that the primary fibroblasts were derived from a male donkey, in the hypothesis that there was no clustering of satellite DNA loci, we expected a total of 53 signals detectable by FISH. Scoring 27 nuclei, the number of satellite DNA signals ranged from 13 to 23, with a mean of 18.2 ± 2.7 signals per nucleus (Figure 4A). Thus, our results clearly demonstrate that satellite DNA sequences tend to coalesce irrespectively of the centromeric function.

Focusing on the intranuclear localization of satellite DNA signals per nucleus, 10.6 ± 2.5 were peripheral, 4.6 ± 1.5 were around nucleoli and 3.0 ± 1.6 were found at internal positions (Figure 4A). In addition, almost the totality of nucleoli was labeled by satellite DNA clusters with a peculiar horseshoe shape, as previously described for the horse (Figure 4B and Figure 5).

The same analysis was performed with the 37cen probe. Unlike the horse, in which 37cen is the CENP-A bound satellite (Piras et al. 2010, Cerutti et al. 2016), 37cen is not strictly centromeric in the donkey. Actually, the 37cen sequence is localized on one telomeric end of seven meta- or submeta-centric chromosome pairs (1p, 7p, 9p, 12p, 13p, 14q and 30p) and in the centromeric region of only two chromosomes (1 and 2), chromosome 1 showing a very large subcentromeric signal; thus, in chromosome 1, this probe recognized both the p arm terminus and the extended subcentromeric heterochromatic region (Piras et al. 2010). Thus, we expected a total of 18 loci detectable by FISH. Scoring 33 nuclei, the number of 37cen ranged from 9 to 15, with a mean of 11.6 ± 1.8 signals per cell, suggesting some degree of clustering in spite of the non centromeric nature of this satellite family. In the nucleus, 5.5 ± 1.6 37cen signals were peripheral, 2.9 ± 1.3 surrounded nucleoli and 3.2 ± 1.6 were internal (Figure 4C). The majority of nucleoli were again marked by the presence of 37cen clusters (Figure 4D and 6). These results suggest that, in the donkey, clustering is mainly due to the presence of satellite DNA rather than to the centromeric function.

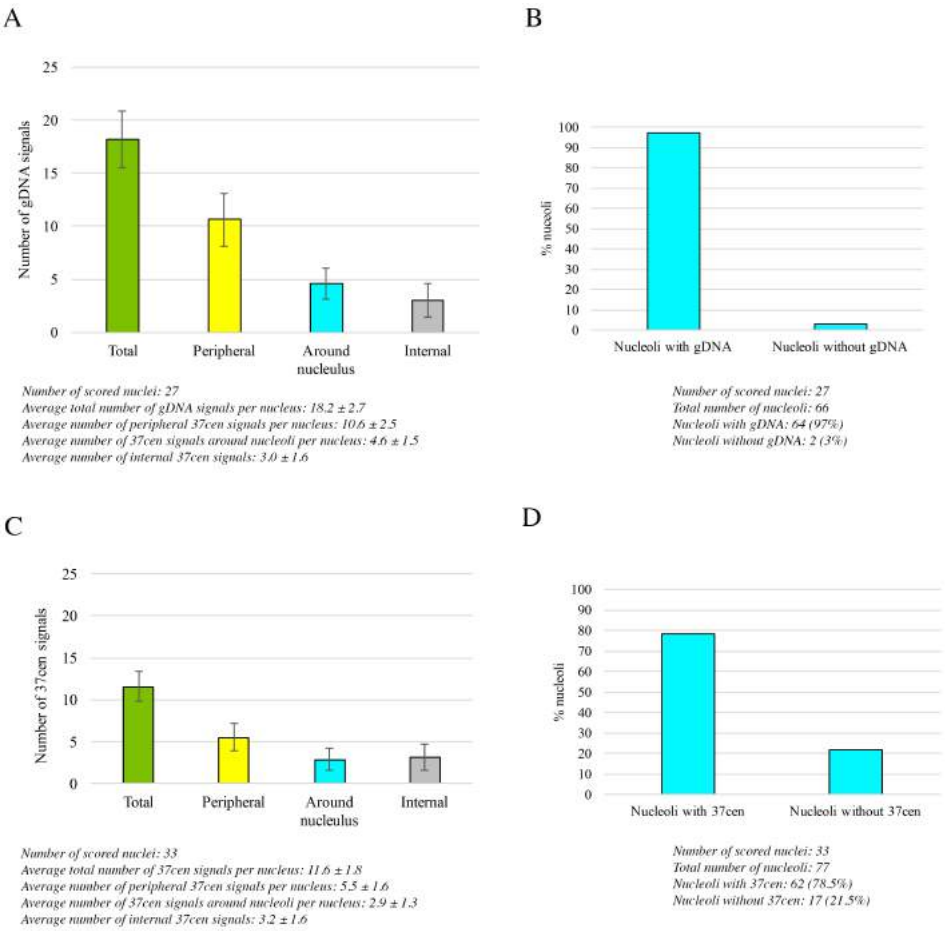


Figure 4. Clustering and intranuclear position of satellite DNA in the donkey. A) Average number of gDNA signals per nucleus classified according to intranuclear position. Error bars represent standard deviations. Number of nuclei = 27. B) Satellite DNA association to nucleoli. Number of nuclei = 27. Number of nucleoli: 66. C) Average number of 37cen signals per nucleus classified according to intranuclear position. Error bars represent standard deviations. Number of nuclei = 33. D) 37cen association to nucleoli. Number of nuclei = 33. Number of nucleoli: 77.

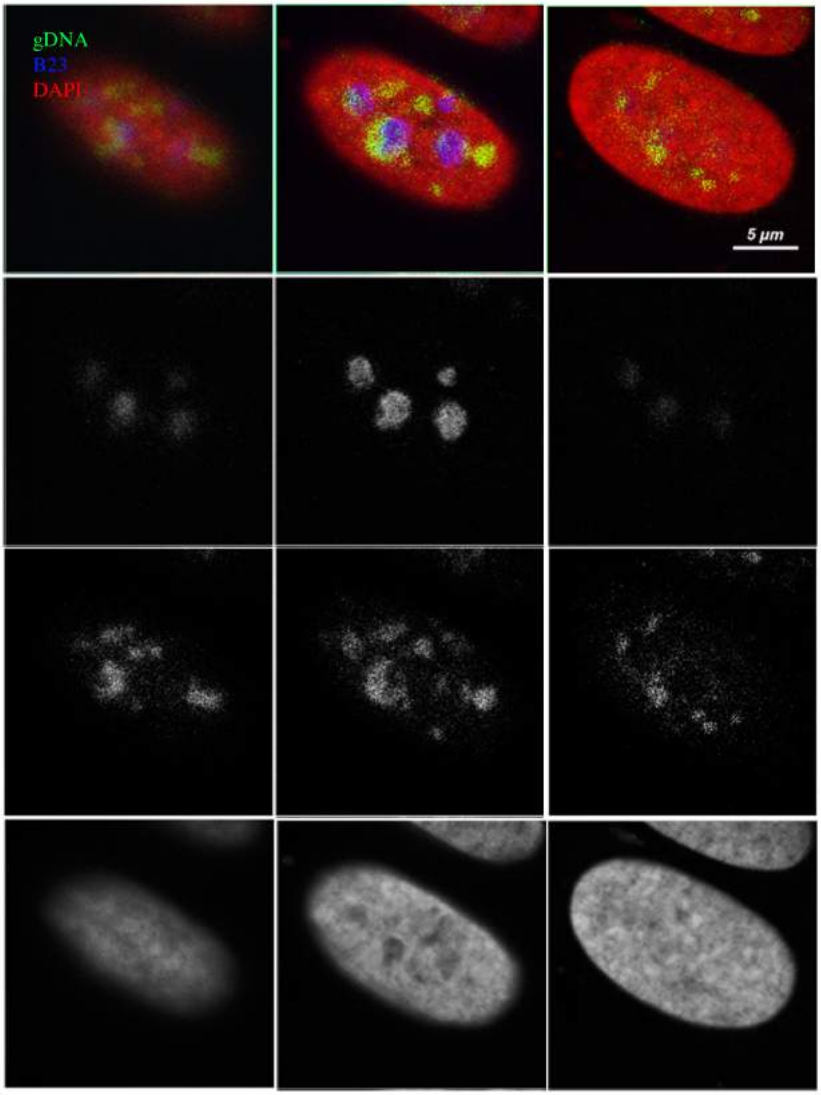


Figure 5. Organization of satellite DNA in donkey primary fibroblasts. Partial series of optical sections of a donkey fibroblast nucleus (red) hybridized with donkey genomic DNA (green). Nucleoli (blue) were immuno-stained with an anti-B23 antibody.

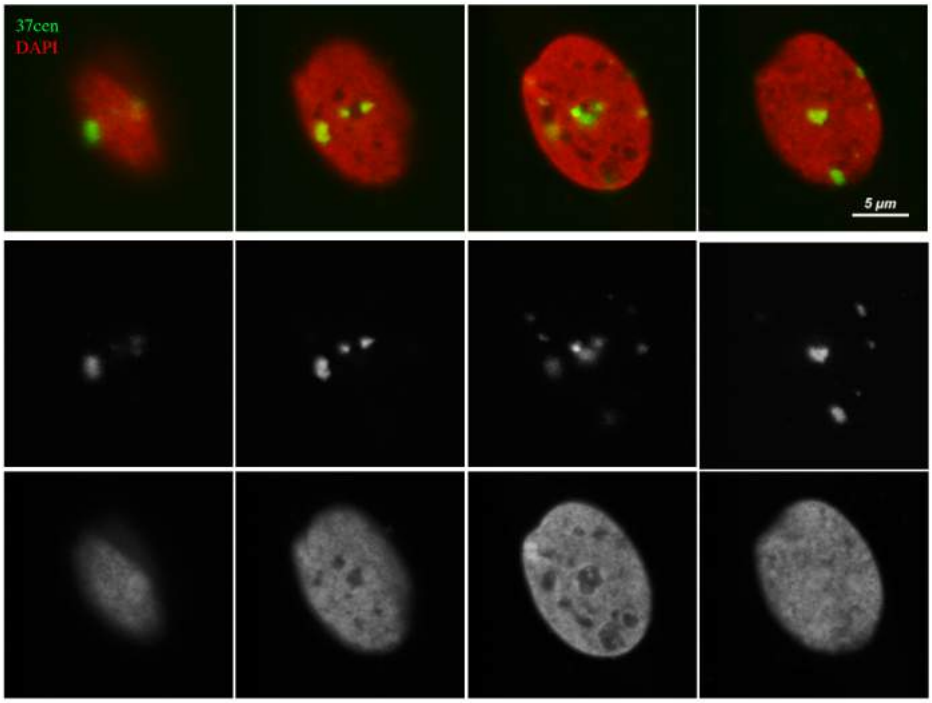


Figure 6. Organization of the 37cen satellite in the donkey. Partial series of optical sections of a donkey fibroblast nucleus (red) hybridized with 37cen (green).

To investigate the nuclear organization of satellite DNA with respect to centromeres, a 3D-immunoFISH was performed on donkey nuclei using a CREST serum and donkey genomic DNA as probe for total satellite DNA (Figure 7). Thirty nuclei were analyzed. Two different signal patterns were observed: centromeric clusters (satellite DNA signals including one or more CREST signals) and non-centromeric signals (satellite DNA signals without any detectable CREST signal). In all nuclei, the number of non-centromeric signals was always very low, ranging from 0 to 5, with 23% of nuclei showing no non-centromeric signals and a mean of 1.5 ± 1.2 non-centromeric signals per cell. In Figure 7 an example is reported. This finding indicates that the majority of non centromeric satellite loci cluster with the centromeric ones.

As shown in Figure 7, mitochondria are immunostained by the CREST serum. It is important to underline that this CREST serum contains also antibodies against mitochondrial proteins. However, we could distinguish the centromeric signals from the ones of mitochondria because of their different position and shape.

All the nuclei showed CREST signals which did not colocalize with satellite DNA. It is likely that these centromeric signals correspond to the donkey satellite-less centromeres, suggesting that clustering relies on the presence of satellite DNA rather than on the centromeric function. No clear association was observed among satellite-less centromeres, which appeared interspersed in the nucleus.

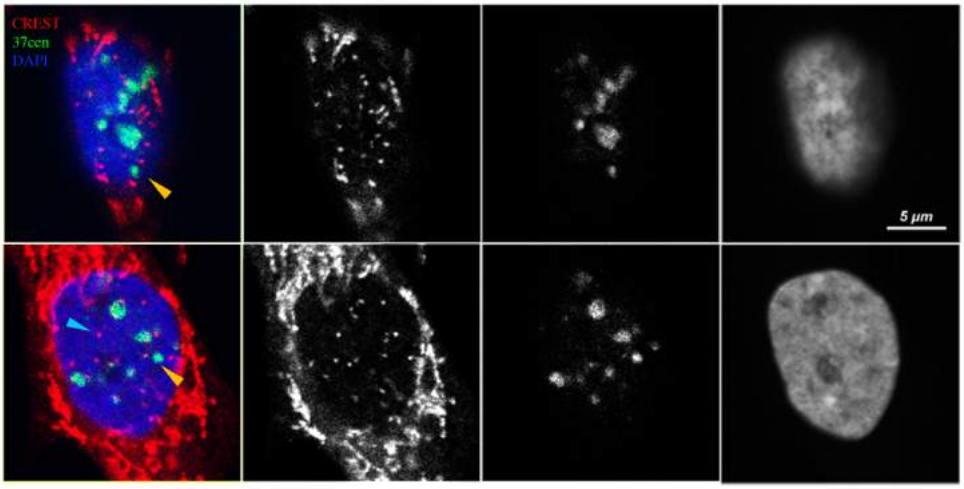


Figure 7. Clustering of satellite-associated centromeres and non centromeric satellite loci in the donkey. Partial series of optical sections of a donkey fibroblast nucleus with CREST-labeled centromeres (red) and total satellite DNA (green). Yellow arrows point two non centromeric satellite loci. The cyan arrow points to an example of satellite-less centromere.

3. NUCLEAR LOCALIZATION OF THE SATELLITE-LESS CENTROMERE OF HORSE CHROMOSOME 11

To assess whether centromere clustering is related to the centromeric function or to the satellite DNA sequence, we investigated the nuclear localization of the sequence of the horse satellite-less centromere of chromosome 11 (ECA11cen) with respect to satellite DNA.

We carried out 3D FISH on horse primary fibroblasts using a BAC probe from the satellite-less centromeric domain and the 37cen probe. We analyzed 54 nuclei and scored 108 ECA11cen signals clearly indicating that the centromeres of ECA11 do not cluster together. In addition, 67% of all ECA11 signals did not colocalize with 37cen (Figure 8 and 9). It is likely that the localization of some ECA11 signals with 37cen is mainly due to the large nuclear space occupied by 37cen clusters.

The lack of clustering between ECA11cen and the horse major satellite was observed also in other cell types, such as those from retina (Figure 10).

In the donkey, the region orthologous to the horse ECA11 centromere is not centromeric but resides on the q arm of donkey chromosome 13 (EAS13). We used the ECA11cen probe to test whether there is any association between this sequence, which is no more centromeric in this species, and 37cen satellite. As mentioned before, also 37cen is not centromeric in this species, being present at only two centromeres and at seven non centromeric chromosomal termini. Nonetheless, as previously described, in the donkey, satellite DNA sequences have the tendency to coalesce irrespectively of their function and, thus, both centromeric and non centromeric satellites predominantly cluster together. The analysis of 33 donkey nuclei revealed that, similarly to the horse, the majority (76%) of the ECA11cen signals did not colocalize with 37cen (Figure 8 and 11). We further evaluated the association between the ECA11cen sequence and total satellite DNA using as probe donkey genomic DNA (Piras et al. 2010): 50% of these signals colocalized with satellite DNA (Figure 8 and 12). These results demonstrate that the association between the ECA11cen sequence, which is not centromeric in the donkey, with satellite DNA is not related to its function but, as discussed below, it may be related to the status of its chromatin.

We then analyzed the intranuclear position of ECA11cen sequence. In the horse, the majority of ECA11 centromeres were localized at peripheral positions of the nucleus, although a substantial fraction was positioned either

around nucleoli or at internal positions. As shown in Figure 8, although we demonstrated the lack of association between this satellite-less centromere with satellite-associated centromeres, ECA11 centromeres share the same nuclear spaces of the satellite clusters. In addition, we observed that ECA11 centromeres which do not cluster with 37cen are preferentially found at the nuclear periphery. On the other hand, 37cen-clustered ECA11 centromeres are mainly found at the nuclear periphery and at the periphery of nucleoli (Figure 8C and 13). It is important to underline that nucleoli are always surrounded by 37cen satellite sequences (Figure 2).

In the donkey, although the ECA11 sequence is not centromeric, its intranuclear localization is similar to the one of the horse (Figure 8). It is worth noting that the ECA11 centromere and its orthologous donkey sequence are embedded in a heterochromatin domain as satellite-based centromeres (Part 3 and PhD thesis by Riccardo Gamba). It is well known that, in the nucleus, heterochromatin is restricted to the periphery and nucleoli (Solovei et al. 2016). Thus, it is likely that the intranuclear position of ECA11cen sequence and satellite DNA is not related to the centromeric function but to their shared epigenetic signature.

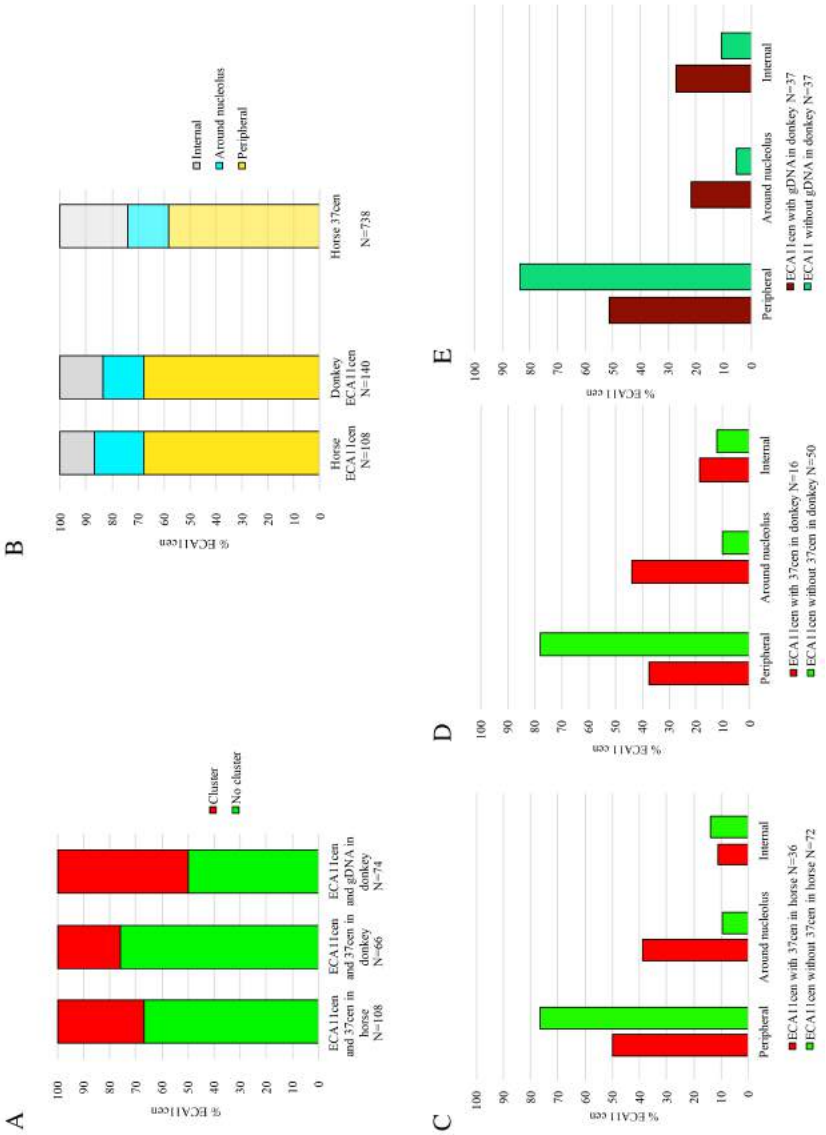


Figure 8. Satellite DNA association and intranuclear localization of ECA11cen sequence in horse and donkey. A) Percentages of absence of clustering (green) and clustering (red) between ECA11cen sequence and satellite DNA in horse and donkey primary fibroblasts. B) Intranuclear localization of ECA11cen in horse and donkey and 37cen-associated centromeres in the horse. C) Different intranuclear localization of ECA11cen according to 37cen association in the horse. D) Different intranuclear localization of ECA11cen according to 37cen association in the donkey. E) Different intranuclear localization of ECA11cen according to total satellite DNA association in the donkey.

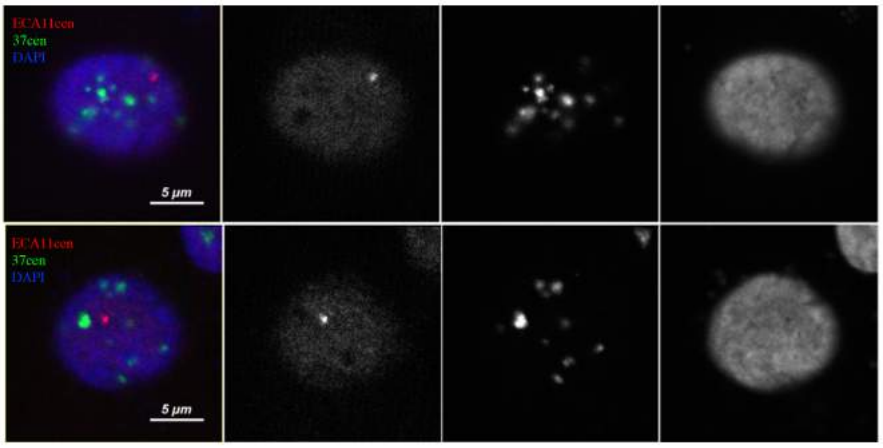
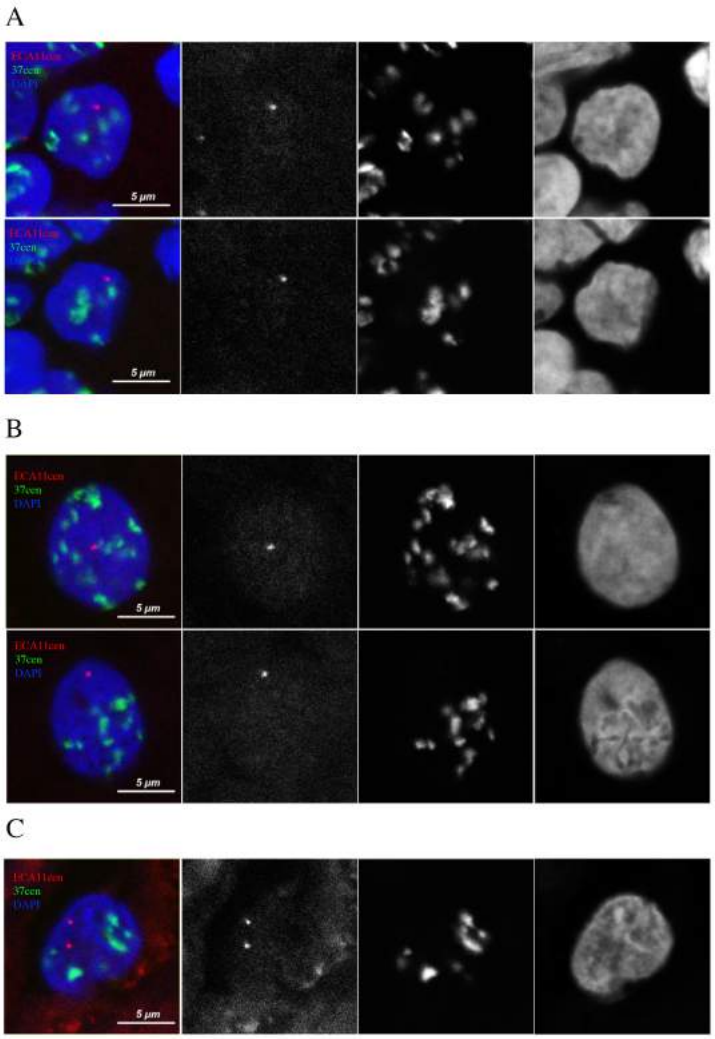


Figure 9. ECA11cen and 37cen in the horse. Horse fibroblast nucleus hybridized with ECA11cen (red) and 37cen (green). The two optical sections showing the focal planes of ECA11cen signals are reported. The two ECA11cen signals do not colocalize with 37cen signals.


A
B
C
5 µm

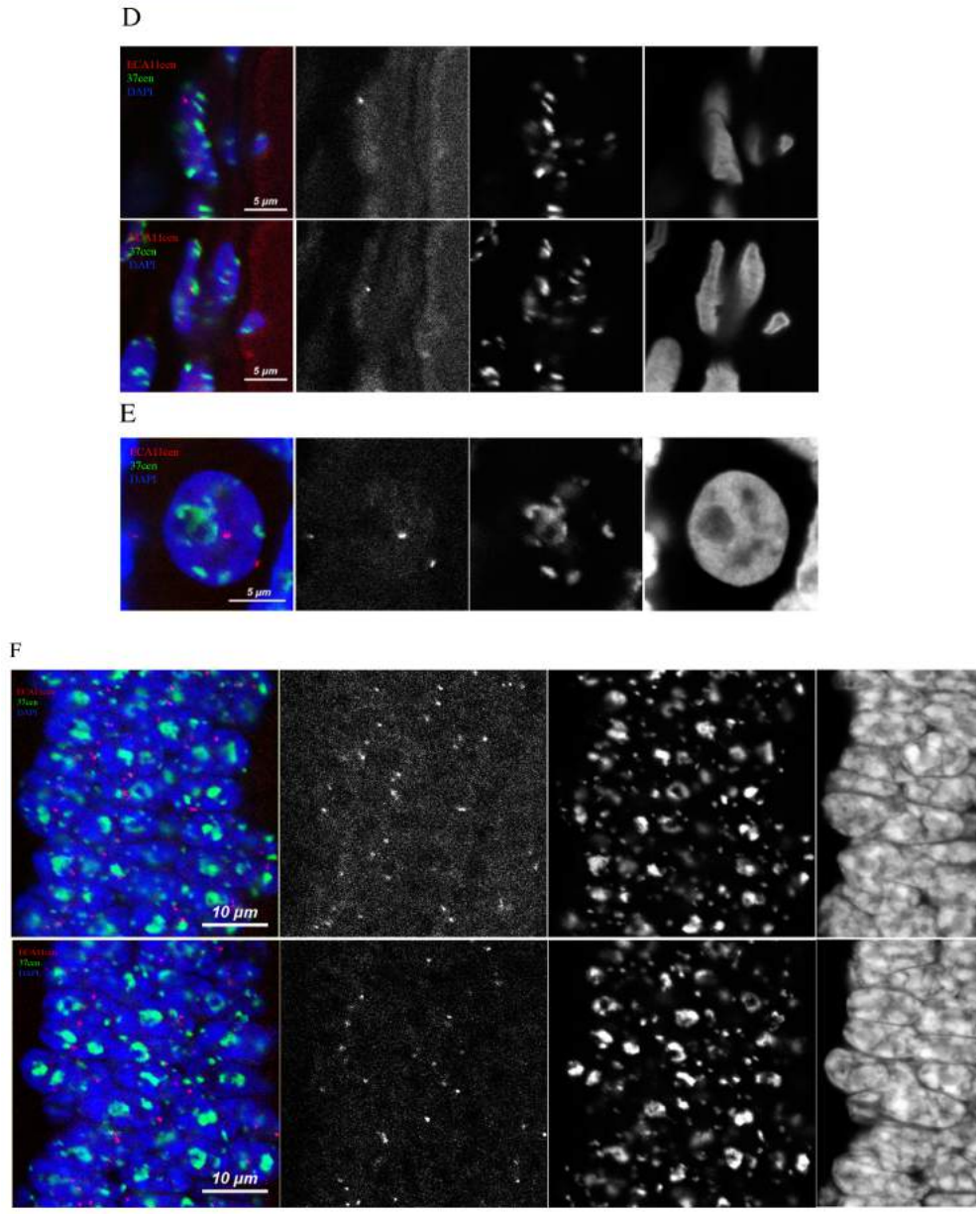


Figure 10. ECA11cen and 37cen in cell types from horse retina. Optical sections of different cell types from a horse retina section hybridized with ECA11cen (red) and 37cen (green). (A-E) For each image, the optical sections showing the focal planes of ECA11cen signals are reported. A) Inner nuclear layer cell. B) Ganglial cell. C) Pigment epithelial cell. D) Smooth muscle cell from retina vessels. E) Horizontal cell. F) Overview of the outer nuclear layer.

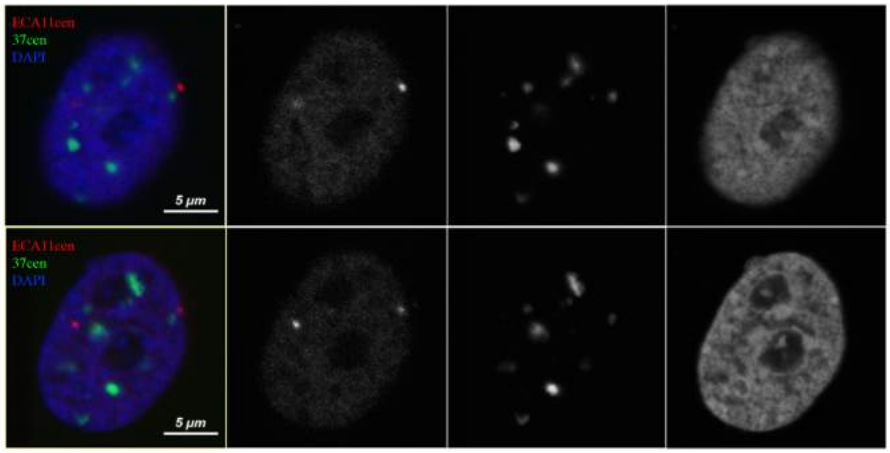


Figure 11. ECA11cen and 37cen in the donkey. Donkey fibroblast nucleus hybridized with ECA11cen (red) and 37cen (green). The two optical sections showing the focal planes of ECA11cen signals are reported. The two ECA11cen signals do not colocalize with 37cen signals.

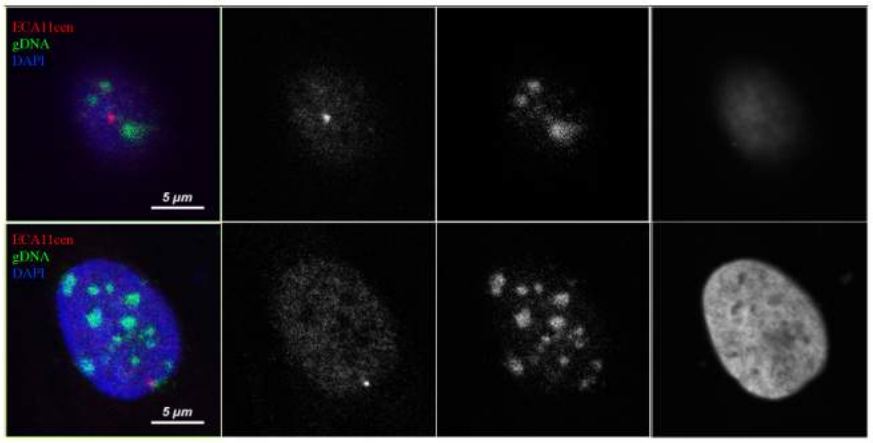


Figure 12. ECA11cen and genomic DNA in the donkey. Donkey fibroblast nucleus hybridized with ECA11cen (red) and donkey genomic DNA (gDNA) for detecting total satellite DNA (green). The two optical sections showing the focal planes of ECA11cen signals are reported. One ECA11cen signal colocalizes with donkey genomic DNA signals.

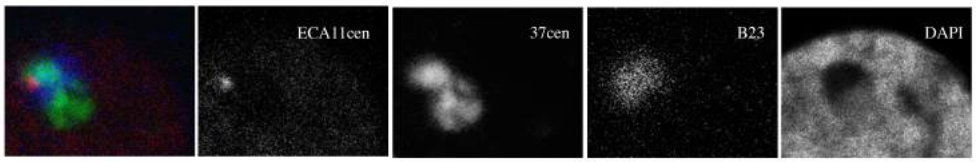


Figure 13. Colocalization between ECA11cen and 37cen at the periphery of nucleolus in the horse. Nucleulus (blue) of a horse fibroblast nucleus hybridized with ECA11cen (red) and 37cen (green). Nucleoli were immunostained with an anti-B23 antibody. On the right, DAPI staining.

4. NUCLEAR LOCALIZATION OF TWO SATELLITE-LESS CENTROMERES OF THE DONKEY

In the donkey, the presence of CREST signals without satellite DNA clearly demonstrated that satellite-less centromeres do not associate to highly repetitive DNA in the tridimensional nuclear architecture.

To confirm the lack of clustering between satellite-less and satellite-based centromeres, we investigated the nuclear localization of two donkey satellite-less centromeres with respect to satellite DNA in the donkey. These satellite-less centromeres were the one of donkey chromosome 5 (EAS5cen) and the one of donkey chromosome 13 (EAS13cen). 3D-FISH experiments were performed using two BAC probes for their centromeric domains on primary fibroblast nuclei. The two BAC probes were labeled with the same fluorochrome. Thus, in each nucleus we could score four BAC signals corresponding to the two pairs of satellite-less centromeres (EAScens). We evaluated the association between these satellite-less centromeres and total satellite DNA or 37cen satellite.

The analysis of 28 nuclei showed that the majority of EAScens signals (64%) did not colocalize with satellite DNA (Figure 14 and 15). The analysis carried out on 18 nuclei with 37cen probe revealed similar results: 75% of EAScens were not associated to this satellite sequence (Figure 14).

The intranuclear position of these satellite-less centromeres was similar to the one of ECA11 centromere. Indeed, these donkey satellite-less centromeres share the same nuclear compartments of satellite DNA clusters (Figure 14B). Moreover, satellite-less centromeres which do not cluster with satellite DNA are preferentially found at the nuclear periphery. Centromeres colocalizing with satellite DNA are mainly found at the nuclear periphery and at the periphery of nucleoli (Figure 14C and D). It is important to underline

that also these centromeres are embedded in a heterochromatic environment (PhD thesis by Riccardo Gamba). These finding confirm that the intranuclear position of satellite-less centromeres is not related to the centromeric function but to their epigenetic signature.

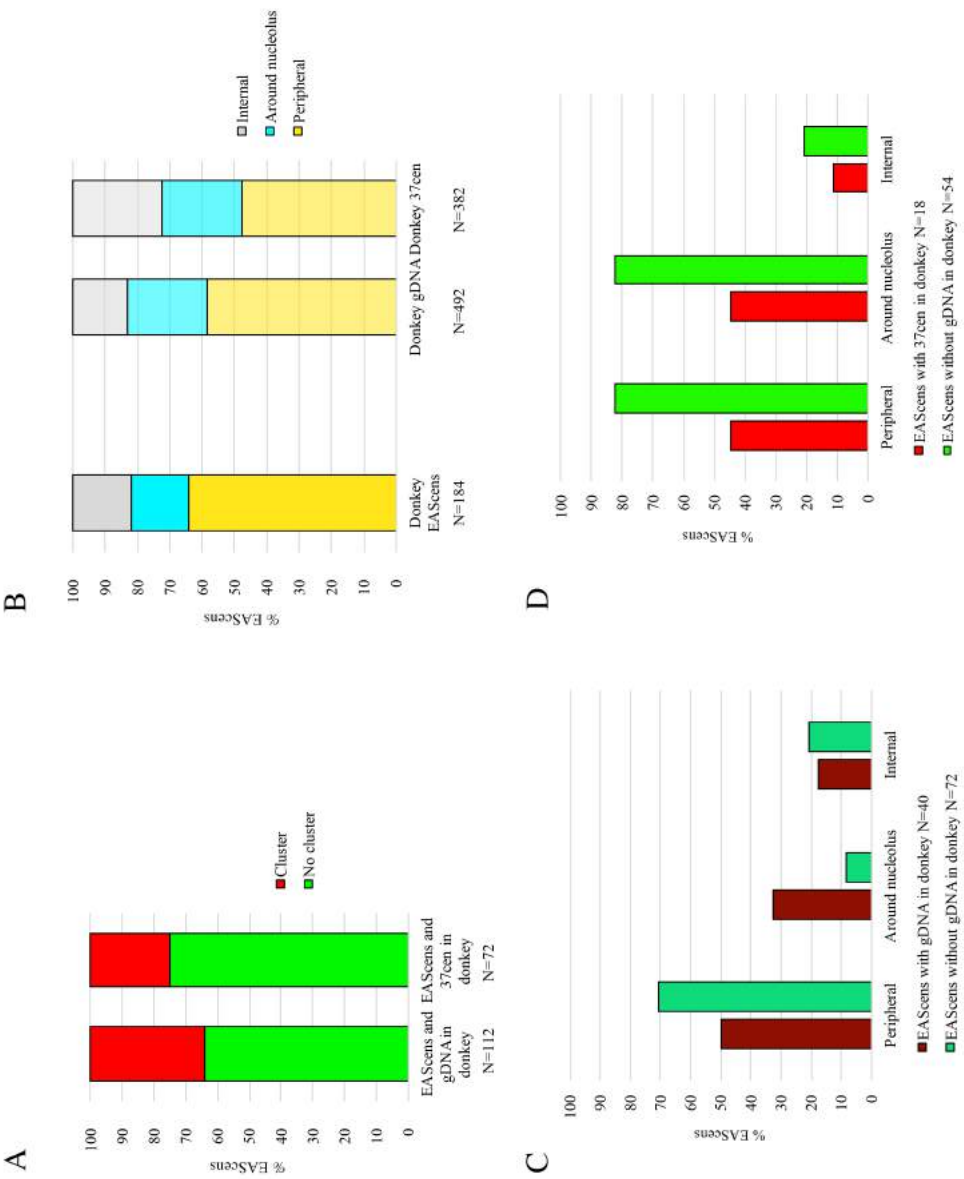


Figure 14. Satellite DNA association and intranuclear localization of two donkey satellite-less centromeres (EAScens) in donkey. A) Percentages of absence of clustering (green) and clustering (red) between ECA11cen sequence and satellite DNA in donkey primary fibroblasts. B) Intranuclear localization of EAScens, total satellite DNA loci and 37cen loci in the donkey. C) Different intranuclear localization of EAScens according to total satellite DNA association in the donkey. D) Different intranuclear localization of EAScens according to 37cen association in the donkey.

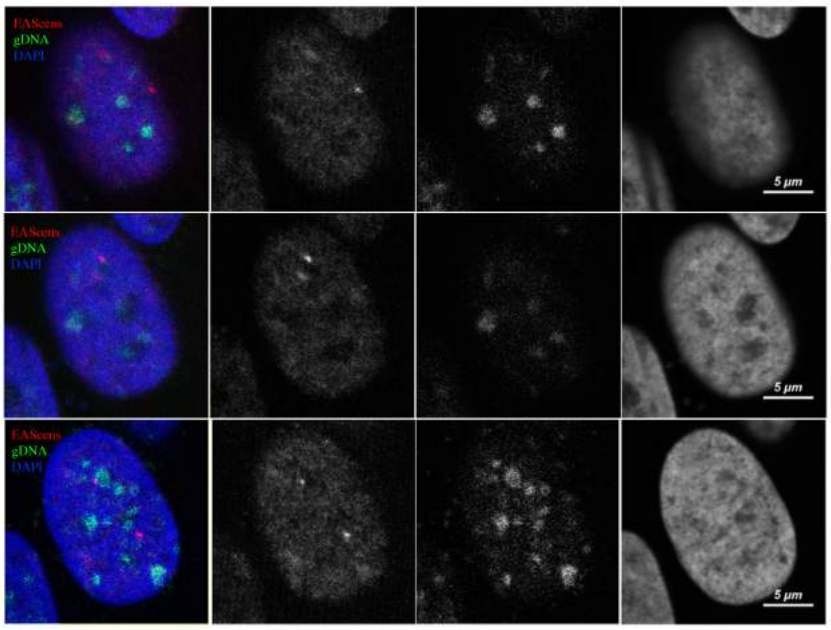


Figure 15. Two donkey satellite-less centromeres (EAScens) and donkey genomic DNA Donkey fibroblasts nucleus hybridized with EAScens (red) and donkey genomic DNA (green) The optical sections showing the focal planes of the four EAScens signals are reported. Only one EAScen signal colocalizes with gDNA.

Discussion

1. Satellite DNA clusters irrespectively of the centromeric function

One of the prominent features of mammalian nuclear architecture is clustering of centromeres in chromocenters. Despite the conservation of centromere clustering across eukaryotes, it was still a matter of debate whether in mammals this phenomenon depends on the presence of satellite DNA at these loci or on the centromeric function.

In the genus *Equus*, satellite DNA is uncoupled from centromeric function: many centromeres are satellite-less and many satellite DNA loci are not centromeric, representing relics of ancestral centromeres or traces of karyotype evolution occurred in this genus. Thus, the genus *Equus* provide us the opportunity to unravel the basis of centromere clustering.

Regarding the relationship between satellite DNA and centromeric function, *E. caballus* and *E. asinus* represent two different scenarios. In the horse, only one chromosome (ECA11) out of 32 is satellite-free, while all the other chromosomes carry satellite DNA at centromeric position, with 37cen satellite as the major centromeric satellite DNA family (Wade et al. 2009, Piras et al. 2010, Cerutti et al. 2016). In the donkey, the degree of uncoupling between satellite DNA and centromeric function is impressive: we recently demonstrated that 16 out of 31 centromeres lack highly repetitive DNA (Nergadze et al. 2018), while satellite DNA resides at the primary constriction of 13 chromosomes, at one non centromeric terminus of 13 chromosomes and at an interstitial position of one chromosome (EASX) (Piras et al. 2010). In addition, 37cen is not the major CENP-A bound satellite family in this species: 18 loci can be detected by FISH but only two reside at primary constrictions.

However, these two species show surprising similarities regarding the tridimensional organization of satellite DNA in the nuclear architecture. In horse nuclei, an average of 21 signals of 37cen satellite, corresponding to 60 centromeric loci, was detected, suggesting a high degree of clustering, in accordance with literature data on centromere clustering in mammals (Weierich et al. 2003, Solovei et al. 2004). In the donkey, hybridization with total genomic DNA for detection of satellite DNA loci revealed the presence of 18 signals, deriving from a total of 53 loci. Differently from the horse, only 26 of these loci reside at primary constrictions, demonstrating that satellite DNA loci tend to coalesce irrespectively of centromeric function. Further

evidence is given by the presence of clusters of 37cen satellite in the donkey in spite of their mainly non centromeric localization.

In addition, *E. caballus* and *E. asinus* share the same intranuclear distribution of satellite DNA, reflecting the typical pattern identified in mammals (Weierich et al. 2003, Solovei et al. 2004): the majority of clusters localize at nuclear periphery and almost all nucleoli were surrounded by clusters of satellite DNA with a peculiar horseshoe shape. A few internal signals, not associated to nucleoli, were detected as well. The same organization was not restricted to fibroblasts, but was found in different cell types, such as those present in the retina.

In the donkey, nearly all satellite DNA clusters contain at least one CREST signal, indicative of functional centromere. This observation suggests two possible scenarios. First, we can speculate that in an imaginary "ancestral nucleus" all the centromeres were satellite-based and clustered. During donkey karyotype evolution marked by exceptionally frequent centromere repositioning events, a number of satellite-based centromeres from each cluster were inactivated but these loci maintained their epigenetic signature and positioning in the tridimensional nuclear architecture, being still associated with the active ones. Alternatively, we can hypothesize that some centromeric protein, beyond its centromeric function, might act as chromocenter-bundling protein, driving satellite DNA association, recognizing also non centromeric and pericentromeric satellite DNA loci. This hypothesis fits the model of chromocenter formation by aggregation of both centromeric and pericentromeric highly repetitive DNA described by Jagannathan and Yamashita (2017).

In conclusion, our results clearly show that satellite DNA sequences coalesce independently from their coupling to centromeric function. Furthermore, these observations are in agreement with the previous notion that in the mammalian nucleus, chromosomal loci tend to associate according to their repeat enrichment as a result of mutual repeat recognition (Solovei et al 2016). Finally, it is well known that the maintenance of large blocks of satellite DNA poses a significant burden for the cell. However, it has been proposed that satellite DNA could have a structural role in shaping and anchoring the tridimensional nuclear architecture (Jagannathan and Yamashita 2017). In this view, we might hypothesize that the non centromeric satellite DNA loci of the donkey have still a role in the tridimensional nuclear organization.

2. Satellite-less centromeres do not cluster with satellite-based centromeres

Satellite-less centromeres represent the other side of the coin in evaluating the basis of centromere clustering. If the basis of clustering was the epigenetically-defined centromeric function, satellite-less centromeres would cluster with the satellite-based ones. On the contrary, our observations on the position of the centromere of ECA11, the only satellite-less centromere of the horse, confute this thesis. Indeed, the majority of ECA11 centromeres (67%) did not cluster with the major CENP-A bound 37cen satellite.

Our results also indicate that the association between a fraction of ECA11 centromeres (33%) and 37cen does not depend on the centromeric function because also the donkey non-centromeric orthologous sequences showed a similar behavior.

Despite the absence of association between satellite-less centromeres and satellite-based ones, the horse ECA11 centromere and its donkey orthologous locus surprisingly show the typical localization of satellite-based centromeres - at the nuclear periphery and adjacent to nucleoli. Since nucleoli are always surrounded by satellite clusters, loci positioned around nucleoli can sometime contact satellite DNA because of spatial limitations. On the other hand, random contacts between satellite-less loci and satellite DNA clusters are less frequent in the wide nuclear periphery.

The uncoupling between satellite-less centromeres and satellite DNA was observed also in the donkey.

Finally, in spite of the absence of centromere clustering, satellite-less centromeres tend to localize in the same nuclear compartments (nuclear periphery and around nucleoli). We know that satellite-less centromeres, as well as satellite-based centromeres, are embedded in heterochromatic domains in which transcription is repressed (Part 3, PhD thesis by Riccardo Gamba). Thus, we might hypothesize that satellite-less centromeres could have formed via epigenetic mechanisms in genome regions that had a nuclear distribution similar to conventional centromeres, having common epigenetic signature.

Conclusions

In conclusion, we demonstrated that, in mammals, satellite DNA clusters irrespectively of the centromeric function. Indeed, the same degree of cluster formation is observed in both the horse, where all satellite DNA loci are centromeric, and in the donkey, where more than half of satellite DNA loci are not centromeric, mainly corresponding to relics of ancient inactivated centromeres.

As for the mammalian species and cell types known so far (Weierich et al. 2003, Solovei et al. 2016, Jagannathan and Yamashita 2017), in both species the clusters localize mainly at the nuclear periphery and around nucleoli. Satellite-less centromeres share this positioning in the tridimensional nuclear architecture but do not cluster with the satellite-based ones, further proving that the phenomenon of centromere clustering relies on the presence of satellite DNA rather on the centromeric function.

PART 6
CENTROMERIC DOMAINS AND RECOMBINATION FOCI IN HORSE MEIOSIS

Results

1. DISTRIBUTION OF MLH1 FOCI ON ECA11

It is known that meiotic recombination is suppressed at centromeres and this phenomenon is called "centromere effect". As mentioned in the Introduction, the centromere of horse chromosome 11 (ECA11) is the only one devoid of satellite-DNA in this species. We wondered whether this centromere exerts the same inhibitory effect on meiotic recombination as a satellite-based centromere. The recombination rates and patterns in horse spermatocytes were already described by Al Jaru and collaborators (Al-Jaru et al. 2014), but still no work was focused on the relationship between meiotic recombination and satellite-less centromeres.

We performed immuno-FISH experiments on horse pachytene spreads to investigate the distribution of recombination foci with respect to centromeres. We used three antibodies: an anti-MLH1, which labels recombination sites, an anti-SYCP3, which labels the synaptonemal complex, and a CREST serum, to label all centromeres. We identified chromosome ECA11 as the only small one lacking 37cen signals (Figure 1).

It is well known that the number of MLH1 foci per chromosome is positively correlated with the synaptonemal complex (SC) length and this relationship was demonstrated also for the horse (Al Jaru et al. 2014). Therefore, to investigate the number and distribution of MLH1 foci on ECA11 using a small metacentric with 37cen signal as internal control. The length of SC complexes was measured using the ImageJ software and, in each spread, the metacentric chromosome with the closest length to ECA11 was chosen as control (Figure 1).

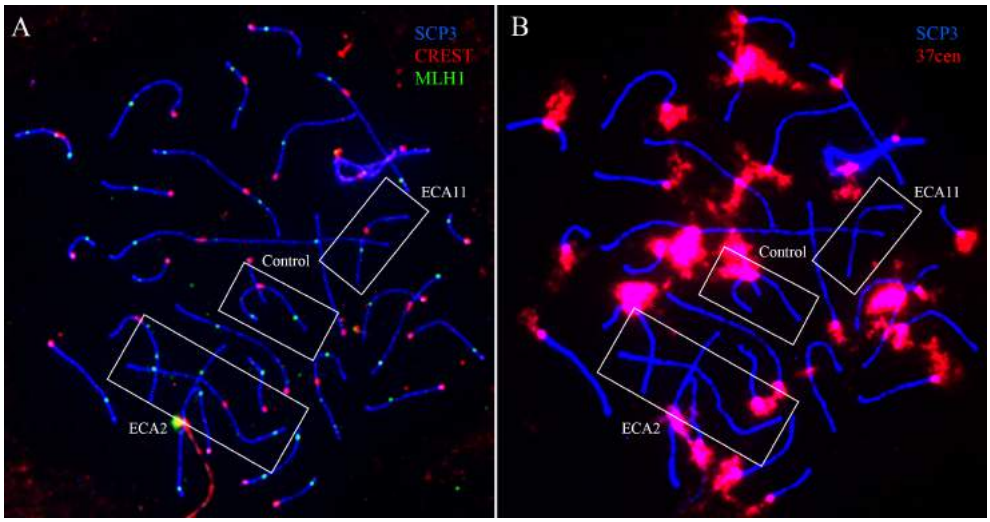


Figure 1. Localization of MLH1 foci and ECA11 identification. A) Triple immunofluorescence with the anti-SYCP3 antibody (blue), CREST serum (red) and an anti-MLH1 antibody (green) on a horse pachytene spread. B) "Reverse" FISH identification of ECA11 centromere using the 37cen satellite probe (red). ECA2 and ECA11 SCs are the only two chromosomes lacking 37cen signals. ECA11 can be recognized because of its shorter length compared to ECA2. The control SC was chosen as the one with the closest length to ECA11 SC.

We analyzed 25 cells in which we could recognize both ECA11 and the control: we detected a total of 36 recombination foci on ECA11 and a total of 34 recombination foci on the control SC, indicating that ECA11 does not differ from a chromosome with a satellite-based centromere. Furthermore, as shown in Figure 2, the number and the distribution between p and q arms of MLH1 foci on ECA11 were very similar to the ones of the control: in the majority of cases, ECA11 and the control displayed only a MLH1 focus on the q arm or two MLH1 foci, one on the q arm and the other on the p arm. This observation is in agreement with the notion that short chromosomes have a low number of recombination foci, but at least one cross-over event is required per bivalent.

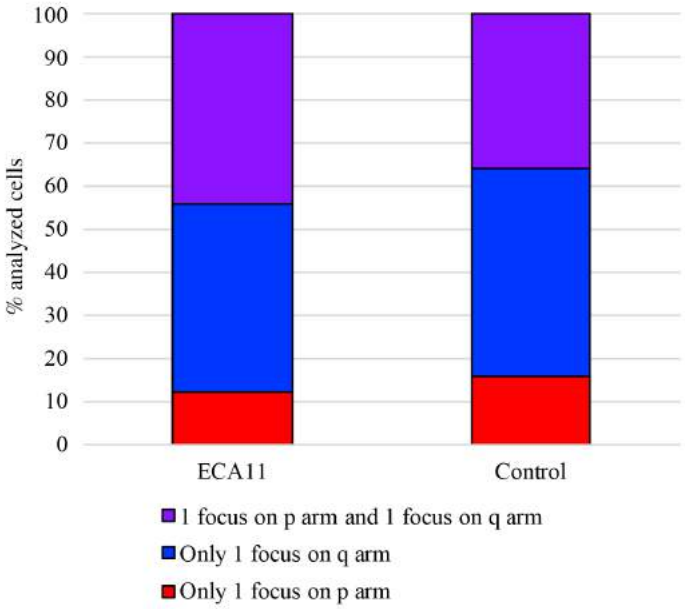


Figure 2. Arm distribution of MLH1 foci on ECA11 and control. Number of analyzed cells = 25.

To test whether the centromere of ECA11 inhibits meiotic recombination at the same level of a satellite-based centromere, we compared the distance between each MLH1 focus and the centromere on both ECA11 and the control. As shown in Figure 3, MLH1 foci are distributed in the same fashion in both ECA11 and control SC, with the majority of MLH1 foci residing in the distal region of the SC arm. Indeed, the differences between ECA11 and the control regarding the percentages of MLH1 foci found in different intervals of the SC arm are not statistically significant. However, since only 25 cells were analyzed, it will be important to increase the number of analyzed cells to avoid bias due to the low sample size. As shown in Figure 3B, in ECA11, no MLH1 foci were detected at a distance from the centromere lower than 19% and 26% of the length of q and p arm, respectively. Thus, these results suggest that ECA11 centromere exerts the same inhibitory effect on meiotic recombination as a classical satellite-based centromere. However, a complementary genetic approach will be required to overcome the resolution limits of this technique and to assess the precise boundaries of centromere inhibition.

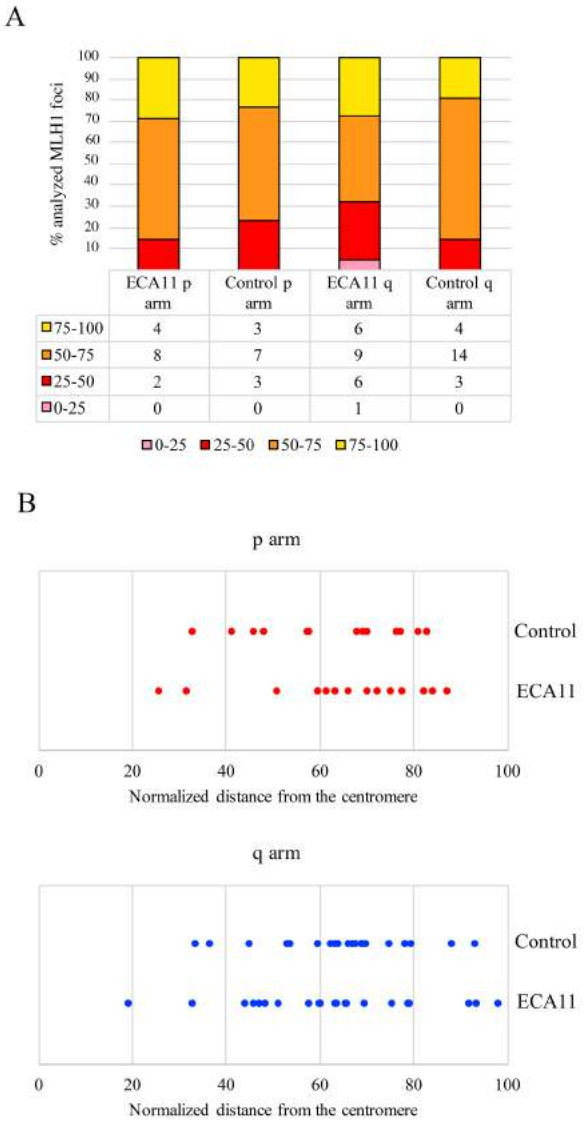


Figure 3. Distribution of MLH1 foci along the arms of ECA11 and control. A) Percentages of MLH1 foci found in four intervals of normalized distances from the centromere: 0-25% of the arm length, 25-50% of the arm length, 50-75% of the arm length and 75-100% of the arm length. In the table below, absolute numbers of foci of each class are reported. Differences are not statistically significant according to Fisher exact test. Number of analyzed cells = 25. B) Distribution of MLH1 foci on the relative length of p and q arms. On x axis, distance from the centromere reported as percentage of the length of SC arm.

2. IDENTIFICATION OF DOUBLE CENP-A SPOTS IN CHROMOSOME BIVALENTS

During preliminary work on meiotic recombination in horse meiosis, we observed by immunofluorescence experiments on horse spermatocytes the presence of double centromeric CREST signals in submeta- or metacentric chromosome bivalents at the pachytene phase (PhD thesis by Claudia Badiale). In particular, 61% of the analyzed cells showed the occurrence of at least one double-spotted centromere. The number of double-spotted bivalents per cell ranged from 0 to 5, although cells with more than two double signals were rarely identified (PhD thesis by Claudia Badiale).

To test whether double-centromeric spots are specifically due to CENP-A or to other centromeric proteins, recognized by the CREST serum, we performed immunofluorescence experiments on horse pachytene spreads from one individual with both a CREST serum and an anti-CENP-A antibody. Double-spotted bivalents were identified in both CREST (Figure 4) and CENP-A (Figure 5) experiments indicating that the double spots are not due to recognition of different proteins. Beyond the classical double signals made by two close dots, we could recognize extended signals longer than all the other single centromeric signals (Figures 4C and 5C). These “stretched” signals were also considered double signals since they likely derived from two spots too close to be resolved separately. However, because of resolution limits, a clear distinction between “two-dots” and “stretched” signals is not always clear. Double-signals were always found on meta- and sub-metacentric chromosomes, while they were never observed on acrocentric bivalents or on the XY body.

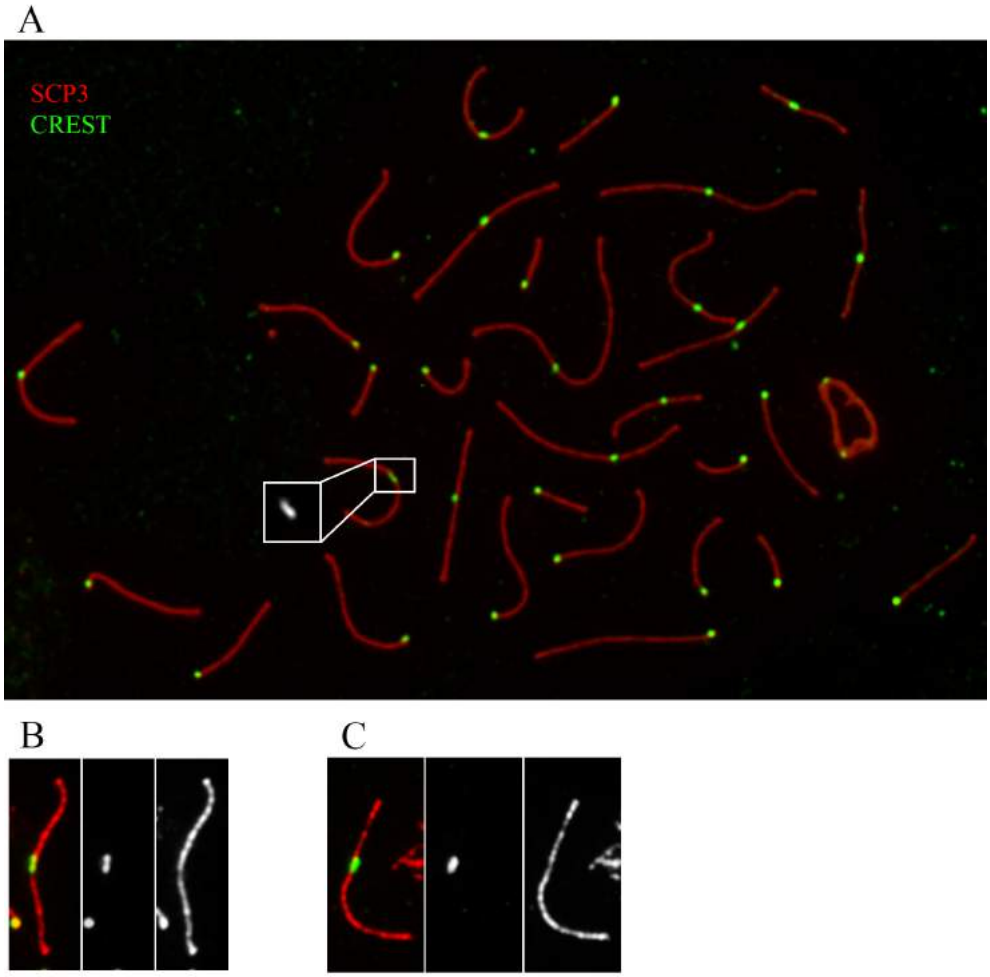


Figure 4. Double-spotted centromeres detected with CREST serum. The synaptonemal complexes (red) was immunostained with an anti-SYCP3 antibody. Centromeric signals (green) were detected with a CREST serum. A) Complete pachytene spread; a double signal is visible in the white rectangle. B) Example of classical "two dots" double signal. C) Example of "stretched" double signal.

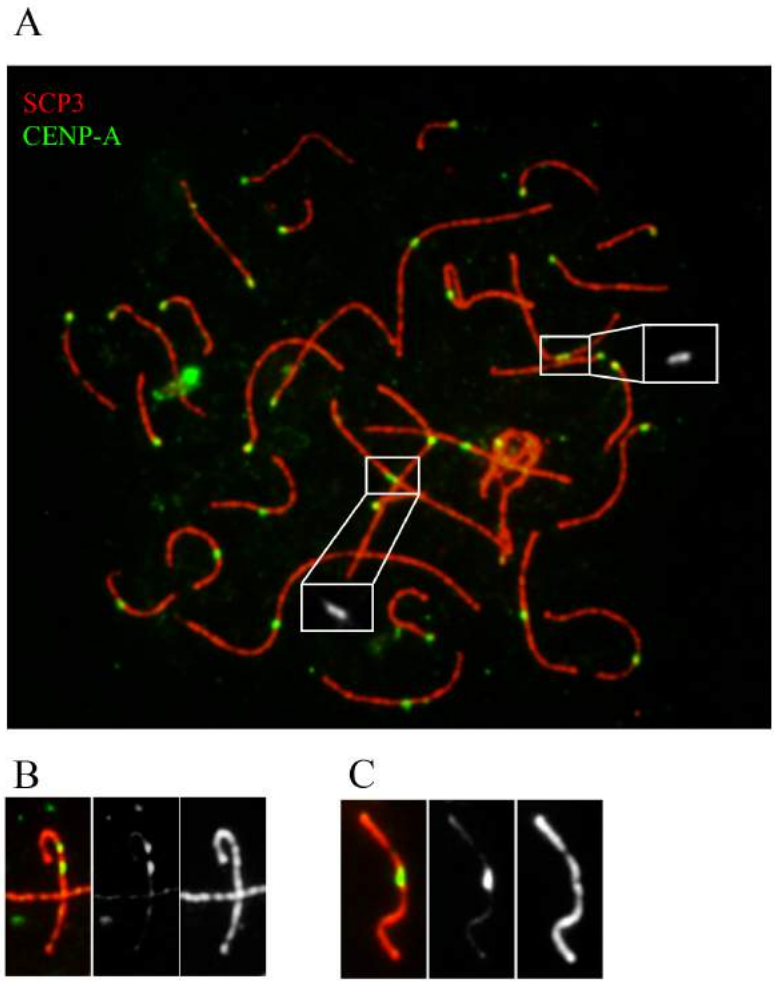


Figure 5. Double-spotted centromeres detected with the anti-CENP-A antibody. The synaptonemal complexes (red) was immunostained with an anti-SYCP3 antibody. Centromeric signals (green) were detected with an anti-CENP-A antibody. A) Complete pachytene spread; a double signal is visible in the white rectangle. B) Example of classical "two dots" double signal. C) Example of "stretched" double signal.

The frequency of double-spotted bivalents detected using CREST serum or the anti-CENP-A antibody was very similar (Figure 6A), demonstrating that the presence of double signals actually relies on CENP-A protein and is not due to other proteins immunostained by the CREST serum. Thus, our findings suggested that the double signals correspond to the two different CENP-A domains of the paired homologs.

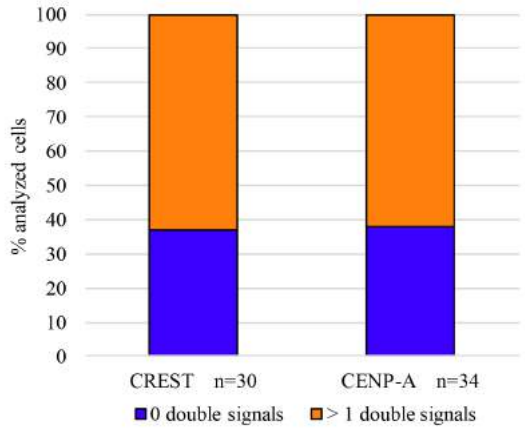


Figure 6. Frequency of double-spotted centromeres detected with CREST serum or anti-CENP-A antibody in a horse individual. Percentage of cells containing no double-spotted bivalents or at least one double-spotted bivalent using the CREST serum (left, number of analyzed pachytene spreads = 30) and the anti-CENP-A antibody (right, number of analyzed pachytene spreads = 33).

2.1. Absence of correlation between the frequency of double-spotted centromeres and the synaptonemal complex length

The variability in the number of centromeric double signals in different cells could be the result of different mechanical stretching of chromatin during the technical treatments for pachytene spread preparation (see Materials and Methods). To test this hypothesis, we evaluated whether the number of CREST or CENP-A double signals positively correlates with the total length of the synaptonemal complexes, used as a parameter to measure the extension of each pachytene cell. As shown in Figure 7, a correlation between the number of double signals and cell extension does not exist. Thus, we can conclude that the different number of double-spotted centromeres is not due to the mechanical stretching of SCs and that these peculiar centromeres are characterized by two separated or partially overlapping domains.

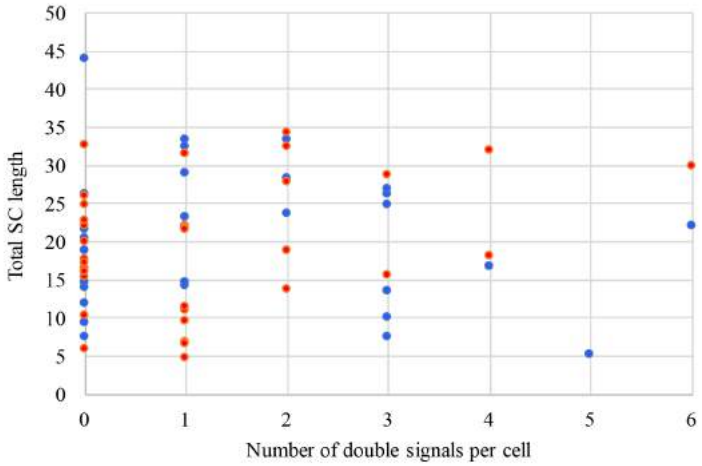


Figure 7. Absence of correlation between the number of double-spotted centromeres and total length of synaptonemal complexes. Total length of synaptonemal complexes per cell (y axis) was measured in inches using Image J software. On the x axis, number of double signals per cell. Blue dots represent cells measured from immunofluorescence experiments with the CREST serum (number of analyzed cells = 30). Red dots represent cells measured from immunofluorescence experiments with the anti-CENP-A antibody (number of analyzed cells = 34).

2.2. Intra- and inter-individual variability of double-spotted centromeres

As mentioned before, preliminary data from our laboratory revealed that the number of double-spotted bivalents per cell varied, although an evaluation of intra-individual and inter-individual variability was not performed. To this end, we performed the same immunofluorescence experiment using the anti-CENP-A antibody on two additional horses. For each horse, we scored the number of double-spotted centromeres per cell. As shown in Figure 8, intra-individual variability was high in all the three individuals, since the number of double-spotted centromere ranged from 0 to 7. In addition, cells with more than three double signals were under-represented or absent in all the horses. Moreover, inter-individual variability was observed regarding the presence or absence of at least one double signal: while in 38% of the cells from horse 1 no double signals could be detected, only 10% and 22% of cells of horse 2 and horse 3, respectively, did not show any double-signal. These differences are statistically significant with a p

value of 0.02 according to Chi-Square test. Therefore, horse 2 is the individual with the highest frequency of cells with double-spotted centromeres. The average number of double spotted-centromeres per cell was 1.2±1.4 in horse 1, 2.3±1.8 in horse 2 and 1.8±1.5 in horse 3. Thus, although the frequency and the distribution of cells with double-spotted centromeres were different in the three individuals, the difference in the average numbers of double-spotted centromeres is not statistically significant. The analysis of a larger number of cells will be required to test the significance of inter-individual variation.

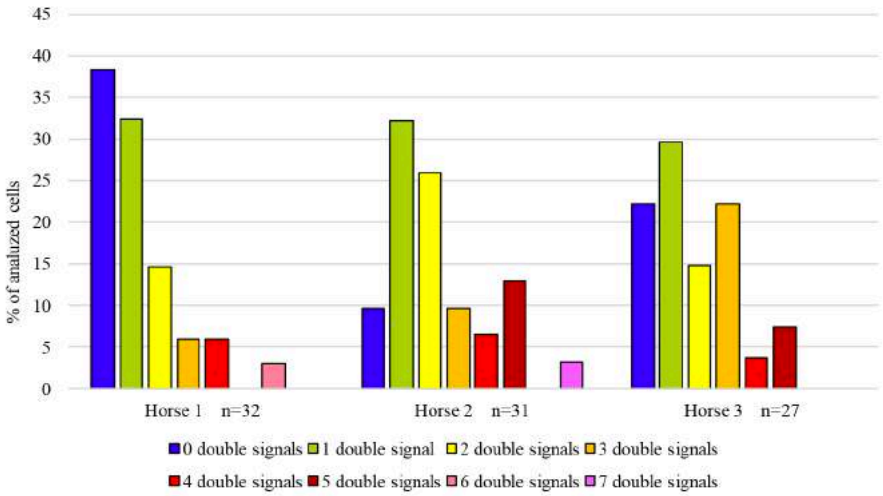


Figure 8. Intra- and inter-individual variability of double-spotted centromeres. For each individual, the percentages of analyzed cells with 0, 1, 2, 3, 4, 5, 6 and 7 double-spotted centromeres are reported. Horse1 is the horse described in the previous paragraphs.

Discussion

1. ECA11 satellite-less centromere and meiotic recombination

The pattern of meiotic recombination of each chromosome depends on chromosomal length, centromere and telomere effect, interference between crossing-over events and DNA sequence features. In particular, the centromere exerts a direct negative effect on meiotic recombination both within itself and proximal regions (Choo 1998).

We took advantage of the unique satellite-less centromere of horse chromosome 11 (ECA11) to test whether a centromere void of satellite DNA suppresses meiotic recombination at the same level of a satellite-based one.

We mapped the recombination foci through the cytogenetic localization of MLH1 protein, a mismatch repair protein of mature recombination sites, along the synaptonemal complexes (SCs). Since the number of recombination foci correlates with the SC length, we compared the distribution of MLH1 foci on ECA11 SC and on another chromosome with comparable meiotic length. The behavior of ECA11 centromere is similar to the one of the compared chromosome with a satellite-based centromere, demonstrating that the inhibitory effect depends on the centromeric function rather than the presence of satellite DNA. These results are in agreement with the common knowledge that crossing over events in the centromeric regions are negatively selected because of disruption of pericentric sister chromatid cohesion or chromosome breakage and loss (Talbert and Henikoff 2010).

2. Identification of double-spotted centromeres at bivalents of horse pachytene phase

One of the key steps of meiosis is the pairing of homologous chromosomes and their connection through the formation of the synaptonemal complex along their length. The formation of chromosome bivalents is accompanied by the pairing of centromeres, as reported for all the eukaryotic species studied so far (Kurdzo and Dawson 2015, Da Ines and White 2005). In the pachytene phase of horse meiosis we uncovered a peculiar phenomenon: centromeres with double CENP-A spots, distributed along the synaptonemal complex length, were frequently observed. Two main morphologies could be recognized: “two dots” signals, in which two distinct CENP-A spots could be resolved, and long “stretched” signals, which were

interpreted as double signals in which the two discrete signals were too close to be distinguished separately. These particular signals were identified on contiguous meta- and sub-metacentric chromosome bivalents. We excluded that the variability in the number of these double signals per cell could be simply the result of different degrees of mechanical stretching of the chromatin during the technical treatment for pachytene spread preparation. Indeed, the number of double-spotted centromeres does not correlate with the total length of the synaptonemal complexes, used as parameter to evaluate the extension of each pachytene cell.

Interestingly, the percentage of cells with at least one double centromeric signal displayed inter-individual variation. Moreover, the number of double-spotted centromeres per cell was highly variable also among cells of the same individual, although we revealed a general trend: the higher the number of double spotted centromeres, the lower their frequency. In addition, these signals were never observed on acrocentric bivalents and XY bodies.

The observation at the cytogenetic level of these peculiar centromeres raises the question on the biological meaning of these double CENP-A domains. We recently demonstrated, taking advantage of the satellite-less centromeres of the genus *Equus*, that the position of the centromere is not fixed but slides, giving rise to different positional alleles, defined "epialleles", which are inherited as Mendelian traits (Purgato et al. 2015, Nergadze et al. 2018, Part 2). Similar polymorphism regarding the position of the CENP-A binding domain was reported also in some human satellite-based centromere, such as the one of HSA17. Indeed, in HSA17, the centromere can assemble on different alpha satellite arrays and individuals with heterozygosity in the position of the centromere were reported (Maloney et al. 2012). Furthermore, satellite DNA arrays display high variation even between homologous chromosomes in single nucleotides polymorphisms within HORs, HOR size and total array size (Waye and Willard 1986, Warburton and Willard 1995, Sullivan et al. 2017). The presence of polymorphism among homologous chromosomes regarding the position of the centromere and the number of tandem repeats suggests a possible interpretation of the occurrence of double-spotted centromeres. As reported in the model presented in Figure 9, we can reasonably hypothesize that misalignment between homologous chromosome frequently occurs in the centromeric and pericentromeric regions during homolog pairing. This misalignment may increase the physical distance between two centromeric domains that are already in different positions on the satellite DNA array. In this view, our double-spotted centromeres would

become visible when the distance between the centromeric domains of the two homologous chromosomes is large enough to be resolved by our method (Figure 8B and 8C). The fact that only submeta- and meta-centric bivalents displayed this peculiar phenomenon could be due to the higher probability of centromere misalignment with respect to acrocentric ones. This is in agreement with the observation of "misaligned" centromeres on one metacentric chromosome bivalent in both the common shrew and the dwarf hamster (Borodin et al. 2008. Bikchurina et al. 2018). However, in our system we identified double-spotted centromeres at a surprising high frequency on different chromosome bivalents, showing a high intra- and inter-individual variability. These findings suggest that the combination of centromere sliding and misalignment of satellite arrays may occur at high frequency in the horse, in agreement with the exceptional centromere plasticity of the equid species.

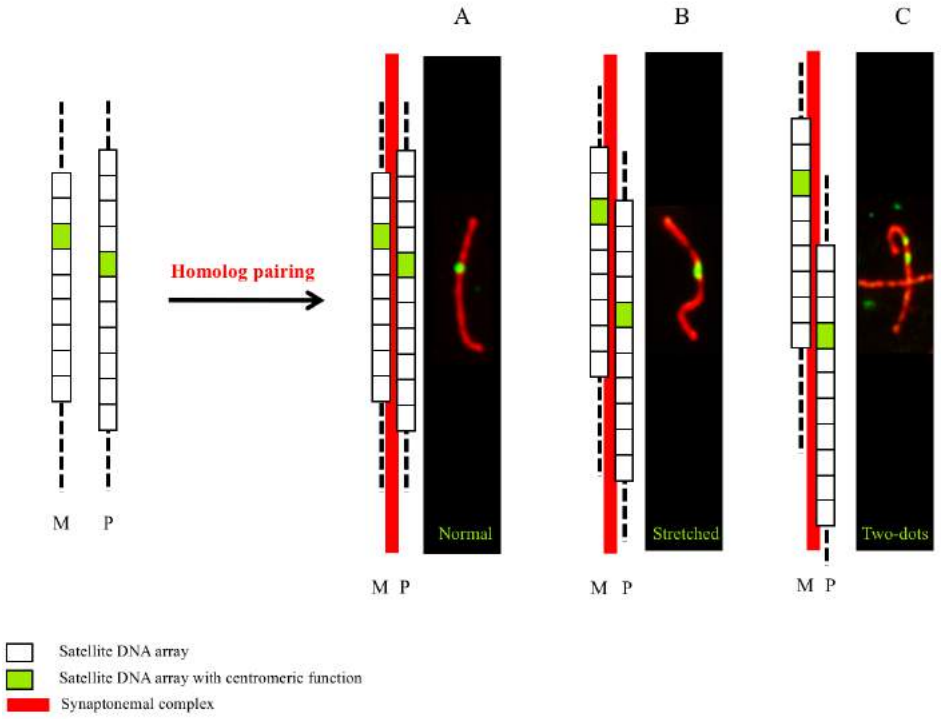


Figure 9. Model for explaining the observation of double-spotted centromeres. On the left, two homologous chromosomes (maternal M and paternal P) with different centromeric position within the satellite DNA arrays with different number of satellite DNA repeats. On the right, possible scenarios after homolog pairing: centromeres remain too close to be resolved separately (A) or become sufficiently distant to be visualized as stretched (B) or "two-dots" (C) signals.

Conclusions

In conclusion, we tested whether a centromere void of satellite DNA exerts the same inhibition on meiotic recombination at the same level of the satellite-based ones. Our results on the distribution of MLH1 recombination foci suggested that the centromere of ECA11 suppresses meiotic recombination as a satellite-based centromere.

Moreover, we identified a peculiar phenomenon during homolog pairing in horse meiosis: a variable number of bivalents, ranging from 0 to 7, displayed double-spotted centromeres. This observation was interpreted as the result of the combination of centromere sliding and misalignment between satellite DNA arrays during homolog pairing. Although single misaligned centromeres were reported in two mammalian species (Borodin et al. 2008, Bikchurina et al. 2018), *E. caballus* showed an exceptional high frequency of these double-spotted centromeres, tempting to speculate that in this species also satellite-based centromeres are highly plastic in CENP-A domain positioning and/or in the length of the satellite-DNA array.

Bibliography

Ahrens E Stranzinger G. Comparative chromosomal studies of E. caballus (ECA) and E. przewalskii (EPR) in a female F1 hybrid. J Anim Breed Genet. 2005;122:97-102

Alexandrov I, Kazakov A, Tumeneva I, Shepelev V, Yurov Y. Alpha-satellite DNA of primates: old and new families. Chromosoma. 2001;110:253-266

Allshire RC, Karpen GH. Epigenetic regulation of centromeric chromatin: old dogs, new tricks? Nat Rev Genet. 2008;9,923-937

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

Anglana M, Bertoni L, Giulotto E. Cloning of a polymorphic sequence from the nontranscribed spacer of horse rDNA. Mamm Genome 1996;7:539-541

Baccarini P. Sulle cinesi vegetative del "Cynomorium coccineum L.". N Giorn Bot Ital N Ser. 1908; 15:189-203

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. Recent segmental duplications in the human genome. Science. 2002;297:1003-1007

Baudat F, Imai Y, de Massy B. Meiotic recombination in mammals: localization and regulation. Nat Rev Genet. 2013;14:794-806

Beadle GW. A Possible Influence of the Spindle Fibre on Crossing-Over in Drosophila. PNAS. 1932;18:160-165

Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27:573-80

Bergmann JH, Martins NMC, Larionov V, Masumoto H, Earnshaw WC. HACking the centromere chromatin code: insights from human artificial chromosomes. Chromosome Res. 2012;20:505-519

Bertoni L, Attolini C, Faravelli M, Simi S, Giulotto E. Intrachromosomal telomere-like DNA sequences in Chinese hamster. Mamm Genome. 1996;7:853-855

Bertoni L, Attolini C, Tessera L, Mucciolo E, Giulotto E. Telomeric and non-telomeric (TTAGGG)n sequences in gene amplification and chromosome stability. Genomics. 1994;24:53-62

Bikchurina TI, Tishakova KV, Kizilova EA, Romanenko SA, Serdyukova NA, Torgasheva AA, Borodin PM. Chromosome Synapsis and Recombination in Male-Sterile and Female-Fertile Interspecies Hybrids of the Dwarf Hamsters (Phodopus, Cricetidae). Genes (Basel). 2018;9

Blackburn EH. Structure and function of telomeres. Nature. 1991;235:305-311

Blower MD, Sullivan BA, Karpen GH. Conserved organization of centromeric chromatin in flies and humans. Dev Cell. 2002;2:319-330

Borodin PM, Karamysheva TV, Belonogova NM, Torgasheva AA, Rubtsov NB, Searle JB. Recombination map of the common shrew, *Sorex araneus* (Eulipotyphla, Mammalia). Genetics. 2008;178:621-632.

Burrack LS, Hutton HF, Matter KJ, Clancey SA, Liachko I, Plemmons AE, Saha A, Power EA, Turman B, Thevandavakkam MA, Ay F, Dunham MJ, Berman J. Neocentromeres provide chromosome segregation accuracy and centromere clustering to multiple loci along a Candida albicans chromosome. PLoS Genet. 2016;12:e1006317

Cacheux L, Ponger L, Gerbault-Seureau M, Richard FA, Escudé C. Diversity and distribution of alpha satellite DNA in the genome of an Old World monkey: Cercopithecus solatus. BMC Genomics. 2016 Nov 14;17:916

Capilla L, Garcia Caldés M, Ruiz-Herrera A. Mammalian Meiotic Recombination: A Toolbox for Genome Evolution. Cytogenet Genome Res. 2016;150:1-16

Carbone L, Nergadze SG, Magnani E, Misceo D, Cardone MF, Roberto R, Bertoni L, Attolini C, Piras MF, de Jong P, Raudsepp T, Chowdhary BP, Guérin G, Archidiacono N, Rocchi M, Giulotto E. Evolutionary movement of centromeres in horse, donkey, and zebra. Genomics 2006;87:777-782

Casola C, Hucks D, Feschotte C. Convergent domestication of pogo-like transposases into centromere-binding proteins in fission yeast and mammals. Mol Biol Evol. 2007;25:29-41

Castillo AG, Pidoux AL, Catania S, Durand-Dubief M, Choi ES, Hamilton G, Ekwall K, Allshire RC. Telomeric repeats facilitate CENP-A(Cnp1) incorporation via telomere binding proteins. PLoS One. 2013;8:e69673

Cerutti F, Gamba R, Mazzagatti A, Piras FM, Cappelletti E, Belloni E, Nergadze SG, Raimondi E, Giulotto E. The major horse satellite DNA family is associated with centromere competence. Mol. Cytogenet. 2016;9:35

Choo KH. Why is the centromere so cold? Genome Res. 1998;8:81-82.

Choo KHA. Centromerization. Trends Cell Biol. 2000;10:182-188

Clarke L. Centromeres: proteins, protein complexes, and repeated domains at centromeres of simple eukaryotes. Curr Opin Genetics Dev. 1998;8:212-218

Cleveland DW, Mao Y, Sullivan KF. Centromeres and kinetochores. Cell. 2003;112:407-421

Clowney EJ, LeGros MA, Mosley CP, Clowney FG, Markenskoff-Papadimitriou EC, Myllys M, Barnea G, Larabell CA, Lomvardas S. Nuclear aggregation of olfactory receptor genes governs their monogenic expression. Cell. 2012;151:724-737

Cooke CA. CENP-B: a major human centromere protein located beneath the kinetochore. J Cell Biol. 1990;110:1475-1488

Da Ines O, White CI. Centromere Associations in Meiotic Chromosome Pairing. Annu Rev Genet. 2015;49:95-114

Dai X, Otake K, You C, Cai Q, Wang Z, Masumoto H, Wang Y. Identification of novel α-n-methylation of CENP-B that regulates its binding to the centromeric DNA. J Proteome Res. 2013;12:4167-4175

Dawe RK, Hiatt EN. Plant neocentromeres: fast, focused, and driven. Chromosome Res. 2004;12, 655-669

Earnshaw WC, Rothfield N. Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma. Chromosoma. 1985;91:313-321

Earnshaw WC, Sullivan KF, Machlin PS, Cooke CA, Kaiser DA, Pollard TD, Rothfield NF, Cleveland DW. Molecular cloning of cDNA for CENP-B, the major human centromere autoantigen. J. Cell Biol. 1987;104:817-829

Earnshaw WC. Discovering centromere proteins: from cold white hands to the A, B, C of CENPs. Nat Rev Mol Cell Biol. 2015;16:443-449

Fachinetti D, Folco HD, Nechemia-Arbely Y, Valente LP, Nguyen K, Wong AJ, Zhu Q, Holland AJ, Desai A, Jansen LE, Cleveland DW. A two-step mechanism for epigenetic specification of centromere identity and function. Nat Cell Biol. 2013;15:1056-1066

Fachinetti D, Han JS, McMahon MA, Ly P, Abdullah A, Wong AJ, Cleveland DW. DNA Sequence-Specific Binding of CENP-B Enhances the Fidelity of Human Centromere Function. Dev Cell. 2015;33:314-327

Fang Y, Spector DL. Centromere positioning and dynamics in living Arabidopsis plants. Mol Biol Cell. 2005;16:5710-5718

Fantaccione S, Pontecorvo G, Zampella V. Molecular characterization of the first satellite DNA with CENP-B and CDEIII motifs in the bat Pipistrellus kuhli. FEBS Lett. 2005;579:2519-2527

Faravelli M, Moralli D, Bertoni L, Attolini C, Chernova O, Raimondi E, Giulotto E. Two extended arrays of a satellite DNA sequence at the centromere and at the short-arm telomere of Chinese hamster chromosome 5. Cytogenet Cell Genet. 1998;83:281-286

Fishman L, Saunders A. Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers. Science. 2008;322:1559-1562

Fowler KJ, Hudson DF, Salamonsen LA, Edmondson SR, Earle E, Sibson MC, Choo KH. Uterine dysfunction and genetic modifiers in centromere protein B-deficient mice. Genome Res. 2000;10:30-41

Fraune J, Brochier-Armanet C, Alsheimer M, Volff JN, Schücker K, Benavente R. Evolutionary history of the mammalian synaptonemal complex. Chromosoma. 2016;125:355-60

Fry K, Salser W. Nucleotide sequences of HS-alpha satellite DNA from kangaroo rat Dipodomys ordii and characterization of similar sequences in other rodents. Cell. 1977;12:1069-1084

Fujita R, Otake K, Arimura Y, Horikoshi N, Miya Y, Shiga T, Osakabe A, Tachiwana H, Ohzeki J, Larionov V, Masumoto H, Kurumizaka H. Stable complex formation of CENP-B with the CENP-A nucleosome. Nucleic Acids Res. 2015;43:4909-4922

Fukagawa T, Earnshaw WC. The centromere: chromatin foundation for the kinetochore machinery. Dev Cell. 2014;30:496-508

Fukagawa T. Critical histone post-translational modifications for centromere function and propagation. Cell Cycle. 2017;16:1259-1265

Funabiki H, Hagan I, Uzawa S, Yanagida M. Cell cycle-dependent specific positioning and clustering of centromeres and telomeres in fission yeast. J Cell Biol. 1993;121:961-976

Gao J, Colaiácovo MP. Zipping and Unzipping: Protein Modifications Regulating Synaptonemal Complex Dynamics. Trends Genet. 2018;34:232-245

Goldberg IG, Sawhney H, Pluta AF, Warburton PE, Earnshaw WC. Surprising deficiency of CENP-B binding sites in African green monkey alpha-satellite DNA: implications for CENP-B function at centromeres. Mol Cell Biol. 1996;16,5156-5168

Grant CE, Bailey TL, Noble WS. FIMO: Scanning for occurrences of a given motif. Bioinformatics. 2011;27:1017-1018

Haaf T, Mater AG, Wienberg J, Ward DC. Presence and abundance of CENP-B box sequences in great ape subsets of primate-specific alpha-satellite DNA. J. Mol. Evol. 1995;41:487-491

Haaf T, Ward DC. Rabl orientation of CENP-B box sequences in Tupaia belangeri fibroblasts. Cytogenet Cell Genet. 1995;70:258-262

He L, Liu J, Torres GA, Zhang H, Jiang J, Xie C. Interstitial telomeric repeats are enriched in the centromeres of chromosomes in Solanum species. Chromosome Res. 2013;21:5-13

Heitz E. Das Heterochromatin der Moose. Jahrb Wiss Botanik. 1928;69:762-818

Henikoff S. The centromere paradox: stable inheritance with rapidly evolving DNA. Science. 2001;293:1098-1102

Hill E, McGivney B, Gu J, Whiston R, MacHugh D. A genome-wide SNP-association study confirms a sequence variant (g. 66493737C> T) in the equine myostatin (MSTN) gene as the most powerful predictor of optimum racing distance for Thoroughbred racehorses. Bmc Genomics. 2010;11:552

Hori T, Shang WH, Takeuchi K, Fukagawa T. The CCAN recruits CENP-A to the centromere and forms the structural core for kinetochore assembly. J Cell Biol. 2013;200:45-60

Hori T, Shang WH, Toyoda A, Misu S, Monma N, Ikeo K, Molina O, Vargiu G, Fujiyama A, Kimura H, Earnshaw WC, Fukagawa T. Histone H4 Lys 20 monomethylation of the CENP-A nucleosome is essential for kinetochore assembly. Dev Cell. 2014;29:740-749

Howman EV, Fowler KJ, Newson AJ, Redward S, MacDonald AC, Kalitsis P, Choo KH. Early disruption of centromeric chromatin organization in centromere protein A (Cenpa) null mice. Proc Natl Acad Sci USA. 2000;97:1148-1153

Hudson DF, Fowler KJ, Earle E, Saffery R, Kalitsis P, Trowell H, Hill J, Wreford NG, de Kretser DM, Cancilla MR, Howman E, Hii L, Cutts SM, Irvine DV, Choo KH. Centromere protein B null mice are mitotically and meiotically normal but have lower body and testis weights. J Cell Biol. 1998;141:309-319

Iwata-Otsubo A, Dawicki-McKenna JM, Akera T, Falk SJ, Chmátal L, Yang K, Sullivan BA, Schultz RM, Lampson MA, Black BE. Expanded satellite repeats amplify a discrete CENP-A nucleosome assembly site on chromosomes that drive in female meiosis. Curr Biol. 2017;27:2365-2373.e8

Jagannathan M, Yamashita YM. Function of Junk: Pericentromeric Satellite DNA in chromosome maintenance. Cold Spring Harb Symp Quant Biol. 2017;82:319-327

Jang CW, Shibata Y, Starmer J, Yee D, Magnuson T. Histone H3.3 maintains genome integrity during mammalian development. Genes Dev. 2015;29:1377-1392

Jin Q, Trelles-Sticken E, Scherthan H, Loidl J. Yeast nuclei display prominent centromere clustering that is reduced in nondividing cells and in meiotic prophase. J Cell Biol. 1998;141:21-29

Jin QW, Fuchs J, Loidl J. Centromere clustering is a major determinant of yeast interphase nuclear organization. J Cell Sci. 2000;113:11

Jones GH, Franklin FC. Meiotic crossing-over: obligation and interference. Cell. 2006;126:246-248.

Jones KW. Chromosomal and nuclear location of mouse satellite DNA in individual cells. Nature 1970;225: 912-915

Kalitsis P, Choo KHA. The evolutionary life cycle of the resilient centromere. Chromosoma. 2012;121:327-340

Kasinathan S, Henikoff S. Non-B-Form DNA Is Enriched at Centromeres. Mol Biol Evol. 2018;35:949-962

Kipling D, Mitchell AR, Masumoto H, Wilson HE, Nicol L, Cooke HJ. CENPB binds a novel centromeric sequence in the Asian mouse Mus caroli. Mol Cell Biol. 1995;15:4009-4020

Kipling D, Warburton PE. Centromeres, CENP-B and Tigger too. Trends in Genet. 1997;13:141-145

Kitagawa K, Masumoto H, Ikeda M, Okazaki T. Analysis of protein-DNA and protein-protein interactions of centromere protein B (CENP-B) and properties of the DNA-CENP-B complex in the cell cycle. Mol Cell Biol. 1995;15:1602-1612

Kouznetsova A, Benavente R, Pastink A, Höög C. Meiosis in mice without a synaptonemal complex. PLoS One. 2011;6:e28255

Krijger PH, de Laat W. Identical cells with different 3D genomes; cause and consequences? Curr Opin Genet Dev. 2013; 23:191-196

Kugou K, Hirai H, Masumoto H, Koga A. Formation of functional CENP-B boxes at diverse locations in repeat units of centromeric DNA in New World monkeys. Sci Rep. 2016;6:27833

Kursel LE, Malik HS. The cellular mechanisms and consequences of centromere drive. Curr Opin Cell Biol. 2018;52:58-65

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357-359

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078-2079

Machanick P, Bailey TL. MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics. 2011;27:1696-1697

Maio JJ. DNA strand reassociation and polyribonucleotide binding in the African green monkey, Cercopithecus aethiops. J Mol Biol. 1971;56:579-595

Malik HS, Bayes JJ. Genetic conflicts during meiosis and the evolutionary origins of centromere complexity. Biochem Soc Trans. 2006;34:569-573

Malik HS. The centromere-drive hypothesis: A simple basis for centromere complexity. Prog Mol Subcell Biol. 2009;48:33-52

Maloney KA, Sullivan LL, Matheny JE, Strome ED, Merrett SL, Ferris A, Sullivan BA. Functional epialleles at an endogenous human centromere. Proc Natl Acad Sci U S A. 2012;109:13704-13709

Manuelidis L. Different central nervous system cell types display distinct and nonrandom arrangements of satellite DNA sequences. Proc Natl Acad Sci USA1984;81:3123-3127

Marshall OJ, Choo HA. Putative CENP-B paralogues are not present at mammalian centromeres. Chromosoma. 2012;121:169-179

Marshall OJ, Chueh AC, Wong LH, Choo KHA. Neocentromeres: new insights into centromere structure, disease development, and karyotype evolution. Am J Hum Genet. 2008;82:261-282

Masumoto H, Masukata H, Muro Y, Nozaki N, Okazaki T. A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. J Cell Biol. 1989;109:1963-1973

Masumoto H, Yoda K, Ikeno M, Kitagawa K, Muro Y, Okazaki T. Properties of CENP-B and its target sequence in a satellite DNA. B.K. Vig, editor. NATO ASI Series. 1993;Vol. H 72. Springer-Verlag, Berlin. 31-43

Masumoto H., Yoda K., Ikeno M., Kitagawa K., Muro Y., Okazaki T. Properties of CENP-B and its target sequence in a satellite DNA. In: Vig B.K. (eds) Chromosome Segregation and Aneuploidy. NATO ASI Series (Series H: Cell Biology) Springer, Berlin, Heidelberg. 1993;72

Mather K. Crossing over and Heterochromatin in the X Chromosome of Drosophila Melanogaster. Genetics. 1939;24:413-435

Mayer R, Brero A, von Hase J, Schroeder T, Cremer T, Dietzel S. Common themes and cell type specific variations of higher order chromatin arrangements in the mouse. BMC Cell Biol. 2005;6:44

McNulty SM, Sullivan LL, Sullivan BA. Human centromeres produce chromosome-specific and array-specific alpha satellite transcripts that are complexed with CENP-A and CENP-C. Dev Cell. 2017;42:226-240

Meyne J, Baker RJ, Hobart HH, Hsu TC, Ryder OA, Ward OG, Wiley JE, Wurster-Hill DH, Yates TL, Moyzis RK. Distribution of nontelomeric sites of (TTAGGG)n telomeric sequences in vertebrate chromosomes. Chromosoma. 1990;99:3-10

Mohibi S, Srivastava S, Wang-France J, Mirza S, Zhao X, Band H, Band V. Alteration/Deficiency in Activation 3 (ADA3) protein, a cell cycle regulator, associates with the centromere through CENP-B and regulates chromosome segregation. J Biol Chem. 2015;290:28299-28310

Montefalcone G, Tempesta S, Rocchi M, Archidiacono N. Centromere repositioning. Genome Res. 1999;9:1184-1188

Moroi Y, Peebles C, Fritzler MJ, Steigerwald J, Tan EM. Autoantibody to centromere (kinetochore) in scleroderma sera. Proc Natl Acad Sci USA. 1980;77:1627-1631

Morozov VM, Giovinazzi S, Ishov AM. CENP-B protects centromere chromatin integrity by facilitating histone deposition via the H3.3-specific chaperone Daxx. Epigenetics Chromatin. 2017;10:63

Muro Y, Masumoto H, Yoda K, Nozaki N, Ohashi M, Okazaki T. Centromere protein B assembles human centromeric alpha-satellite DNA at the 17-bp sequence, CENP-B box. J Cell Biol. 1992;116: 585-596

Musacchio A, Salmon ED. The spindle-assembly checkpoint in space and time. Nat Rev Mol Cell Biol. 2007;8:379-393

Musilova P, Kubickova S, Vahala J, Rubes J. Subchromosomal karyotype evolution in Equidae. Chromosome Res. 2013;21:175-187

Musilova P, Kubickova S, Zrnova E, Horin P, Vahala J, Rubes J. Karyotypic relationships among Equus grevyi, Equus burchelli and domestic horse defined using horse chromosome arm-specific probes. Chromosome Res. 2007;15:807-813

Myka JL, Lear TL, Houck ML, Ryder OA, Bailey E. Homologous fission event(s) implicated for chromosomal polymorphisms among five species in the genus Equus. Cytogenet Genome Res. 2003;102:217-221

Nagaki K. Visualization of diffuse centromeres with centromere-specific histone H3 in the holocentric plant Luzula nivea. Plant Cell. 2005;17:1886-1893

Nanda I, Schrama D, Feichtinger W, Haaf T, Schartl M, Schmid M. Distribution of telomeric (TTAGGG)(n) sequences in avian chromosomes. Chromosoma. 2001;111:215-227

Navrátilová A, Koblízková A, Macas J. Survey of extrachromosomal circular DNA derived from plant satellite repeats. BMC Plant Biol. 2008;8:90

Nergadze SG, Belloni E, Piras FM, Khoriauli L, Mazzagatti A, Vella F, Bensi M, Vitelli V, Giulotto E, Raimondi E. Discovery and comparative analysis of a novel satellite, EC137, in horses and other equids. Cytogenet Genome Res. 2014;144:114-123

Nergadze SG, Farnung BO, Wischnewski H, Khoriauli L, Vitelli V, Chawla R, Giulotto E, Azzalin CM. CpG-island promoters drive transcription of human telomeres. RNA. 2009;15:2186-2194

Nergadze SG, Piras FM, Gamba R, Corbo M, Cerutti F, McCarter JGW, Cappelletti E, Gozzo F, Harman RM, Antczak DF, Miller D, Scharfe M, Pavesi G, Raimondi E, Sullivan KF, Giulotto E. Birth, evolution and transmission of satellite-free mammalian centromeric domains. Genome Res. 2018;28:789-799

Nokkala S, Puro J. Cytological evidence for a chromocenter in Drosophila melanogaster oocytes. Hereditas. 1976;83:265-268

Ohzeki JI, Nakano M, Okada T, Masumoto H. CENP-B box is required for de novo centromere chromatin assembly on human alphoid DNA. J Cell Biol. 2002;159:765-775

Okada T, Ohzeki J, Nakano M, Yoda K, Brinkley WR, Larionov V, Masumoto H. CENP-B controls centromere formation depending on the chromatin context. Cell. 2007;131:1287-1300

Olszak AM, van Essen D, Pereira AJ, Diehl S, Manke T, Maiato H, Saccani S, Heun P. Heterochromatin boundaries are hotspots for de novo kinetochore formation. Nat Cell Biol. 2011;13:799-808

Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, Schubert M, Cappellini E, Petersen B, Moltke I, Johnson PL, Fumagalli M, Vilstrup JT, Raghavan M, Korneliussen T et al. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature 2013;499:74-78

Padeken J, Mendiburo MJ, Chlamydas S, Schwarz HJ, Kremmer E, Heun P. The nucleoplasmin homolog NLP mediates centromere clustering and anchoring to the nucleolus. Mol Cell. 2013;50:236-249

Peters AH, Plug AW, van Vugt MJ, de Boer P. A drying-down technique for the spreading of mammalian meiocytes from the male and female germline. Chromosome Res. 1997;5:66-68

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

Piras FM, Nergadze SG, Poletto V, Cerutti F, Ryder OA, Leeb T, Raimondi E, Giulotto E. Phylogeny of horse chromosome 5q in the genus Equus and centromere repositioning. Cytogenet Genome Res. 2009;126:165-172

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

Plohl M, Meštrović N, Mravinac B. Centromere identity from the DNA point of view. Chromosoma. 2014;123:313-325

Plohl M, Meštrović N, Mravinac B. Satellite DNA evolution. Genome Dyn. 2012;7:126-152

Politz JC, Scalzo D, Groudine M. Something silent this way forms: the functional organization of the repressive nuclear compartment. Annu Rev Cell Dev Biol. 2013;29:241-270

Purgato S, Belloni E, Piras FM, Zoli M, Badiale C, Cerutti F, Mazzagatti A, Perini G, Della Valle G, Nergadze SG, Sullivan KF, Raimondi E, Rocchi M, Giulotto E. Centromere sliding on a mammalian chromosome. Chromosoma. 2015;124:277-287

Rhoades MM, Vilkomerson H. On the anaphase movement of chromosomes. Proc Natl Acad Sci USA. 1942;28:433-436

Ribeiro SA, Vagnarelli P, Dong Y, Hori T, McEwen BF, Fukagawa T, Flors C, Earnshaw WC. A super-resolution map of the vertebrate kinetochore. Proc Natl Acad Sci USA. 2010;107:10484-10489

Rocchi M, Archidiacono N, Schempp W, Capozzi O, Stanyon R. Centromere repositioning in mammals. Heredity. 2012;108:59-67

Ronneberger O, Baddeley D, Scheipl F, Verveer PJ, Burkhardt H, Cremer C, Fahrmeir L, Cremer T, Joffe B. Spatial quantitative analysis of fluorescently labeled nuclear structures: problems, methods, pitfalls. 2008;16:523-562

Rošić S, Erhardt S. No longer a nuisance: long non-coding RNAs join CENP-A in epigenetic centromere regulation. Cell Mol Life Sci. 2016;73:1387-1398

Saffery R, Irvine DV, Griffiths B, Kalitsis P, Wordeman L, Choo KH. Human centromeres and neocentromeres show identical distribution patterns of >20 functionally important kinetochore-associated proteins. Hum Mol Genet. 2000;9:175-185

Saitoh H, Tomkiel J, Cooke CA, Ratrie H, Maurer M, Rothfield NF, Earnshaw WC. CENP-C, an autoantigen in scleroderma, is a component of the human inner kinetochore plate. Cell. 1992;70:115-125

Salser W, Bowen S, Browne D, el-Adli F, Fedoroff N, Fry K, Heindell H, Paddock G, Poon R, Wallace B, Whitcome P. Investigation of the organization of mammalian chromosomes at the DNA sequence level. Fed Proc. 1976;35:23-35

Sambrook J, Russel DW. Molecular Cloning: a laboratory manual, 3rd edition. Vol II. Cold Spring Harbor Laboratory Press. 2001.

Santaguida S, Musacchio A. The life and miracles of kinetochores. EMBO J. 2009;28:2511-2531

Sanyal K, Baum M, Carbon J. Centromeric DNA sequences in the pathogenic yeast Candida albicans are all different and unique. Proc Natl Acad Sci U S A. 2004;101:11374-11379

Saxena A, Saffery R, Wong LH, Kalitsis P, Choo KHA. Centromere proteins Cenpa, Cenpb, and Bub3 interact with poly(ADP-

ribose) polymerase-1 protein and are poly(ADP-ribosyl)ated. J Biol Chem. 2002;277:26921-26926.

Schueler MG, Sullivan BA. Structural and functional dynamics of human centromeric chromatin. Annu Rev Genomics Hum Genet. 2006;7:301-313

Shampay J Schmitt M, Bassham S. A novel minisatellite at a cloned hamster telomere. Chromosoma. 1995;104:29-38

Shepelev VA, Alexandrov AA, Yurov YB, Alexandrov IA. 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. 2009;5:e1000641

Solovei I, Cremer M. 3D-FISH on cultured cells combined with immunostaining. Methods Mol Biol. 2010;659:117-126

Solovei I, Grandi N, Knoth R, Volk B, Cremer T. Positional changes of pericentromeric heterochromatin and nucleoli in postmitotic Purkinje cells during murine cerebellum development. Cytogenet Genome Res. 2004;105:302-310

Solovei I, Kreysing M, Lanctôt C, Kösem S, Peichl L, Cremer T, Guck J, Joffe B. Nuclear architecture of rod photoreceptor cells adapts to vision in mammalian evolution. Cell. 2009;137(2):356-68.

Solovei I, Thanisch K, Feodorova Y. How to rule the nucleus: divide et impera. Curr Opin Cell Biol. 2016;40:47-59

Stitou S, Díaz de la Guardia R, Jiménez R, Burgos M. Isolation of a species-specific satellite DNA with a novel CENP-B-like box from the North African rodent Lemniscomys barbarus. Exp Cell Res. 1999;250:381-386

Straub T. Heterochromatin dynamics. PLoS Biol. 2003;1:E14

Sugimoto K, Yata H, Muro Y, Himeno M. Human centromere protein C (CENP C) is a DNA-binding protein which possesses a novel DNA-binding motif. J Biochem. 1994;116:877-881

Sujiwattanarat P, Thapana W, Srikulnath K, Hirai Y, Hirai H, Koga A. Higher-order repeat structure in alpha satellite DNA occurs in New World monkeys and is not confined to hominoids. Sci Rep. 2015;5:10315

Sullivan KF, Glass CA. CENP-B is a highly conserved mammalian centromere protein with homology to the helix-loop-helix family of proteins. Chromosoma. 1991;100:360-370

Sullivan LL, Chew K, Sullivan BA. α satellite DNA variation and function of the human centromere. Nucleus. 2017;8:331-339

Suntronpong A, Kugou K, Masumoto H, Srikulnath K, Ohshima K, Hirai H, Koga A. CENP-B box, a nucleotide motif involved in centromere formation, occurs in a New World monkey. Biol Lett. 2016;12:20150817

Suzuki N, Nakano M, Nozaki N, Egashira S, Okazaki T, Masumoto H. CENP-B interacts with CENP-C domains containing Mif2 regions responsible for centromere localization. J Biol Chem. 2004;279, 5934-5946

Syrjänen JL, Pellegrini L, Davies OR. A molecular model for the role of SYCP3 in meiotic chromosome organisation. Elife. 2014;3:eLife.02963

Talbert PB, Henikoff S. Centromeres convert but don't cross. PLoS Biol. 2010;8(3):e1000326

Tan T, Chen Z, Lei Y, Zhu Y, Liang Q. A regulatory effect of INMAP on centromere proteins: antisense INMAP induces CENP-B variation and centromeric halo. PLoS ONE. 2014;9:e91937

Tanaka Y, Kurumizaka H, Yokoyama S. CpG methylation of the CENP-B box reduces human CENP-B binding. FEBS J. 2005;272:282-289

Tanaka Y, Nureki O, Kurumizaka H, Fukai S, Kawaguchi S, Ikuta M, Iwahara J, Okazaki T, Yokoyama S. Crystal structure of the CENP-B protein-DNA complex: the DNA-binding domains of CENP-B induce kinks in the CENP-B box DNA. EMBO J. 2001;20:6612-6618

Tawaramoto MS, Park SY, Tanaka Y, Nureki O, Kurumizaka H, Yokoyama S. Crystal structure of the human centromere protein B (CENP-B) dimerization domain at 1.65-A resolution. J Biol Chem. 2003;278:51454-51461

Tjong H, Li W, Kalhor R, Dai C, Hao S, Gong K, Zhou Y, Li H, Zhou XJ, Le Gros MA, Larabell CA, Chen L, Alber F. Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. Proc Natl Acad Sci U S A. 2016;113:E1663-1672

Tolomeo D, Capozzi O, Stanyon RR, Archidiacono N, D'Addabbo P, Catacchio CR, Purgato S, Perini G, Schempp W, Huddleston J, Malig M, Eichler EE, Rocchi M. Epigenetic origin of evolutionary novel centromeres. Sci Rep. 2017;7:41980

Trifonov VA, Musilova P, Kulemsina AI. Chromosome evolution in Perissodactyla. Cytogenet Genome Res. 2012;137:208-217

Trifonov VA, Stanyon R, Nesterenko AI, Fu B, Perelman PL, O'Brien PC, Stone G, Rubtsova NV, Houck ML, Robinson TJ, Ferguson-Smith MA, Dobigny G, Graphodatsky AS, Yang F. Multidirectional cross-species painting illuminates the history of karyotypic evolution in Perissodactyla. Chromosome Res. 2008;16:89-107

Van de Werken HJG1, Haan JC, Feodorova Y, Bijos D, Weuts A, Theunis K, Holwerda SJB, Meuleman W, Pagie L, Thanisch K, Kumar P, Leonhardt H, Marynen P, van Steensel B, Voet T, de Laat W, Solovei I, Joffe B. Small chromosomal regions position themselves autonomously according to their chromatin class. Genome Res. 2017;27:922-933

Ventura M, Weigl S, Carbone L, Cardone MF, Misceo D, Teti M, D'Addabbo P, Wandall A, Björck E, de Jong PJ, She X, Eichler EE, Archidiacono N, Rocchi M. Recurrent sites for new centromere seeding. Genome Res. 2004 Sep;14:1696-1703

Vidale P, Magnani E, Nergadze SG, Santagostino M, Cristofari G, Smirnova A, Mondello C, Giulotto E. The catalytic and the RNA subunits of human telomerase are required to immortalize equid primary fibroblasts. Chromosoma. 2012;121:475-488

Voullaire LE, Slater HR, Petrovic V, Choo KHA. A functional marker centromere with no detectable alpha-satellite, satellite III, or CENP-B protein: activation of a latent centromere? Am J Hum Genet. 1993;52:1153-1163

Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, Lear TL, Adelson DL, Bailey E, Bellone RR, Blöcker H, Distl O, Edgar RC, Garber M, Leeb T et al. Genome sequence, comparative analysis, and population genetics of the domestic horse. Science. 2009; 326:865-867

Walter J, Joffe B, Bolzer A, Albiez H, Benedetti PA, Muller S, Speicher MR, Cremer T, Cremer M, Solovei I. Towards many colors in FISH on 3D-preserved interphase nuclei. Cytogenet Genome Res. 2006;114:367-378

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

Warburton PE, Willard HF. Interhomologue sequence variation of alpha satellite DNA from human chromosome 17: evidence for concerted evolution along haplotypic lineages. J Mol Evol. 1995;41:1006-1015

Waye JS, Willard HF. Molecular analysis of a deletion polymorphism in alpha satellite of human chromosome 17: evidence for homologous unequal crossing-over and subsequent fixation. Nucleic Acids Res. 1986;14:6915-6927

Weierich C, Brero A, Stein S, von Hase J, Cremer C, Cremer T, Solovei I. Three-dimensional arrangements of centromeres and telomeres in nuclei of human and murine lymphocytes. Chromosome Res. 2003;11:5

Westermann S, Schleiffer A. Family matters: structural and functional conservation of centromere-associated proteins from yeast to humans. Trends Cell Biol. 2013;23:260-269

Willard HF, Waye JS. Hierarchical order in chromosomespecific human alpha satellite DNA. Trends Genet. 1987;3:192-198

Willard HF. Evolution of alpha satellite. Curr Opin Genet Dev. 1991;1:509-514

Wu ZA, Liu WX, Murphy C, Gall J. Satellite 1 DNA sequence from genomic DNA of the giant panda Ailuropoda melanoleuca. Nucleic Acids Res. 1990;18:1054

Yoda K, Ando S, Okuda A, Kikuchi A, Okazaki T. In vitro assembly of the CENP-B/alpha-satellite DNA/core histone complex: CENP-B causes nucleosome positioning. Genes Cells. 1998;3:533-548

Yoda K, Kitagawa K, Masumoto H, Muro Y and Okazaki T. A human centromere protein, CENP-B, has a DNA binding domain containing four potential alpha helices at the NH2 terminus, which is separable from dimerizing activity. J Cell Biol. 1992;119:1413-1427

Yoda K, Nakamura T, Masumoto H, Suzuki N, Kitagawa K, Nakano M, Shinjo A, Okazaki T. Centromere protein B of African green monkey cells: gene structure, cellular expression, and centromeric localization. Mol Cell Biol. 1996;16:5169-5177

Zickler D, Kleckner N. Meiotic chromosomes: integrating structure and function. Annu Rev Genet. 1999;33:603-754

Attached publications

Cerutti F*, Gamba R*; Mazzagatti A*, Piras FM, **Cappelletti E**, Belloni E, Nergadze SG, Raimondi E, Giulotto E. The major horse satellite DNA family is associated with centromere competence. Molecular Cytogenetics, 2016.

Nergadze SG*, Piras FM*, Gamba R*, Corbo M*, Cerutti F, McCarter JGW, **Cappelletti E**, Gozzo F, Harman RM, Antczak DF, Miller D, Scharfe M, Pavesi G, Raimondi E, Sullivan KF, Giulotto E. Birth, evolution and transmission of satellite-free mammalian centromeric domains. Genome Research, 2018.

List of meeting abstracts

Cappelletti E, I. Solovei I, Piras FM, Corbo M, Nergadze SG, Giulotto E. Satellite DNA is responsible for centromere clustering in mammals. XV Congress of the Italian Federation of Life Sciences, Rome, Italy, 18-21 September 2018 (poster presentation).

Cappelletti E, Corbo M, Piras FM, Rausa A, Di Mauro RM, Bailey E, Nergadze SG, Giulotto E. Functional annotations of horse centromeres. 12th Dorothy Russell Havemeyer International Horse Genome Workshop – Pavia, Italy, 12-15 September 2018 (oral presentation).

Corbo M, Piras FM, **Cappelletti E**, Faravelli S, Colantuoni M, Bailey E, Raimondi E, Nergadze SG, Giulotto E. Birth, evolution and transmission of equid centromeres. 12th Dorothy Russell Havemeyer International Horse Genome Workshop – Pavia, Italy, 12-15 September 2018.

Del Giudice S (1), **Cappelletti E** (2), Khoriauli L, Piras FM, Nergadze SG, Giulotto E. 1) Investigating the function of telomeric-repeat containing RNA. 2) Satellite DNA is responsible for centromere clustering in mammals. 2nd Joint Annual Symposium of the Departments of Biology Biotechnology, Molecular Medicine and CNR-Institute of Molecular Genetics – Pavia, Italy, 20-22 June 2018 (oral presentation).

Corbo M, Roberti A, Piras FM, **Cappelletti E**, Bensi M, Nergadze SG, Raimondi E, Giulotto E. The epigenetic landscape of mammalian centromeres. 2nd Joint Annual Symposium of the Departments of Biology Biotechnology, Molecular Medicine and CNR-Institute of Molecular Genetics – Pavia, Italy, 20-22 June 2018.

Roberti A, Nergadze SG, Bensi M, Gamba R, Corbo M, Piras FM, **Cappelletti E**, Giulotto E, Raimondi E. Epigenetic modifications at satellite-less evolutionarily new centromeres. Congress of the Italian Geneticists Association AGI – Cortona, Italy, 7-9 September 2017.

Nergadze SG, Gamba R, Piras FM, **Cappelletti E**, Corbo M, Gozzo F, Miller D, Antczak D, Raimondi E, Sullivan K, Giulotto E. Epigenetic characterization of centromeric chromatin in equids. ISAG 36th International Society for Animal Genetics Conference – Dublin, Ireland 16-21 July 2017.

Gamba R, Nergadze SG, Piras FM, **Cappelletti E**, Corbo M, Gozzo F, McCarter J, Boero E, Tavella S, Miller D, Antczak D, Raimondi E, Sullivan K, Giulotto E. The epigenetic landscape of equid centromeres: a molecular approach. XIV Congress of the Italian Federation of Life Sciences – Rome, Italy, 20-23 September 2016.

Nergadze SG, Gamba R, Piras FM, Corbo M, **Cappelletti E**, Mazzagatti A, Gozzo F, McCarter J, Miller D, Antczak D, Raimondi E, Sullivan K, Giulotto E. Functional organization of centromeric chromatin in the absence of satellite DNA: the equid model system. EMBO Workshop: Chromosome segregation and aneuploidy – Galway, Ireland, 25-29 June 2016.

Cappelletti E. Characterization of the DNA binding sites of the centromeric protein CENP-B in the genus Equus. EPIGEN workshop: Data on the beach – Rimini, Italy, 5-6 May 2016 (oral presentation).

Nergadze SG, Cerutti F, Gamba R, Piras FM, Corbo M, Badiale C, **Cappelletti E**, McCarter J, Miller D, Antczak D, Raimondi E, Sullivan K, Giulotto E. Functional organization and inheritance of satellite-less equid centromeric domains. Plant and Animal Genome XXIV Conference – San Diego, California, 9-13 January 2016.

Nergadze SG, Cerutti F, Piras FM, Gamba G, Mazzagatti A, Corbo M, Badiale C, **Cappelletti E**, Raimondi E, Giulotto E. Functional organization of horse centromeres: a genome wide analysis. 11th Dorothy Russell Havemeyer Foundation International Equine Genome Mapping Workshop – Hannover, Germany, 22-25 July 2015.

Nergadze SG, Cerutti F, Gamba R, Piras FM, Mazzagatti A, Corbo M, Badiale C, **Cappelletti E**, McCarter J, Sullivan K, Raimondi E, Giulotto E. Functional organization of satellite-less equid centromeres. 10th European Cytogenetics Conference – Strasbourg, France, 4-7 July 2015.

Badiale C, Nergadze SG, Cerutti F, Gamba R, Piras FM, Mazzagatti A, Corbo M, **Cappelletti E**, McCarter J, Sullivan K, Raimondi E, Giulotto E. Epigenetic specification of the centromeric function in the absence of satellite DNA. 11th Seminar of the Italian Society for Biophysics and Molecular Biology: "From Genomes to Functions" – Turin, Italy. 1-3 July 2015.

Cerutti et al. Molecular Cytogenetics (2016) 9:35
DOI 10.1186/s13039-016-0242-z

Molecular Cytogenetics

RESEARCH Open Access

The major horse satellite DNA family is associated with centromere competence

Federico Cerutti[†^], Riccardo Gamba[†], Alice Mazzagatti[†], Francesca M. Piras, Eleonora Cappelletti, Elisa Belloni, Solomon G. Nergadze, Elena Raimondi[*] and Elena Giulotto[*]

Abstract

Background: The centromere is the specialized locus required for correct chromosome segregation during cell division. The DNA of most eukaryotic centromeres is composed of extended arrays of tandem repeats (satellite DNA). In the horse, we previously showed that, although the centromere of chromosome 11 is completely devoid of tandem repeat arrays, all other centromeres are characterized by the presence of satellite DNA. We isolated three horse satellite DNA sequences (37cen, 2P1 and EC137) and described their chromosomal localization in four species of the genus Equus.

Results: In the work presented here, using the ChIP-seq methodology, we showed that, in the horse, the 37cen satellite binds CENP-A, the centromere-specific histone-H3 variant. The 37cen sequence bound by CENP-A is GC-rich with 221 bp units organized in a head-to-tail fashion. The physical interaction of CENP-A with 37cen was confirmed through slot blot experiments. Immuno-FISH on stretched chromosomes and chromatin fibres demonstrated that the extension of satellite DNA stretches is variable and is not related to the organization of CENP-A binding domains. Finally, we proved that the centromeric satellite 37cen is transcriptionally active.

Conclusions: Our data offer new insights into the organization of horse centromeres. Although three different satellite DNA families are cytogenetically located at centromeres, only the 37cen family is associated to the centromeric function. Moreover, similarly to other species, CENP-A binding domains are variable in size. The transcriptional competence of the 37cen satellite that we observed adds new evidence to the hypothesis that centromeric transcripts may be required for centromere function.

Keywords: Horse genome, Centromere, Satellite DNA, Next generation sequencing, High resolution cytogenetics

Background

In mammals, a significant fraction of the genome is constituted by extended stretches of tandemly repeated DNA. It was shown that these highly repetitive sequences can give rise to satellite bands in gradient centrifugation experiments when they have a different GC content compared to bulk genomic DNA [1]; therefore, they were defined "satellite" DNA. In most eukaryotic chromosomes, these non-coding sequences are the main DNA component of centromeric and pericentromeric heterochromatin [2–6].

Although the centromeric function is highly conserved through eukaryotes, centromeric satellite DNA is rapidly evolving, often being species specific [6–8]. Moreover, following our initial description of a centromere completely devoid of satellite DNA in the horse [9], other examples of naturally occurring satellite-less centromeres were observed in plants and animals [10–13]. These observations raise the challenging question whether centromeric and pericentromeric satellites have a functional role. A number of hypotheses have been proposed to explain the recruitment, by the majority of eukaryotic centromeres, of large stretches of satellite DNA. Satellite DNA may facilitate binding of the centromere specific histone CENP-A (the main epigenetic mark of centromere function) to centromeric chromatin

* Correspondence: elena.raimondi@unipv.it; elena.giulotto@unipv.it
[†]Equal contributors
[^]Deceased
Dipartimento di Biologia e Biotecnologie, Università di Pavia, Via Ferrata 1, 27100 Pavia, Italy

[14]. In addition, centromeric repetitive DNA, typically devoid of active genes, may aid the formation of a heterochromatic environment which would favour the stability of the chromosome during mitosis and meiosis [6, 7, 15]. In several species, centromeric satellite DNA is transcribed and it has been suggested that these transcripts may play a role in heterochromatin formation. Transcription of the centromeric regions seems to be important for chromatin opening and CENP-A loading; these transcripts are believed to provide a flexible scaffold that allows assembly or stabilization of the kinetochore proteins and may act *in trans* on all or on a subset of chromosomes, independently of the primary DNA sequence [16–18].

In a previous work, we isolated two horse satellites, 37cen and 2PI, from a genomic library in lambda phage [19], and investigated their chromosomal distribution in four equid species [10]. More recently [20], we described a new horse satellite, EC137, which is less abundant than 37cen and 2PI and mostly pericentromeric. In the horse, 37cen, 2PI and EC137 are present, together or individually, at all primary constrictions, with the exception of the centromere of chromosome 11 which is completely satellite-free [9, 10, 21]. In this work, we applied next-generation DNA sequencing and high-resolution cytogenetic approaches to identify the satellite repeat bearing the centromeric function in the horse and we proved that this satellite is transcriptionally active.

Results and discussion

Molecular identification of the functional centromeric satellite DNA

The aim of the present work was to define the satellite DNA repeats bearing the centromeric function in the horse. To this purpose, an anti-CENP-A antibody [9, 21] was used in immunoprecipitation experiments with chromatin from horse skin primary fibroblasts. DNA purified from immunoprecipitated and from control non-immunoprecipitated chromatin (input) was paired-end sequenced through an Illumina HiSeq 2000 platform. A total of 78,207,302 and 41,155,660 high-quality reads were obtained from ChIP and input samples, respectively. It is important to remind that most mammalian centromeres are not assembled due to their highly repetitive nature and that all mammalian genome data bases include a "virtual" chromosome, named "unplaced", composed of contigs containing highly repetitive DNA sequences (a number of which are located at the centromeres) that lack chromosome assignment. Therefore, in the EquCab2.0 reference genome, we expected to identify most of the centromeric repeats binding CENP-A in "unplaced" contigs. Each contig is identified by a number which is unrelated to its genomic location.

Sequence reads were aligned through Bowtie 2.0 [22] to the horse reference genome (EquCab2.0, 2007 release). Peak-calling was performed with the default parameters of MACS 2.0.10 software [23] using the input reads as control dataset and applying stringent criteria (see Materials and Methods) to select significantly enriched regions [24]. A total of 1705 regions mapping on 1462 unplaced contigs were significantly enriched, as shown in Additional file 1: Table S1.

The sequence of the 1705 enriched regions was downloaded from the nucleotide database [25] and compared, with the MultAlin software [26], to all known equine repetitive elements, retrieved from the Repbase database [27, 28]; 97 % (1653/1705) of these repetitive fragments consisted of the 37cen satellite (SAT_EC at [28]). In all these regions the 37cen 221 bp units were organized in a head-to-tail fashion.

We then aligned the reads from input and from immunoprecipitated chromatin with the consensus sequence of 37cen (SAT_EC at [28]), of the pericentromeric satellite 2PI (SAT2pl at [28]) and of the ERE-1 retrotransposon, that is interspersed throughout the genome (ERE1 at [28]); we also aligned them with the sequence of the pericentromeric satellite EC137 (GenBank JX026961, [20]). The alignment was performed using the Razers3 software [29] allowing 20 % of mismatches. The number of reads was normalised to take into account the total number of reads in each sample and the length of the consensus sequence; raw read counts are reported in Additional file 2: Table S2. To quantify the enrichment of these sequences in CENP-A bound chromatin, we calculated the ratio between normalized read counts in the immunoprecipitated and in the input DNA (Fig. 1a, left panel). A 6.5-fold enrichment was observed for the 37cen satellite; 2PI and EC137 were under-represented in the immunoprecipitated chromatin, while ERE1 was equally represented in the two fractions. These results demonstrate that 37cen is the main functional centromeric satellite sequence.

To better define the sequence actually bound by CENP-A, we deduced a consensus from the 33,902,776 reads mapping on the 37cen reference (Additional file 2: Table S2). The consensus is shown as logo in Fig. 1a right panel. Although 20 % of mismatches were allowed in selecting the 37cen reads, the newly defined consensus is very similar to the previously reported consensus suggesting that 37cen units are highly conserved both in CENP-A bound and unbound DNA.

AT richness has been considered a typical feature of centromeric chromatin [30], however, this idea has been recently a subject of debate [8]. The GC content of 37cen is 53 % thus confirming that GC richness is compatible with the centromeric function.

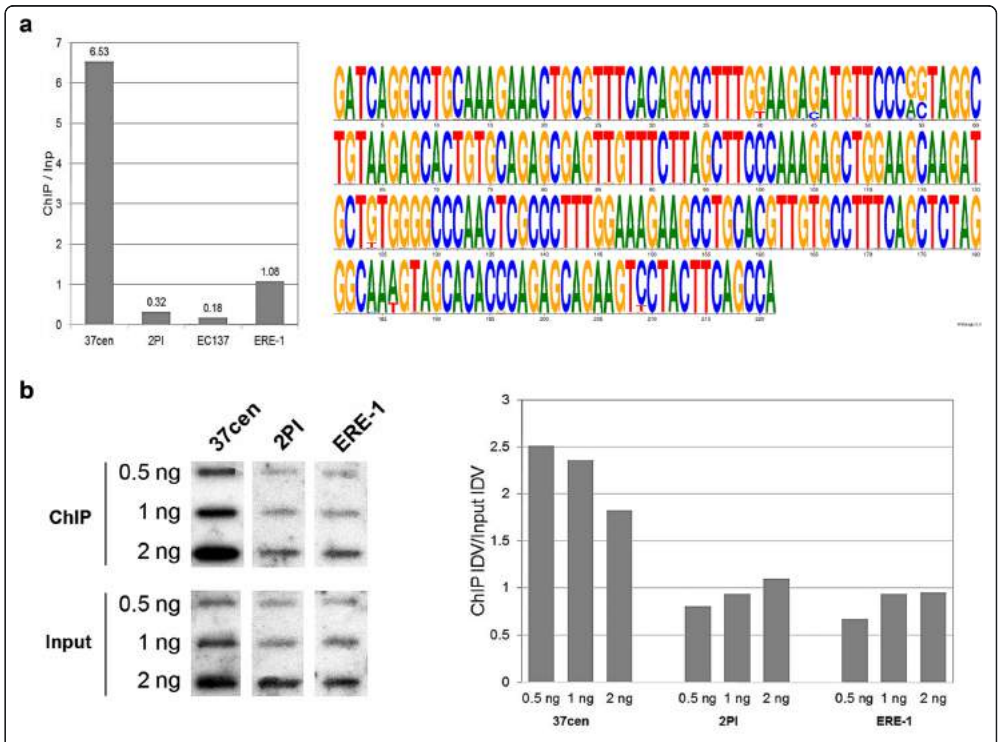


Fig. 1 Identification and analysis of the CENP-A bound horse satellite. **a**: In the left panel, the enrichment of the 37cen, 2P1 and EC37 satellites was measured as ratio between normalized read counts in immunoprecipitated and in input DNA. The same calculation was performed for the ERE1 retrotransposon sequence. The right panel shows the 221 bp consensus sequence of the CENP-A bound 37cen satellite. **b**: Slot-blot analysis. Left panel: hybridization of P^{32} labelled probes (37cen, 2PI and ERE-1) with DNA purified from chromatin immunoprecipitated with CENP-a (*top*) and from non immunoprecipitated chromatin (*bottom*). Right panel: densitometric analysis of slot-blot hybridizations

To further confirm the association of the 37cen satellite DNA with centromeric function, horse chromatin was immunoprecipitated with the anti-CENP-A antibody [9, 21]. Purified immunoprecipitated and input DNA was blotted and hybridized with probes for 37cen, 2PI and ERE-1 repeats (Fig. 1b). The results showed that the 37cen hybridization signal was more intense in immunoprecipitated than in input DNA; conversely, the signal intensity obtained after hybridization with the 2PI and ERE-1 probes was comparable or even lower in immunoprecipitated than in input DNA blots. The Integrated Densitometric Value (IDV) of signals was calculated with the ImageJ 1.48v software [31]. As reported in Fig. 1b, right panel, the ratio between immunoprecipitated and input values for 37cen was comprised between 1.8 and 2.5 confirming that this satellite is enriched in CENP-A bound chromatin. On the opposite, no enrichment of 2PI and ERE-1 repeats was observed.

These results demonstrate that, although at horse centromeric and pericentromeric regions the different satellite families form a complex mosaic of intermingled segments [20], only the 37cen family is involved in the centromeric function. This situation is similar to that previously described in other species, such as humans, where alpha satellite only is bound by CENP-A whereas other satellite families seems to play an accessory function [6].

Transcription of the 37cen satellite

A large body of evidence demonstrates that centromeric and pericentromeric satellite DNA is transcribed in a number of species from yeast to mammals [18]. We analysed, by means of RNA-seq, the transcriptome profile of a horse fibroblast cell line in order to search for 37cen transcripts. Out of the 59,090,294 RNA-seq reads analysed, we detected 9803 reads corresponding to the consensus sequences of 37cen (Fig. 2). The alignment with a

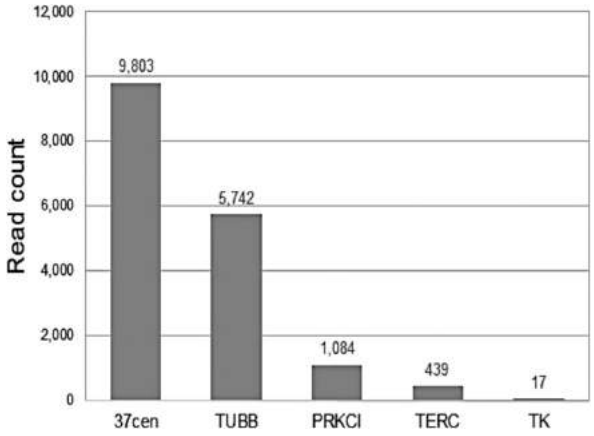


Fig. 2 Transcription of the 37cen satellite by RNA-seq. The graph reports the number of reads corresponding to the consensus sequences of 37cen, tubulin beta (TUBB), protein kinase C iota (PRKCI), telomerase RNA component (TERC), thymidine kinase (TK)

37cen dimer was performed using the Razers3 software [29] and allowing 20 % of mismatches. We also counted the number of reads corresponding to 442 nt long transcripts from four genes: *TUBB (tubulin beta), PRKCI (protein kinase C iota), TERC (telomerase RNA component), TK (thymidine kinase)* (Fig. 2). The results show that the number of 37cen reads is comparable or higher than that observed for the analysed genes.

From these data we cannot infer the transcription level of single 37cen units nor the fraction of transcriptionally active units. It has been suggested that centromeric transcripts may have an impact on development, cell differentiation, and response to environmental stimuli [4, 6] and it is generally agreed that transcription competence is a prerequisite for centromere functioning and kinetochore assembly [32–34]. Emerging evidence suggests that satellite transcripts may act both *in cis* and *in trans* [5, 35]. Therefore, in the horse system, it is tempting to speculate that 37cen RNA may play a role not only at satellite-based centromeres but also at the satellite-less centromere of chromosome 11.

High resolution cytogenetic analysis

Our previous FISH analyses, on stretched chromosomes and combed DNA fibres, demonstrated that horse centromeric and pericentromeric regions display a mosaic arrangement of different satellite DNA families [20]. To analyse the physical organization of the centromeric domains, we carried out immuno-FISH experiments on mechanically stretched chromosomes using 37cen as FISH probe (red in Fig. 3a) and a previously tested [21] CREST serum (green in Fig. 3a) to mark the centromeric domain. A total number of 99 stretched chromosomes (46 meta- or submeta-centric and 53 acrocentric) was examined, a representative panel of which is shown in Fig. 3a. Although the results of this type of experiments can only be considered semi-quantitative, the abundance of the 37cen sequence appeared highly variable among chromosomes, extending in some instances over a large pericentromeric region (white arrows in Fig. 3a) or being apparently confined to the primary constriction. As expected, the CREST signals always colocalized with the 37cen fluorescence, however, no clear correlation seemed to exist between intensity and extension of the 37cen and the CREST signals.

These results suggest that, at horse centromeres, the size of CENP-A binding domains is not related to the extent of satellite DNA stretches; these finding are in agreement with the well described inter- and intra-specific variability of the molecular organization of eukaryotic centromeres [6].

To define more precisely the relationship between 37cen and the centromeric function, a higher-resolution immuno-FISH analysis was performed on horse chromatin fibres. A total number of 25 extended fibres was analysed, some representative examples of which are reported in Fig. 3b. Different arrangements of CENP-A domains were observed: although 60 % of the fibres (15/25) showed CENP-A binding covering the whole length of the 37cen positive region (I in Fig. 3b), in 28 % (7/25) of the cases (II in Fig. 3b) CENP-A domains appeared as blocks of variable length intermingled into 37cen stretches. The observation of the discontinuous presence of CENP-A at centromeres

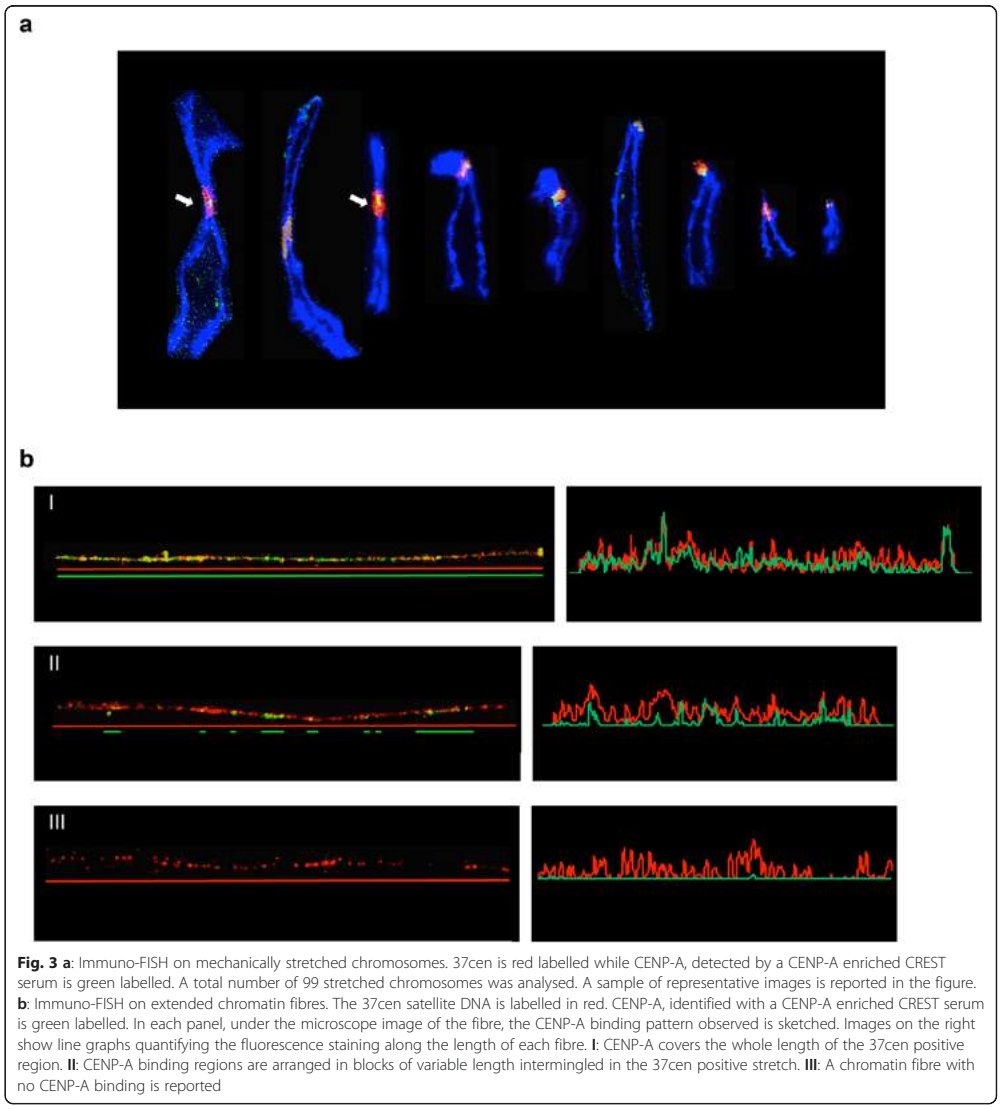


Fig. 3 a: Immuno-FISH on mechanically stretched chromosomes. 37cen is red labelled while CENP-A, detected by a CENP-A enriched CREST serum is green labelled. A total number of 99 stretched chromosomes was analysed. A sample of representative images is reported in the figure. **b**: Immuno-FISH on extended chromatin fibres. The 37cen satellite DNA is labelled in red. CENP-A, identified with a CENP-A enriched CREST serum is green labelled. In each panel, under the microscope image of the fibre, the CENP-A binding pattern observed is sketched. Images on the right show line graphs quantifying the fluorescence staining along the length of each fibre. **I**: CENP-A covers the whole length of the 37cen positive region. **II**: CENP-A binding regions are arranged in blocks of variable length intermingled in the 37cen positive stretch. **III**: A chromatin fibre with no CENP-A binding is reported

resembles the chromatin organization observed using the same high resolution morphological approach in human cells and in *Drosophila* [36]. Our ChIP results (see Fig. 1a) demonstrated that only a fraction of all genomic 37cen repeats is associated with centromere function; the detection of the FISH signal without CENP-A binding (III in Fig. 3b) on 12 % (3/25) of the fibres further confirmed this result; this fraction of fibres may derive from pericentromeric locations, that were shown to contain the 37cen satellite by our analysis on stretched chromosomes (Fig. 3a).

Conclusions

The primary constriction of mammalian chromosomes is typically embedded in a constitutive heterochromatic environment characterized by long arrays of tandemly

repeated satellite DNA. Centromeric satellite repeats are extremely variable in length and composition, not only between and within species but also among chromosomes of the same individual [7]. The horse is peculiar among mammalian species because the centromere of chromosome 11 is completely devoid of satellite DNA [9, 10, 21]. Satellite-based horse centromeres are constituted by the two major classes of equid satellite DNA, 37cen and 2PI, flanked by the pericentromeric accessory satellite EC137 [20]. In the present paper, we proved that only the GC rich 37cen sequence is associated with the centromeric function and is transcriptionally active. We also showed that the horse shares with other species a similar molecular organization of centromeres, relying on CENP-A blocks of variable length immersed in long satellite DNA stretches [36].

The significance of satellite DNA at mammalian centromeres has so far been elusive because satellite-less centromeres are perfectly functional [9, 21]. In the horse, the presence of satellite-based together with a satellite-less centromere makes this species a particularly suitable model for future studies on the role of centromeric tandem repeats.

Methods

Ethics statement

Horse DNA, RNA, chromosomes and chromatin samples were obtained from previously established primary fibroblast cell lines [21]. These cell lines were established from skin samples taken from animals not specifically sacrificed for this study; the animals were being processed as part of the normal work of the abattoirs.

Cell lines

Horse skin primary fibroblasts were were cultured in DMEM medium (EuroClone) supplemented with 20 % foetal bovine serum, 2 mM L-glutamine, 1 % penicillin/streptomycin and 2 % non-essential amino acids at 37 °C with 5 % CO_2. Cytogenetic analysis demonstrated that the cell lines had a diploid modal chromosome number (2n = 64) and a normal karyotype.

Chromatin Immuno-Precipitation (ChIP) and sequencing (ChIP-seq)

Chromatin was prepared from horse primary fibroblasts, following cross-linking with 1 % formaldehyde and sonication. Immunoprecipitation was performed using a purified CENP-A polyclonal [9, 21], raised against the N-terminus of human CENP-A, kindly provided by Prof. Mariano Rocchi (University of Bari). The immunocomplex was purified using A/G beads (nProtein A Sepharose™ 4 Fast Flow/Protein G Sepharose™ 4 Fast Flow, GE Healthcare). After reverse cross-linking, carried out overnight at 65 °C, immunoprecipitated and input DNAs were extracted with the "Wizard Genomic DNA Purification Kit" (Promega) according to the manufacturer's instructions.

Immunoprecipitated and input DNAs were then paired-end sequenced through an Illumina HiSeq2000 platform by IGA Technology Services [37]. Sequence reads were aligned to the horse reference genome (EquCab2.0, 2007 release) with Bowtie 2.0 [22] and peak-calling was performed through the software MACS version 2.0.10 20120605 [23], using default parameters. Stringent criteria [24] were applied to identify significantly enriched regions: fold enrichment > 5, pile-up > 100, $-\log_{10}$(p-value) > 100, $-\log_{10}$(q-value) > 100.

To quantify the number of reads corresponding to each repetitive element, the reads from immunoprecipitated DNA and input DNA were mapped to a reference constituted by the consensus sequences of 37 cen ("SAT_EC" on repbase, [27, 28]), 2PI ("SAT2pl" on repbase), ERE-1 ("ERE1" on repbase) and EC137 (GenBank JX026961). The alignment was performed with the Razers3 software [29] using all of the reads from the paired-end sequencing as a whole single-end dataset; the mapping was carried out using default parameters with exception of percent identity threshold (-i option) which was set to 80. For each sequence type analysed, read counts from immunoprecipitated and input DNA were calculated with the "SAM/BAM to Counts 1.0.0" tool, available on the Galaxy platform [38]. Each read count value was normalized with respect to the total number of reads and to the length of the reference sequence. To measure enrichment due to immunoprecipitation with CENP-A, the ratio between normalized read counts in the immunoprecipitated and input samples was calculated.

Slot-blot analysis

DNA purified from chromatin imunoprecipitated with the anti CENP-A antibody [9, 21] and input DNA were transferred to nylon membranes (Amersham HybondTM-N, GE Healthcare) through a Minifold II apparatus (Schleicher and Schuell) and denatured. The membranes were hybridized at 64 °C for 18 h in Church buffer containing one of the following ^{32}P-α[dCTP]-labelled probes, generated by random primer labelling: a 7 kb EcoRI/SacI 37cen fragment and a 7.2 kb EcoRI/SacI 2PI fragment [10]; a 441 bp PCR-amplified fragment from horse genomic DNA, containing an ERE-1 insertion [39].

After hybridization, the membranes were washed twice in 2× SSC, 0.5 % SDS for 15 min at 64 °C and once in 0.2× SSC, 0.5 % SDS for 30 min at 64 °C. Radioactive signals were detected using a phosphorimager (Cyclone, Packard) and the densitometric analysis was performed with the ImageJ 1.48v software [31].

RNA extraction and sequencing (RNA-seq)

RNA extraction from whole cells was performed using QIAzol Lysis Reagent (QIAGEN) according to the manufacturer's instructions. To eliminate DNA contaminations, RNA was treated twice with RNase-free DNase-I (Promega), and then purified with the RNA Clean and Concentration kit (ZYMO Research). After library preparation using Illumina TruSeq Stranded Total RNA with Ribo-Zero GOLD, the resulting cDNA was paired-end sequenced by IGA Technology Services [37] through an Illumina HiSeq2000 platform.

RNA-seq reads were mapped, with the same Razers3 parameters as the ChIP and input datasets, on a reference composed of a dimer of the 37cen consensus sequence ("SAT_EC" on repbase) and on 442 bp long portions of the following transcripts: TUBB (XM_001491178.5, nucleotides 488 to 929), PRKCI (XM_014732748.1, nucleotides 605 to 1046), TERC (NR_001566.1 nucleotides 9 to 450), TK (XM_001491081.5 nucleotides 26 to 467). The same length was used for each sequence in order to have comparable read counts without normalization.

Immuno-FISH

Mechanically stretched chromosomes and extended chromatin fibres were prepared as previously described [20, 21]. Immunofluorescence was carried out using a CENP-A enriched CREST serum [21] for CENP-A detection, and a plasmid containing the 37cen satellite as FISH probe [20]; immuno-FISH experiments on stretched chromosomes and chromatin fibres were carried out as previously described [21]. Digital grey-scale images were acquired with a fluorescence microscope (Zeiss Axioplan) equipped with a cooled CCD camera (Photometrics). Pseudocoloring and merging of images were performed using the IpLab software (Scanalytics Inc.). For fluorescence quantification of 37cen (red signal) and CENP-A (green signal) on chromatin fibres, separate channel digital images were converted in text images using ImageJ 1.48v [31]. The mean fluorescence intensity of each antibody spot was calculated point by point along the fibre length and plotted in a line chart.

Additional files

Additional file 1: Table S1. Enriched regions found on the unplaced contigs. The columns represent: the accession number of the contigs, the start and end position of the enriched region within the contig, the length of the region, and the statistical parameters calculated by the peak caller [pile-up, fold enrichment, -log10(p-value), -log10(q-value)]. Regions are listed according to their contig number. (XLS 211 kb)

Additional file 2: Table S2. un-normalized read counts from ChIP-seq experiment and input control. (XLS 27 kb)

Abbreviations

CENP-A: centromere protein A; ChIP: chromatin immunoprecipitation; ChIP-seq: ChIP sequencing; CREST: calcinosis, Raynaud's phenomenon, esophageal dysmotility, sclerodactyly, and telangiectasia; ERE-1: equine repetitive element 1; FISH: fluorescence *in situ* hybridization; IDV: integrated densitometric value; MACS: model-based analysis of ChIP-seq; PRKCI: protein kinase C iota; RNA-seq: RNA sequencing; TERC: telomerase RNA component; TK: thymidine kinase; TUBB: tubulin beta.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

FC and RG carried out bioinformatics analysis of ChIP-seq and RNA-seq experiments, contributed to manuscript drafting and figure preparation. AM carried out cytogenetic analyses and contributed to figure preparation. FMP and EC carried out immunoprecipitations and slot-blot experiments. EB contributed to cytogenetic experiments. SGN supervised molecular biology experiments and bioinformatics analyses. ER designed and supervised cytogenetic experiments and participated to manuscript preparation. EG conceived the study, supervised the sequencing experiments and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Giulio Pavesi (Bioinformatics, Evolution and Comparative Genomics Lab Department of Biosciences, University of Milan, Italy) for helpful suggestions on bioinformatics analysis, Mariano Rocchi (Department of Biology, University of Bari, Italy) for providing the anti-CENP-A antibody and Claudia Alpini (Fondazione I.R.C.C.S. Policlinico San Matteo, Pavia, Italy) for the CREST serum. F.C. was supported by a fellowship from Fondazione di Piacenza e Vigevano. This work was supported by grants from Consiglio Nazionale delle Ricerche (CNR-Progetto Bandiera Epigenomica) and from Ministero dell'Istruzione dell'Università e della Ricerca (MIUR-PRIN).
This paper is dedicated to the memory of Federico Cerutti who left us on May 30, 2015. We will never forget his warm smile, kind personality and bright intelligence.

Received: 21 March 2016 Accepted: 1 April 2016
Published online: 27 April 2016

References

1. Szybalski W. Use of cesium sulfate for equilibrium density gradient centrifugation. Methods Enzymol. 1968;12:330–60.
2. Volpe TA, Kidner C, Hall IM, Teng G, Grewal SIS, Martienssen RA. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. Science. 2002;297:1833–7.
3. Vourc'h C, Biamonti G. Transcription of Satellite DNAs in Mammals. Prog Mol Subcell Biol. 2011;51:95–118.
4. Gent JI, Dawe RK. RNA as a structural and regulatory component of the centromere. Annu Rev Genet. 2012;46:443–53.
5. Quénet D, Dalal Y. A long non-coding RNA is required for targeting centromeric protein A to the human centromere. Elife. 2014;3:e03254.
6. Plohl M, Meštrović N, Mravinac B. Centromere identity from the DNA point of view. Chromosoma. 2014;123:313–25.
7. Plohl M, Luchetti A, Mestrović N, Mantovani B. Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. Gene. 2008;409:72–82.
8. Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, et al. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol. 2013;14:R10.
9. Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, et al. Genome sequence, comparative analysis, and population genetics of the domestic horse. Science. 2009;326:865–7.
10. Piras FM, Nergadze SG, Magnani E, Bertoni L, Attolini C, Khoriauli L, et al. Uncoupling of Satellite DNA and Centromeric Function in the Genus Equus. PLoS Genet. 2010;6:e1000845.
11. Shang W-H, Hori T, Martins NMC, Toyoda A, Misu S, Monma N, et al. Chromosome engineering allows the efficient isolation of vertebrate neocentromeres. Dev Cell. 2013;24:635–48.

12. Gong Z, Wu Y, Koblízková A, Torres GA, Wang K, Iovene M, et al. Repeatless and repeat-based centromeres in potato: implications for centromere evolution. Plant Cell. 2012;24:3559–74.
13. Rocchi M, Archidiacono N, Schempp W, Capozzi O, Stanyon R. Centromere repositioning in mammals. Heredity (Edinb). 2012;108:59–67.
14. Steiner FA, Henikoff S. Diversity in the organization of centromeric chromatin. Curr Opin Genet Dev. 2015;31:28–35.
15. Marshall OJ, Chueh AC, Wong LH, Choo KHA. Neocentromeres: new insights into centromere structure, disease development, and karyotype evolution. Am J Hum Genet. 2008;82:261–82.
16. Rošić S, Köhler F, Erhardt S. Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division. J Cell Biol. 2014;207:335–49.
17. Biscotti MA, Canapa A, Forconi M, Olmo E, Barucca M. Transcription of tandemly repetitive DNA: functional roles. Chromosome Res. 2015;23:463–77.
18. Rošić S, Erhardt S. No longer a nuisance: long non-coding RNAs join CENP-A in epigenetic centromere regulation. Cell Mol Life Sci. 2016;73:1387–98.
19. Anglana M, Bertoni L, Giulotto E. Cloning of a polymorphic sequence from the nontranscribed spacer of horse rDNA. Mamm Genome. 1996;7:539–41.
20. Nergadze SG, Belloni E, Piras FM, Khoriauli L, Mazzagatti A, Vella F, et al. Discovery and comparative analysis of a novel satellite, EC137, in horses and other equids. Cytogenet Genome Res. 2014;144:114–23.
21. Purgato S, Belloni E, Piras FM, Zoli M, Badiale C, Cerutti F, et al. Centromere sliding on a mammalian chromosome. Chromosoma. 2015;124:277–87.
22. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
23. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137.
24. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, et al. Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. PLoS Comput Biol. 2013;9:e1003326.
25. Home - Nucleotide - NCBI. http://www.ncbi.nlm.nih.gov/nucleotide. Accessed 11 Mar 2016.
26. Corpet F. Multiple sequence alignment with hierarchical clustering. Nucl Acids Res. 1988;16:10881–90.
27. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110:462–7.
28. Repbase - GIRI. http://www.girinst.org/repbase/index.html. Accessed 11 Mar 2016.
29. Weese D, Holtgrewe M, Reinert K. RazerS 3: Faster, fully sensitive read mapping. Bioinformatics. 2012;28:2592–9.
30. Henikoff S, Ahmad K, Malik HS. The centromere paradox: stable inheritance with rapidly evolving DNA. Science. 2001;293:1098–102.
31. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Meth. 2012;9:671–5.
32. Hall LE, Mitchell SE, O'Neill RJ. Pericentric and centromeric transcription: a perfect balance required. Chromosome Res. 2012;20:535–46.
33. Bergmann JH, Martins NMC, Larionov V, Masumoto H, Earnshaw WC. HACking the centromere chromatin code: insights from human artificial chromosomes. Chromosome Res. 2012;20:505–19.
34. Chan FL, Wong LH. Transcription in the maintenance of centromere chromatin identity. Nucl Acids Res. 2012;40:11178–88.
35. Bergmann JH, Rodríguez MG, Martins NMC, Kimura H, Kelly DA, Masumoto H, et al. Epigenetic engineering shows H3K4me2 is required for HJURP targeting and CENP-A assembly on a synthetic human kinetochore: H3K4me2 and kinetochore maintenance. EMBO J. 2011;30:328–40.
36. Blower MD, Sullivan BA, Karpen GH. Conserved Organization of Centromeric Chromatin in Flies and Humans. Dev Cell. 2002;2:319–30.
37. IGA Technology Services. http://www.igatechnology.com. Accessed 11 Mar 2016.
38. Galaxy. https://usegalaxy.org. Accessed 11 Mar 2016.
39. Santagostino M, Khoriauli L, Gamba R, Bonuglia M, Klipstein O, Piras FM, et al. Genome-wide evolutionary and functional analysis of the Equine Repetitive Element 1: an insertion in the myostatin promoter affects gene expression. BMC Genet. 2015;16:126.

Research

Birth, evolution, and transmission of satellite-free mammalian centromeric domains

Solomon G. Nergadze,[1,6] Francesca M. Piras,[1,6] Riccardo Gamba,[1,6] Marco Corbo,[1,6] Federico Cerutti,[1,†] Joseph G.W. McCarter,[2] Eleonora Cappelletti,[1] Francesco Gozzo,[1] Rebecca M. Harman,[3] Douglas F. Antczak,[3] Donald Miller,[3] Maren Scharfe,[4] Giulio Pavesi,[5] Elena Raimondi,[1] Kevin F. Sullivan,[2] and Elena Giulotto[1]

[1]*Department of Biology and Biotechnology "Lazzaro Spallanzani," University of Pavia, 27100 Pavia, Italy;* [2]*Centre for Chromosome Biology, School of Natural Sciences, National University of Ireland, Galway, H91 TK33, Ireland;* [3]*Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, New York 14850, USA;* [4]*Genomanalytik (GMAK), Helmholtz Centre for Infection Research (HZI), 38124 Braunschweig, Germany;* [5]*Department of Biosciences, University of Milano, 20122 Milano, Italy*

Mammalian centromeres are associated with highly repetitive DNA (satellite DNA), which has so far hindered molecular analysis of this chromatin domain. Centromeres are epigenetically specified, and binding of the CENPA protein is their main determinant. In previous work, we described the first example of a natural satellite-free centromere on *Equus caballus* Chromosome II. Here, we investigated the satellite-free centromeres of *Equus asinus* by using ChIP-seq with anti-CENPA antibodies. We identified an extraordinarily high number of centromeres lacking satellite DNA (16 of 31). All of them lay in LINE- and AT-rich regions. A subset of these centromeres is associated with DNA amplification. The location of CENPA binding domains can vary in different individuals, giving rise to epialleles. The analysis of epiallele transmission in hybrids (three mules and one hinny) showed that centromeric domains are inherited as Mendelian traits, but their position can slide in one generation. Conversely, centromere location is stable during mitotic propagation of cultured cells. Our results demonstrate that the presence of more than half of centromeres void of satellite DNA is compatible with genome stability and species survival. The presence of amplified DNA at some centromeres suggests that these arrays may represent an intermediate stage toward satellite DNA formation during evolution. The fact that CENPA binding domains can move within relatively restricted regions (a few hundred kilobases) suggests that the centromeric function is physically limited by epigenetic boundaries.

[Supplemental material is available for this article.]

Chromosome segregation during mitosis and meiosis is directed by the centromere, the chromosomal locus that specifies kinetochore assembly during cell division (Cleveland et al. 2003; McKinley and Cheeseman 2015). Although the mechanism of kinetochore function in mitosis is highly conserved, centromere-associated DNA sequences are highly variable in evolution, a situation that has been referred to as the centromere paradox (Eichler 1999; Henikoff et al. 2001). In most multicellular organisms, centromeres are associated with large arrays of tandemly iterated satellite DNA sequences, typified by alpha-satellite DNA of primates in which a 171-bp sequence is present in arrays of up to megabase size at the primary constriction of mitotic chromosomes (Hayden et al. 2013). Despite this common theme, the sequences of the centromeric satellite DNA are divergent and are estimated to be among the most rapidly evolving components of the genome (Plohl et al. 2014). Direct evidence that DNA sequence is not the sole factor in determining centromere position or function was originally derived from examination of human chromosomal abnormalities. Dicentric chromosomes possessing kinetochore activity at only one of two alpha-satellite loci revealed that satellite DNA is not sufficient for centromere specification (Earnshaw and Migeon 1985). Identification of analphoid chromosomes, that nonetheless possessed fully functional centromeres, demonstrated that satellite DNA is not necessary for centromere function (Voullaire et al. 1993). Rather than DNA sequence, the common feature that links centromere function in most eukaryotes is the presence of a distinctive histone H3 variant, CENPA, which can directly confer centromere function to a locus when tethered experimentally (Palmer et al. 1991; Stoler et al. 1995; Mendiburo et al. 2011). These observations have led to the proposal that centromere identity is established and maintained through epigenetic mechanisms, and CENPA functions as a central component in centromere specification (Karpen and Allshire 1997; Panchenko and Black 2009; McKinley and Cheeseman 2015).

The evolutionary plasticity of centromeres is exemplified by the phenomenon of centromere repositioning (Montefalcone et al. 1999). By detailed molecular characterization of karyotypic relationships among primate species, it was observed that centromere position can change without a corresponding change in DNA organization (Montefalcone et al. 1999; Cardone et al. 2006; Ventura et al. 2007). In these cases, referred to as

[6]These authors contributed equally to this work.
[†]Deceased.
Corresponding authors: elena.giulotto@unipv.it, kevin.sullivan@nuigalway.ie, elena.raimondi@unipv.it
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.231159.117.

evolutionarily new centromeres (ENCs), centromere evolution seems to be driven by forces other than the surrounding DNA.

A relationship between ENCs and the analphoid neocentromeres observed in human clinical samples emerged from analysis of the positions in which these events occur. For example, human neocentromeres at Chromosomes 3, 9, and 6 occur in the same genomic regions as ENCs observed in some primates, indicating that certain regions of the genome have a propensity to form centromeres (Ventura et al. 2004; Capozzi et al. 2008, 2009). Thus, regions of the genome may harbor "latent" centromere potential (Voullaire et al. 1993). The observation that the primate ENCs possessed typical arrays of alpha-satellite DNA led to the hypothesis that epigenetic marks can drive the movement of centromere function to new genomic sites, which can subsequently mature through the acquisition of satellite DNA sequences (Amor and Choo 2002; Piras et al. 2010; Kalitsis and Choo 2012). Following their original discovery in primates, a surprisingly large number of ENCs were identified in the genus *Equus* (Carbone et al. 2006; Piras et al. 2009), and some examples were also observed in other animals (Ferreri et al. 2005; Kobayashi et al. 2008) and in plants (Han et al. 2009), indicating that centromere repositioning is a widespread force for karyotype evolution.

A fundamental step in understanding centromere biology was the discovery that the ENC at horse Chromosome 11 is completely devoid of satellite DNA (Wade et al. 2009). This observation revealed, for the first time, that a satellite-free centromere can be present in all individuals of a vertebrate species as a normal karyotype component. This centromere is established on a segment of DNA, conserved in vertebrates, which is free of genes as well as of satellite DNA, providing an example of an evolutionarily "young" ENC that has not acquired repetitive sequences. Satellite-free centromeres were subsequently observed in chicken (Shang et al. 2010), orangutan (Locke et al. 2011), and potato (Gong et al. 2012).

Examination of the centromere of horse Chromosome 11 in several individuals revealed that the satellite-free centromeric domains are present in each case, but the precise location of the CENPA binding region (~100 kb in length) differs among individuals and even between the two homologous chromosomes of a single individual (Purgato et al. 2015). Centromere activity could be associated with any sequence within a ~500-kb domain in the centromere forming region of Chromosome 11. Therefore, this "centromere sliding" is DNA sequence independent, as expected for an epigenetically defined locus. Thus, centromeres exhibit large-scale relocalization (centromere repositioning) during evolution as well as short-range relocalization (centromere sliding) within a population (Giulotto et al. 2017).

The genus *Equus* comprises eight extant species (two horses, three donkeys, and three zebras) that diverged from a common ancestor ~4 million years ago (Mya) (Steiner et al. 2012; Orlando et al. 2013). In a previous work, we analyzed the karyotype of four *Equus* species by in situ hybridization with satellite DNA probes and revealed that, in the domestic donkey (*E. asinus*) and in two zebras (*E. burchelli* and *E. grevyi*), a large number of centromeres lack detectable satellite DNA (Piras et al. 2010; Geigl et al. 2016), whereas in the horse, Chromosome 11 is the only one.

The aim of this work was to verify the presence of satellite-free centromeres in *E. asinus*, using ChIP-seq with anti-CENPA antibodies, to analyze their DNA sequence organization, positional stability, and transmission.

Results

Satellite-free CENPA binding domains in *Equus asinus*

Our previous work identified several donkey centromeres that lack detectable satellite repeats (Piras et al. 2010). Here, to identify the DNA sequences at these centromeres, ChIP-seq experiments were carried out on donkey primary skin fibroblasts. Two different antibodies were used to immunoprecipitate formaldehyde cross-linked chromatin fragments: a rabbit antiserum against CENPA (Wade et al. 2009) and a human CREST serum with high titer against CENPA (Purgato et al. 2015; Cerutti et al. 2016). DNA purified from immunoprecipitated and input chromatin was then subjected to paired-end Illumina sequencing. Since we previously demonstrated the presence of a satellite-free centromere on horse Chromosome 11 by ChIP-on-chip (Wade et al. 2009; Purgato et al. 2015), as positive control, we carried out the same ChIP-seq experiment with chromatin from horse skin fibroblasts. The horse and donkey genomes share an average of >98% sequence identity (Orlando et al. 2013; Huang et al. 2015) and chromosome orthologies are well described (Yang et al. 2004; Musilova et al. 2013). Since only draft sequences of the donkey genome comprising unassembled scaffolds are available (Orlando et al. 2013; Huang et al. 2015), we aligned both the horse and the donkey reads to the horse reference genome (EquCab2.0). Sequencing and alignment statistics of the ChIP-seq experiments are reported in Supplemental Table S1. Figure 1 reports the graphical representation of the enrichment peaks, corresponding to the centromere of horse Chromosome 11 from one individual, here called HorseS (Fig. 1A), and to the 16 donkey satellite-free centromeric domains from one individual, here called DonkeyA (Fig. 1B). The two antibodies recognized essentially identical sequence domains and exhibited largely similar patterns of protein binding.

The 16 donkey regions spanned 54–345 kb and contained one or two CENPA binding domains. Similar to what we described for horse Chromosome 11 (Purgato et al. 2015), the presence of two peaks is related to different epialleles on the two homologs, as demonstrated below on the basis of single nucleotide variant (SNV) analysis. Although some peaks showed a Gaussian-like regular shape (such as EAS4 and EAS30), other peaks were irregular (such as EAS8 and EAS14), contained gaps (such as EAS7 and EAS14), or exhibited a narrow, spike-like distribution (such as EAS9 and EAS19).

The satellite-based donkey centromeres are not described here because their corresponding ChIP-seq reads cannot be precisely mapped on specific chromosomes in the horse reference genome. These centromeres are probably organized similarly to the great majority of typical mammalian centromeres, as already shown for satellite-based horse centromeres (Nergadze et al. 2014; Cerutti et al. 2016).

CENPA binding domains correspond to primary constrictions in 16 *E. asinus* chromosomes

Cytogenetic analysis was carried out to map the 16 donkey CENPA binding regions relative to the primary constrictions of horse and donkey chromosomes. CENPA binding domain coordinates were used to select a set of horse BACs from the CHORI-241 library (Supplemental Table S2; Leeb et al. 2006). These were used as probes for in situ hybridization on metaphase spreads of horse and donkey skin fibroblasts. Examples of in situ hybridization

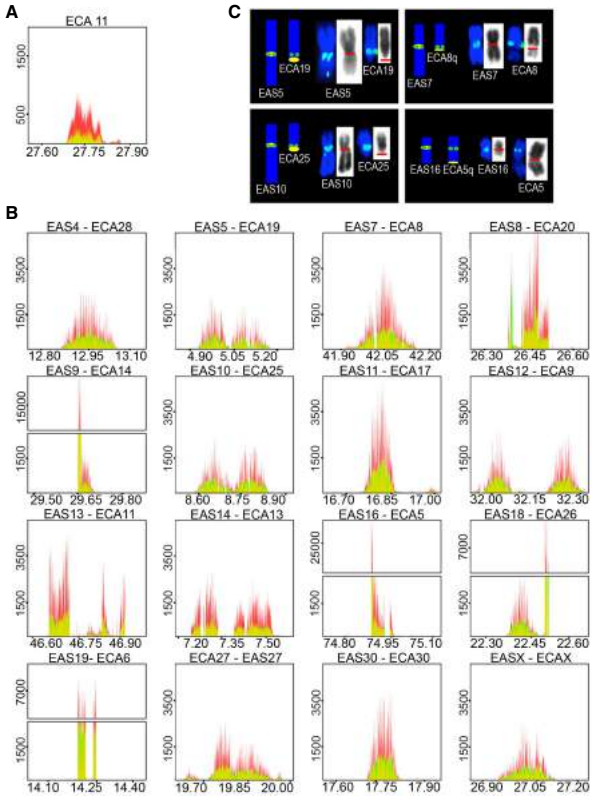


Figure 1. Identification of satellite-free centromeres in *Equus asinus*. ChIP-seq reads from primary fibroblasts of HorseS (*A*) and DonkeyA (*B*) were mapped on the EquCab2.0 horse reference genome. Immunoprecipitation was performed with an antibody against human CENPA (red) or with a CREST serum (green). Peak overlapping appears in yellow. The *y*-axis reports the normalized read counts, whereas the *x*-axis reports the genomic coordinates (Mb). The *E. caballus* satellite-free centromere from Chromosome 11 (*A*) and the 16 satellite-free *E. asinus* centromeres (*B*) are shown; for each *E. asinus* (EAS) chromosome, the number of the orthologous *E. caballus* chromosome (ECA) is reported. (*C*) FISH with BAC probes covering the genomic regions identified by ChIP-seq. Four examples (EAS) along with their orthologous horse chromosomes (ECA) are shown; the remaining chromosomes are reported in Supplemental Figure S1. On the *left* of each panel, a sketch of the orthology between *E. caballus* and *E. asinus* chromosomes (Yang et al. 2004; Musilova et al. 2013) is shown, with BAC signals represented as green dots, and the position of the cytogenetically determined primary constriction represented as a yellow oval. On the *right* of each panel, metaphase chromosomes are shown with FISH signals in green, and the primary constriction is marked by a red line on the reverse DAPI images (gray).

results are shown in Figure 1C with remaining data presented in Supplemental Figure S1. Each of the BAC probes identified a unique locus on the donkey karyotype, and its location was always consistent with the location of the primary constriction. Notably, the FISH signal on the orthologous horse chromosome was never centromeric, suggesting that the 16 satellite-free donkey centromeres were repositioned during evolution. We conclude that the 16 CENPA binding domains identified by ChIP-seq analysis are ENCs located within the respective cytogenetically defined primary constrictions.

Sequence assembly of satellite-free centromere domains and comparison with orthologous horse genomic regions

Several CENPA binding domains showed read-free gaps and distorted shapes when mapped to the horse reference genome, suggesting differences in DNA sequence between the two species (Fig. 1B). The actual DNA sequence corresponding to the donkey centromeres was determined by assembling Illumina reads and carrying out Sanger sequencing of selected regions to resolve gaps in the assembly. For each centromeric region, genomic segments ranging in size between 157 and 358 kb were assembled (Supplemental Table S3).

In the majority of donkey satellite-free centromeres, multiple rearrangements (deletions, insertions, and inversions) were observed compared to the horse orthologous sequence (EAS4, EAS5, EAS7, EAS10, EAS11, EAS12, EAS13, EAS14, EAS27, EAS30) (Supplemental Fig. S2). The number and size of these rearrangements varies at different centromeres, but deletions are the most prevalent type. In donkey Chromosome 5, we observed several deletions; given the small size of these deletions, no gaps in the peak profile were observed. Conversely, donkey Chromosome 7 contains three relatively large deletions coinciding with gaps in the peak profile. The organization of the centromere of donkey Chromosome 13 is more complex, including a large deletion (110 kb) and a translocation, giving rise to a large gap in the central region (deletion) and an off-site peak outside the right border (translocation). In EAS14, which shows a two-peak profile, four relatively extended deletions coincide with gaps in the peak profile. No rearrangements were evident in the centromere of donkey Chromosome X. The centromeric domain identified by ChIP-seq is contained within the previously described large pericentric inversion of donkey Chromosome X (Raudsepp et al. 2002).

To determine more precisely the organization of CENPA distribution at satellite-free centromeres, we constructed a chimeric reference genome by inserting the assembled centromeric donkey contigs in EquCab2.0 to replace their orthologous horse sequences (Supplemental Table S3). The result was a virtual reference genome named EquCabAsiA.

ChIP-seq reads were then mapped on the EquCabAsiA genome (Supplemental Fig. S3). Comparison of the peak profiles obtained with the two reference genomes (Fig. 1B; Supplemental Fig. S3) shows that large gaps and irregular profiles that were observed in Figure 1B (EAS7, EAS13, EAS14, EAS16, EAS19) were no longer

detected following the new alignment. These results demonstrate that the CENPA binding domains of the satellite-free donkey centromeres are uninterrupted, and their architectural organization resembles that of horse Chromosome 11 (Fig. 1A; Wade et al. 2009).

Tandem repetitions associated with some satellite-free centromeres

For five donkey centromeres (EAS8, EAS9, EAS16, EAS18, and EAS19), we detected novel tandem repetitions of sequences that are single copy in the horse genome. In particular, reads spanning junctions between adjacent units of tandem arrays directly demonstrated their presence. For EAS18 and EAS19, the amplified sequences contain a deletion relative to the horse genomic sequence (Supplemental Fig. S2). Due to their repetitive nature, these five regions could not be precisely assembled. To prove the presence of tandem repetitions at these centromeres and to determine their copy number, three independent approaches were taken (Fig. 2). Sequence amplification was initially tested by comparative Southern blotting (Fig. 2A). Four individuals were analyzed: one horse (HorseS), two donkeys (DonkeyA and DonkeyB), and a mule (MuleA), offspring of DonkeyB. Signal intensity of the bands clearly indicated increased copy number of these sequences in the donkeys compared to the horse. The copy number increase is particularly marked for EAS9 and EAS18. As expected, in the mule, signal intensity was intermediate between the donkey parent and the horse sample. At the EAS19 centromeric domain, signal intensity was different in the two donkey samples, suggesting polymorphism in the population.

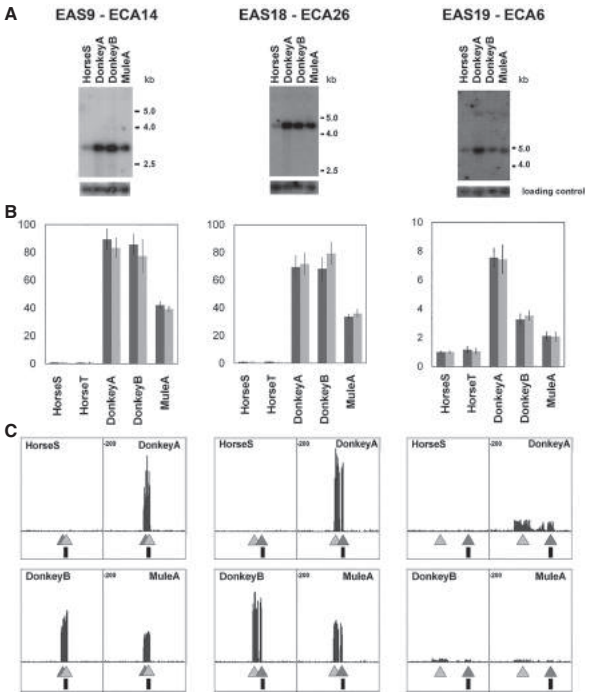


Figure 2. DNA sequence amplification at the centromeres of *E. asinus* Chromosomes 9, 18, and 19. The number of the *E. asinus* chromosome (EAS) and of its ortholog in *E. caballus* (ECA) is reported on *top*. (*A*) Southern blot analysis of genomic DNA from one horse, two donkeys, and a mule (MuleA, offspring of DonkeyB). The probes were obtained by PCR-amplification of a portion of the unit repeated in the donkey (Supplemental Table S4). Map positions of the probes are indicated as vertical black rectangles in *C*. (*B*) Quantitative PCR performed on DNA from two horses, two donkeys, and one mule. Each centromere was analyzed with two primer pairs (dark and light gray bars) (Supplemental Table S4). (*C*) Profile of input reads from one horse, two donkeys, and one mule aligned on the horse reference genome. The genomic regions shown are 29,593,109–29,725,206 for Chromosome 9; 22,441,448–22,572,314 for Chromosome 18; and 14,157,787–14,289,525 for Chromosome 19. Peaks represent regions amplified in the donkey genome compared to the horse genome. Light and dark gray triangles indicate the location of the fragments amplified in the quantitative PCR assay (*B*).

To quantify copy number variation, quantitative PCR (qPCR) experiments were performed, including a second horse individual (HorseT) (Fig. 2B). The results confirm sequence amplification in the two donkeys, particularly marked at the EAS9 and EAS18 centromeres (about 70- to 90-fold compared to the horses); in the mule, the copy number corresponds to about half the value of its DonkeyB father. At EAS19, the number of repeats is relatively low and differs in the two donkeys; in the mule, fold enrichment values are between those of the horses and the donkey father.

A third independent method directly compared read counts between horse and donkey input samples, aligned to the horse reference genome EquCab2.0 (Fig. 2C). The presence of peaks in the donkey centromere domains and their absence in the horse confirm that these regions are amplified in the donkey. Peak height is greater in the donkeys with respect to the mule, and the degree of amplification is lower in EAS19 compared to the other two chromosomes. Quantitative PCR experiments and input read count comparisons were also carried out to analyze the variation of copy number at the centromeres of EAS16 and EAS8 (Supplemental Fig. S4), revealing sequence amplification and copy number variation.

Taken together, these results confirm the occurrence of tandem sequence amplification at a subset of centromeres in the donkey, with evidence for marked inter-individual variation in copy number at some of these loci.

DNA sequence analysis of the satellite-free centromeric domains

DNA sequence features of the satellite-free donkey centromeres were compared with the corresponding regions in the horse genome (Supplemental Fig. S5). The five centromeres containing amplifications were excluded from this analysis because we could not define their complete sequence. The percentage of SINEs, LINEs, LTR-derived sequences, and transposable DNA elements

at the donkey centromere domains did not differ from the orthologous horse sequences. The GC content at these loci was also similar in the two species. Since the horse genome sequence is not well annotated and no annotation of the donkey genome is available, we are not able to provide an accurate analysis of gene content in the satellite-free centromeric regions.

We then compared the abundance of transposable elements at the centromeric regions with the average genome-wide values obtained from a draft donkey genome (Huang et al. 2015). Donkey centromeres were significantly poor in SINEs ($P < 0.00001$), whereas LINE elements were enriched ($P = 0.0057$); LTRs and DNA elements showed the same abundance in all samples. As expected, centromeric satellite sequences (Piras et al. 2010; Cerutti et al. 2016) were totally absent from the 16 centromeres examined here. Finally, donkey centromeres showed a 36.2% GC content as opposed to the genome-wide average of 41.3%, indicating that these satellite-free centromeres are AT rich.

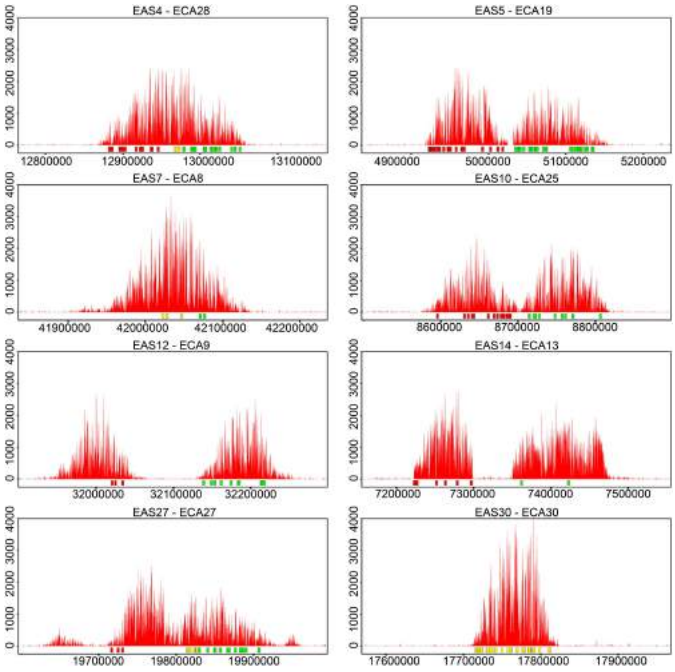


Figure 3. Identification of epialleles through SNV analysis. The positions of single nucleotide variants (SNVs), located within each centromeric domain, are represented as colored rectangles under each ChIP-seq profile. Reads were mapped on the chimeric EquCabAsiA reference genome. The *y*-axis reports the normalized read counts, and the *x*-axis reports the genomic coordinates. Red or green rectangles indicate positions where only one nucleotide variant was enriched in the immunoprecipitated reads, and yellow rectangles indicate positions where both SNVs were present.

Centromere sliding occurs in *Equus asinus*

The double peaks observed on several chromosomes (EAS5, EAS10, EAS12, EAS14, and EAS18) suggested the presence of epialleles on the homologous pairs in the donkey similarly to what we reported for horse Chromosome 11 (Purgato et al. 2015). To verify the presence of epialleles, we used a single nucleotide variant (SNV) based approach. We identified heterozygous nucleotide positions, SNVs, within each centromeric domain using a high coverage input library (Supplemental Table S1). These heterozygous positions would allow us to resolve the two homologs in the reads obtained from CENPA immunoprecipitated chromatin: If the two CENPA domains were present on both homologs, immunoprecipitated chromatin would contain similar amounts of the two SNV alleles; alternatively, if each homolog contained a single CENPA domain, only one of the two SNV alleles would be enriched in immunoprecipitated chromatin. The results of this analysis are shown in Figure 3 and Supplemental Table S5. The SNV analysis was informative for eight of the 16 centromeres (EAS4, 5, 7, 10, 12, 14, 27, and 30). The X Chromosome was excluded because this animal is a male; the five chromosomes with tandem repetitions at centromeres were excluded due to incomplete sequence definition; finally, at EAS11 and EAS13, centromeres informative SNVs were not identified. On EAS5, 10, 12, and 14 centromeres with two clearly separated peaks, a single variant was highly enriched at all positions in the immunoprecipitated DNA, demonstrating that each homolog contains a single functional domain in different positions on the two homologs (Fig. 3). On EAS4, 7, and 27, different results were obtained when SNVs at the edges or at the center of the peak were analyzed. At the edges, only one variant was observed; on the contrary, both nucleotides were found at the center of the peak; the interpretation of this result is that CENPA binds to slightly different but overlapping regions in the two homologs. On EAS30, at all positions both single nucleotide variants were detected, suggesting that the two homologs contain a very similar epiallele, giving rise to overlapping CENPA binding domains.

The size of individual epialleles was estimated by taking into account the borders of each peak and the distribution of SNVs (Fig. 3). This measurement is not precise, particularly when two epialleles overlap (EAS4, EAS7, and EAS27), giving rise to an approximate size of 100 kb.

To further investigate the individual variability of the donkey satellite-free centromeric domains, we analyzed an additional unrelated donkey (DonkeyB) by ChIP-seq with the same anti-CENPA antibody used for DonkeyA (Supplemental Fig. S6). To compare the two individuals, the reads of both animals were mapped on the horse reference sequence (EquCab2.0). Of the 16 satellite-free centromeres identified in DonkeyA, only 15 proved to be satellite-free in the DonkeyB: No enrichment of the ChIP-seq reads was observed on EAS8. It may be that, in DonkeyB, the centromere occurs on satellite repeats. A situation like this was recently

described in orangutan (Tolomeo et al. 2017), and we may be seeing a polymorphism in the donkey population at Chromosome 8.

A marked variability in the position of CENPA binding domains between the two individuals was observed at six chromosomes (Supplemental Figure S6), indicating that CENPA binding domains can move within regions of up to 600 kb. The remaining nine satellite-free centromeres showed little or no positional variability between these two animals.

Germline and somatic transmission of centromeric domains

The observation of positional instability of satellite-free centromeres raises the question of when such movement of the CENPA domain can occur. The stability of centromeres across generations was examined by crossing DonkeyB with three mares (HorseA, HorseB, and HorseC) by in vitro fertilization. Embryonic fibroblasts were established from the resultant mule concepti (MuleA, MuleB, and MuleC). Adult skin fibroblast cell lines were established from DonkeyB and from two of the three mares (HorseA and HorseC; cells from HorseB were not available). In addition, skin fibroblasts cell lines were obtained from a male horse (HorseD) and from the hinny derived from its cross with a female donkey (female donkey cells not available). The genetic relationships among the individuals used in this study are reported in Figure 4A. All the cell lines from the two families were subjected to ChIP-seq analysis using anti-CENPA antibody. Since the mule and hinny cells contain two haploid genomes, one from *E. caballus* and one from *E. asinus*, the transmission of individual centromere alleles could be easily followed. From the DonkeyB and the mule cell lines, three replicate ChIP-seq data sets were obtained (Methods; Supplemental Table S1).

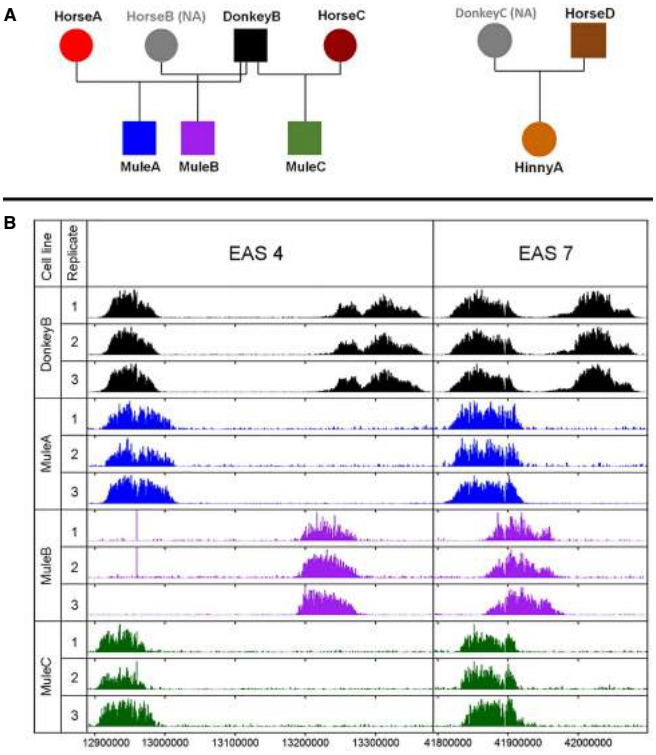


Figure 4. Transmission of satellite-free centromeric domains in hybrids. (*A*) Family trees reporting the genetic relationships among the individuals used in this study. Each color represents an individual, and the same color code is used in *B*. Cell lines from the individuals in gray were not available (NA). (*B*) ChIP-seq analysis performed with the anti-CENPA antibody on chromatin from the DonkeyB cell line and the cell lines from its offspring MuleA, MuleB, and MuleC. For each cell line, the results of three experiments are shown. The centromeres of donkey Chromosomes 4 (EAS4) and 7 (EAS7) are shown as examples, and the other centromeres are reported in Supplemental Figure S7. The EquCabAsiB chimeric genome was used as reference.

To facilitate centromere mapping in these samples, a DonkeyB-derived chimeric genome was assembled from reads as described above for EquCabAsiA. The resulting EquCabAsiB chimeric reference sequence (Supplemental Table S3) was used to map reads deriving from DonkeyB and mule cell lines (Fig. 4B; Supplemental Fig. S7). The irregular shape of some peaks may be due to (1) inaccurate sequence assembly; (2) presence of subpopulations of cells with slightly different centromeric domains; or (3) irregular distribution of CENPA containing nucleosomes.

Figure 4B shows, as examples, the centromeric domains of Chromosomes 4 and 7 in three replicate ChIP-seq experiments carried out with the DonkeyB, MuleA, MuleB, and MuleC cell lines. The centromeres of Chromosomes 4 and 7 (Fig. 4B) showed two distinct peaks in DonkeyB, whereas each mule inherited only one, revealing independent assortment of epialleles and normal monoallelic transmission. For Chromosome 4, the most likely interpretation is that, in MuleA, the left peak was inherited in the same position; in MuleB, the right peak was inherited but shifted by ~50 kb; and, in MuleC, the left peak was inherited with a minor, if any, movement. At Chromosome 7, the left domain seems to have been transmitted to all three mules with a relevant shift of ~50 kb in MuleB. In Supplemental Figure S7, inheritance of the other informative DonkeyB centromeric domains and of horse Chromosome 11 is shown. This analysis revealed additional examples of centromeres that exhibit a striking change in the position or structure of the epiallele in mule or hinny offspring.

In conclusion, we analyzed centromeric domain segregation of 10 donkey centromeres in three mules for a total of 30 independent events. In addition, horse Chromosome 11 centromere was analyzed in three instances. Altogether, we observed clear

positional movement in 5 of 33 transmission events. In the remaining cases, little or no movement was detected.

To test whether centromere sliding can occur during propagation in culture, we examined positional stability in six clonal cell lines isolated from TERT-TERC immortalized fibroblasts (Vidale et al. 2012) derived from MuleA. Following establishment of an immortal cell population, single cells were isolated and expanded for about 40 population doublings and subjected to CENPA ChIP-seq. As shown in Figure 5 and in Supplemental Figure S8, for 10 informative centromeres, no relevant change in peak position and shape was detected among the clones nor between the clones and the immortal parental cell line. These results suggest that the position of centromeres in the immortal cell population was homogeneous in spite of the high number of cell divisions in culture required for immortalization. In addition, during their independent growth for about 40 population doublings, centromere position remained unaltered in all the clones. In light of these observations, we can reasonably exclude in vitro cell culturing as the source of the positional instability observed in the families.

Discussion

Identification and DNA sequence composition of satellite-free centromeres

Here, we have demonstrated, at the sequence level, that an exceptionally high number of *E. asinus* centromeres are devoid of satellite DNA. If more than half of the donkey chromosomes can be stable in the species while being devoid of centromeric satellite DNA, the role of these sequences becomes even more puzzling than previously supposed (Wade et al. 2009; Fukagawa and Earnshaw 2014; Plohl et al. 2014). The 16 satellite-free donkey centromeric domains do not correspond to centromeres on the orthologous horse genomic regions; therefore, they derived from centromere repositioning events that occurred after the separation of the donkey lineage from the horse/donkey common ancestor. Thus, these centromeres are evolutionarily new (ENCs).

The large number of sequenced satellite-free centromeres allowed us to investigate the properties of "centromerizable" genomic regions in a mammal. Our analysis pointed out that satellite-free centromeres are AT and LINE rich. In addition, most satellite-free centromeres contain structural rearrangements relative to *E. caballus* and, interestingly, five of 16 show sequence amplification.

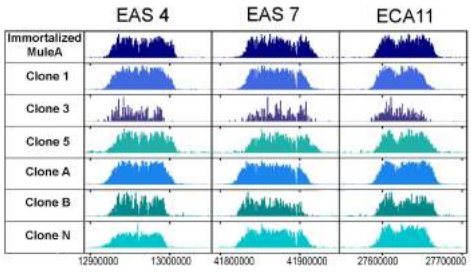


Figure 5. Transmission of satellite-free centromeric domains in clonal cell lines. ChIP-seq analysis of the immortalized cell line obtained from MuleA primary fibroblasts and six clonal derivative cell lines. Three centromeric domains taken as examples are shown (EAS4, EAS7, and ECA11). Results from the remaining centromeres are reported in Supplemental Figure S8. The EquCabAsiB chimeric genome was used as reference.

Sequence analysis of the 16 satellite-free centromeric loci revealed that they are AT rich, LINE rich, and SINE poor (Supplemental Fig. S5; Huang et al. 2015). AT richness is a common feature of centromeres in a number of organisms (Clarke and Carbon 1985; Marshall et al. 2008; Chueh et al. 2009). However, it does not seem to be a necessary requirement (Melters et al. 2013), nor was it seen at the centromere of horse Chromosome 11 (Wade et al. 2009). Enrichment of LINE-1 sequences has been detected in natural human centromeres (Plohl et al. 2014) as well as in clinical neocentromeres (Chueh et al. 2005; Capozzi et al. 2008; Marshall et al. 2008). On the other hand, no association of LINEs was observed in experimentally induced neocentromeres in chicken cell lines (Shang et al. 2010) or in the evolutionary neocentromere of horse Chromosome 11 (Wade et al. 2009). It is not clear whether these features contribute directly to establishment of "centromerizable" genomic domains. The observation that LINE/LTR-rich domains are clustered within the nucleus suggests that this arrangement may be related to function (van de Werken et al. 2017). In this scenario, the sequence composition of the satellite-free donkey centromeres may allow them to partition into subnuclear domains that promote the functional activation of centromeric chromatin.

Comparison between the satellite-free donkey centromeric loci and their horse noncentromeric counterparts demonstrated the presence of rearrangements in most instances (deletions, amplifications, insertions, and inversions) (Supplemental Fig. S2). Although we do not know whether these rearrangements occurred before or after centromere formation, chromosome breakage may promote CENPA binding, as suggested by the observation that CENPA can be recruited at DNA breaks (Zeitlin et al. 2009). Huang et al. (2015) used the BAC locations, mapped in our early work on centromere repositioning (Carbone et al. 2006), to identify donkey scaffolds spanning very extended regions surrounding six neocentromeres. Although they did not detect any obvious increase in chromosome rearrangements over extended (several megabases long) regions, we precisely identified sequence rearrangements contained within functional, CENPA binding, centromeric domains in this work.

Five donkey centromeres exhibit tandem repetition of sequences present in single copy in the horse genome (Fig. 2; Supplemental Figs. S2, S4). These amplified genomic sequences are unrelated to one another, with amplified units ranging in size from 5.3 (EAS16) to 138 (EAS8) kb. These repeated units are AT rich (about 65%) and SINE poor, and four of five are LINE rich. The repeat copy number was variable in the two individuals analyzed, suggesting the existence of polymorphism in the population. On the basis of our estimates, we predict that the amplified regions range in size from 100 up to 800 kb of genomic DNA. It is tempting to speculate that these amplified arrays represent an intermediate stage toward satellite DNA formation.

The presence of "ongoing" amplification at some donkey neocentromeres allows us to propose a new model (Fig. 6) for the maturation of a centromere during evolution, including different routes, some of which involve sequence amplification. According to the model, the presence of amplified sequences at a neocentromere is an indication of its more mature stage compared to nonamplified centromeres. It remains to be demonstrated whether amplification is a necessary step toward centromeric satellite DNA formation. Although the classical definition of satellite

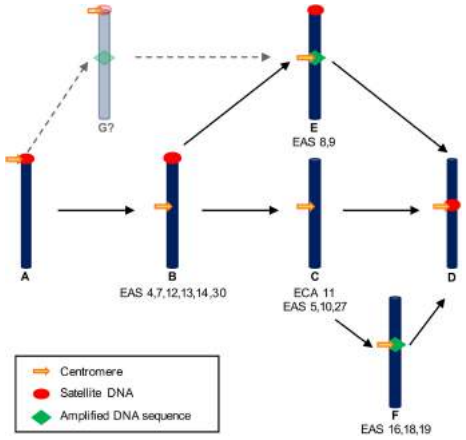


Figure 6. Model for the maturation of a centromere during evolution. Different pathways can be envisaged leading to a fully mature satellite-based repositioned centromere (D) from an ancestral centromere with satellite repeats (A) through satellite-free intermediates (B,C,E,F). The first route (A–D) follows the previously proposed model (Piras et al. 2010): a neocentromere arises in a satellite-free region; satellite repeats may then colonize this repositioned centromere at a later stage, giving rise to a "mature" centromere; meanwhile the ancestral satellite DNA is lost. Alternative routes (A, B, E, D or A, B, C, F, D) imply that, at an already functional satellite-free centromere, amplification occurs as an intermediate step toward complete maturation of the neocentromere. In this model, neocentromere maturation and loss of satellite DNA from the old centromere site are independent events that can occur at different stages during evolution. Donkey chromosomes exemplifying each step are listed, taking into account the position of satellite DNA as previously described (Piras et al. 2010). Horse Chromosome 11 is also reported since its evolutionary stage (C) was previously analyzed (Wade et al. 2009). We cannot exclude that sequence amplification may precede neocentromere formation (G?) but we have no data supporting this possibility.

DNA refers to clusters of tandem repetitions extending for several megabases, the tandem repeat expansions that we observed at these five centromeres may well be considered as an early seed of chromosome-specific centromeric satellites. In this view, these five neocentromeres cannot be considered as bona fide satellite free. To our knowledge, our results represent the first evidence supporting the hypothesis that amplification-like mechanisms can trigger the formation of tandemly repeated DNA sequences within the centromere core.

The heterogeneity of the amplified centromeric units that we observed is compatible with the molecular mechanism proposed for the multistep evolution of amplified DNA in drug-resistant mammalian cell lines (Giulotto et al. 1986). Large domains are amplified initially and, during the following steps, the copy number increases by amplification of subregions of the repeated unit, giving rise to highly condensed arrays of relatively short DNA fragments (Saito et al. 1989).

Although the systems and the time scale are extremely different, similar recombination-based mechanisms (Mondello et al. 2010) might generate novel satellite DNA families following amplification of large segments at neocentromeres. We propose that, in early stages of centromere formation, tandem duplications may arise and evolve through recombination-based meiotic or mitotic mechanisms as demonstrated for primate alpha-satellite families (Schueler and Sullivan 2006; Cacheux et al. 2016).

In the model depicted in Figure 6, satellite DNA recruitment is a late event in centromere maturation. It has been proposed that satellite DNA increases segregation fidelity through binding with specific kinetochore proteins, such as CENPB (Fachinetti et al. 2015). The positional instability of satellite-free centromeres (discussed below) suggests that repetitive DNA arrays may contribute to centromere stability by reducing the impact of positional flexibility.

Positional variability and transmission of satellite-free centromeric domains

The position of centromeric domains can vary between individuals at satellite-free (Purgato et al. 2015) and satellite-bearing (Maloney et al. 2012) centromeres. Here, we show extensive positional allelism, verified by SNV analysis, at most donkey satellite-free centromeres (Fig. 3). Comparison of two donkey individuals (Supplemental Fig. S6) shows that centromere position can vary within genomic regions spanning several hundred kilobases, whereas independent assortment of epialleles in hybrids (Fig. 4B; Supplemental Fig. S7) provides direct proof that each chromosome carries a single centromeric domain. Despite their different positions and associated sequences, all epialleles are rather homogeneous in size, measuring ~100 kb, similar to those of horse Chromosome 11 (Purgato et al. 2015). We can reasonably propose that the sliding phenomenon is common to all satellite-free centromeres, because the analysis of only two individuals allowed us to observe evidence of more than one allele at the majority of informative centromeres (Fig. 3).

An intriguing result obtained from the analysis of the transmission of CENPA binding domains in hybrids was positional movement in five of 33 transmission events. These results demonstrate, for the first time, that centromere sliding can occur in one generation. The extent of this movement is never extreme. Indeed, the centromeric domain in the offspring is always at least partially overlapping the domain of the parent, suggesting that a fraction of CENPA nucleosomes maintains its position, and centromeres do not jump to a completely new location. We can envisage that, in the course of several generations, slight movements accumulate giving rise to nonoverlapping epialleles. In the transmission experiments reported here, we observed instances of substantial centromere movement, on the order of 50–80 kb, that occurred in a single generation. On the other hand, different epialleles at a given centromere are contained within limited regions occupying up to ~600 kb. These observations are consistent with the existence of some sort of boundaries, such as specific patterns of chromatin marks (Sullivan and Karpen 2004; Martins et al. 2016), limiting the region through which CENPA binding domains can move.

The movement of centromeric domains, observed in the family analysis, does not seem to be due to in vitro culturing (Fig. 5; Supplemental Fig. S8) in agreement with the behavior of centromeres in chicken DT40 cell lines (Hori et al. 2017). The stability of the centromeric domains in cultured cells is consistent with a spatially conserved transmission and replenishment mechanism for CENPA nucleosomes (McKinley and Cheeseman 2015; Ross et al. 2016) that, during the mitotic cell cycle, ensures that new CENPA nucleosomes are inserted at centromeric location with high fidelity. The sliding that we observed in the hybrids presumably took place during germline differentiation, meiotic division,

fertilization, or early developmental stages. It is possible that CENPA is mobilized during the extensive chromatin remodeling and epigenetic reprogramming characterizing these stages.

A well-described mechanism of chromatin reorganization is the replacement of histones with protamines (protamine transition) during spermatogenesis. Although CENPA is quantitatively maintained during this process (Palmer et al. 1990), it might slide into adjacent histone-depleted regions. Notably, we observed centromere sliding in both an oocyte-derived horse Chromosome 11 (Supplemental Fig. S7) as well as in several sperm-derived chromosomes in the hybrid offspring (Fig. 4B; Supplemental Fig. S7). Another process which may cause shift of centromeric domains is the meiotic division itself, during which the fidelity of CENPA deposition is poorly understood (McKinley and Cheeseman 2015). In addition, early embryonic cell cycles are highly dynamic in terms of active DNA demethylation and histone modifications and remodeling (Mayer et al. 2000; Santos et al. 2005; Probst and Almouzni 2011). We do not know at which stage centromere sliding may occur, but it is clear that the normally stringent maintenance of CENPA position can become relaxed between generations, possibly during the unique epigenetic transactions of meiosis and early embryogenesis.

Conclusions

We identified satellite-free centromeres at 16 of the 31 chromosome pairs of the donkey. Nearly one-third of the evolutionarily new centromeres of donkey exhibit tandem DNA sequence amplification. These centromeres may be in the process of selecting novel satellite DNA sequences, eventually leading to mature satellite-based centromeres (Fig. 6).

Centromeres can slide by a substantial fraction of their total size in one generation. This mobility appears to be an intrinsic property of CENPA chromatin domains in the equids. Satellite DNA may function to constrain the mobility of the centromere and enforce specific locus identity.

The presence of so many satellite-free centromeres may be due to the fact that the donkey lineage separated recently (about 3 Mya) from the common *Equus* ancestor, and there was not enough evolutionary time for satellite DNA accumulation and centromere maturation (Fig. 6). The observation of centromeres with sequence amplification intermediates supports this hypothesis. An alternative hypothesis, based on the centromere drive model (Malik and Bayes 2006; Henikoff and Furuyama 2010), can be proposed: Although large centromeres with expanded blocks of satellite DNA should be stronger than small ones (Iwata-Otsubo et al. 2017), a selective pressure against satellite DNA accumulation may operate in the donkey.

Methods

Cell lines

Primary fibroblast cell lines from HorseS and DonkeyA were established from the skin of slaughtered animals. Fibroblasts from DonkeyB, HorseA, HorseC, and Hinny were established from skin biopsies of adult animals from Cornell University. HorseD fibroblasts were obtained from testicular tissue of a freshly castrated animal from Cornell. MuleA, MuleB, and MuleC cell lines were derived from three mule conceptuses from normal pregnancies recovered on days 32–34 after ovulation via uterine lavage, as described (Adams and Antczak 2001).

Immortalization of the MuleA fibroblast cell line was carried out as described in Vidale et al. (2012) and in Supplemental Methods.

Horses, donkeys, and (horse × donkey) hybrids from the families used for the study of centromere transmission were maintained at the Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University. Animal care and experiments were carried out in accordance with the guidelines set forth by the Institutional Animal Care and Use Committee of Cornell University under protocol 1986-0216, Douglas F. Antczak PI.

The DonkeyA and HorseS fibroblast cell lines were established from skin samples taken from animals not specifically sacrificed for this study; the animals were being processed as part of the normal work of the abattoirs.

Chromatin Immunoprecipitation (ChIP)

Chromatin was cross-linked with 1% formaldehyde, extracted, and sonicated to obtain DNA fragments ranging from 200 to 800 bp. Immunoprecipitation was performed as previously described (Cerutti et al. 2016) by using a polyclonal antibody against human CENPA protein (Wade et al. 2009) or a human CREST serum (Purgato et al. 2015). Sequencing was performed as described in Supplemental Methods.

Cytogenetic analysis

FISH experiments on horse and donkey metaphase spreads were carried out with a panel of BAC clones (Supplemental Table S2) from the horse library CHORI-241 as previously described (Raimondi et al. 2011; for details, see Supplemental Methods).

Assembly of centromeric regions, sequence analysis, and construction of the chimeric reference genomes

The de novo assembly of the donkey centromeric regions and the construction the chimeric EquCabAsiA and EquCabAsiB references was performed as described in the Supplemental Methods.

Bioinformatic analysis of ChIP-seq data

Reads were aligned to the horse reference genome or to the EquCabAsiA or EquCabAsiB references with Bowtie 2.0 (Langmead and Salzberg 2012). Peak calling was performed with the software MACS 2.0.10 (Zhang et al. 2008). ChIP-seq data were normalized with the deepTools package using a subtractive method (Ramírez et al. 2014). ChIP-seq enrichment plots were obtained with the R software package Sushi (Phanstiel et al. 2014). Data sets were mapped on EquCab2.0 and plotted with Integrative Genomics Viewer (IGV) (Robinson et al. 2011). Details are reported in Supplemental Methods.

SNV analysis

To identify single nucleotide variants (SNVs) in the DonkeyA centromeric regions, we used the IGV software (Robinson et al. 2011) with the EquCabAsiA genome as reference, analyzing the BAM file resulting from read mapping (for details, see Supplemental Methods).

Southern blotting and quantitative PCR (qPCR)

Southern blotting was performed under standard conditions using probes prepared by PCR as described in Supplemental Methods.

For quantitative qPCR amplification, levels were calculated as previously described (Purgato et al. 2015). See Supplemental Methods for details.

Data access

Raw sequencing data from this study have been submitted to the NCBI BioProject database (http://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA385275. De novo assembled centromeric regions of DonkeyA and DonkeyB from this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/genbank/) under accession numbers MF344597–MF344627.

Acknowledgments

We thank Silvia Bione and Paolo Cremaschi (IGM-CNR, Pavia, Italy) for helpful suggestions on the initial bioinformatic analysis; Mariano Rocchi (Department of Biology, University of Bari, Italy) for providing the anti-CENPA antibody; and Claudia Alpini (Fondazione I.R.C.C.S. Policlinico San Matteo, Pavia, Italy) for the CREST serum. The E.G. laboratory was supported by grants from Consiglio Nazionale delle Ricerche (CNR-Progetto Bandiera Epigenomica, Subproject 4.9), from Ministero dell'Istruzione dell'Università e della Ricerca ((MIUR): PRIN Grant No. 2015RA7XZS_002); Dipartimenti di Eccellenza Program (2018–2022) – Dept. of Biology and Biotechnology "L. Spallanzani," University of Pavia (to S.G.N, F.M.P., M.C., E.C., E.R, and E.G.). The K.F.S. laboratory was supported by the Science Foundation Ireland under Grant No.12/A/1370. Funding bodies had no role in the design of the study and collection, analysis and interpretation of data, and in writing the manuscript.

Author contributions: E.G. conceived the study and supervised all experiments. E.G., K.F.S., and E.R. designed the research and wrote the manuscript. S.G.N., F.M.P., R.G., and M.C. carried out most molecular and cell biology experiments and bioinformatic analyses and contributed to result interpretation and figure preparation. J.G.W.McC. carried out bioinformatic analyses. Federico Cerutti, who tragically died on May 30, 2015, gave an essential contribution in the early phases of the study. E.C., F.G., R.M.H., D.F.A., D.M., M.S., and G.P. provided materials and data. D.M., R.M.H., and D.F.A. provided cells and tissues. E.G., K.F.S., E.R., S.G.N., F.M.P., R.G., M.C., F.C., J.G.W.McC., E.C., and G.P. participated in discussions and result interpretation. All authors read and approved the final manuscript.

References

Adams AP, Antczak DF. 2001. Ectopic transplantation of equine invasive trophoblast. *Biol Reprod* **64:** 753–763.

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

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* **17:** 916.

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.

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.

Carbone L, Nergadze SG, Magnani E, Misceo D, Francesca Cardone M, Roberto R, Bertoni L, Attolini C, Francesca Piras M, de Jong P, et al. 2006. Evolutionary movement of centromeres in horse, donkey, and zebra. *Genomics* **87:** 777–782.

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.

Cerutti F, Gamba R, Mazzagatti A, Piras FM, Cappelletti E, Belloni E, Nergadze SG, Raimondi E, Giulotto E. 2016. The major horse satellite DNA family is associated with centromere competence. *Mol Cytogenet* **9:** 35.

Chueh AC, Wong LH, Wong N, Choo KH. 2005. Variable and hierarchical size distribution of L1-retroelement-enriched CENP-A clusters within a functional human neocentromere. *Hum Mol Genet* **14:** 85–93.

Chueh AC, Northrop EL, Brettingham-Moore KH, Choo KH, Wong LH. 2009. LINE retrotransposon RNA is an essential structural and functional epigenetic component of a core neocentromeric chromatin. *PLoS Genet* **5:** e1000354.

Clarke L, Carbon J. 1985. The structure and function of yeast centromeres. *Annu Rev Genet* **19:** 29–55.

Cleveland DW, Mao Y, Sullivan KF. 2003. Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling. *Cell* **112:** 407–421.

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

Eichler EE. 1999. Repetitive conundrums of centromere structure and function. *Hum Mol Genet* **8:** 151–155.

Fachinetti D, Han JS, McMahon MA, Ly P, Abdullah A, Wong AJ, Cleveland DW. 2015. DNA sequence-specific binding of CENP-B enhances the fidelity of human centromere function. *Dev Cell* **33:** 314–327.

Ferreri GC, Liscinsky DM, Mack JA, Eldridge MD, O'Neill RJ. 2005. Retention of latent centromeres in the Mammalian genome. *J Hered* **96:** 217–224.

Fukagawa T, Earnshaw WC. 2014. The centromere: chromatin foundation for the kinetochore machinery. *Dev Cell* **30:** 496–508.

Geigl EM, Bar-David S, Beja-Pereira A, Cothran EG, Giulotto E, Hrabar H, Oyunsuren T, Pruvost M. 2016. Genetics and paleogenetics of equids. In *Wild equids* (ed. Ransom JI, Kaczensky P), pp. 87–104. Johns Hopkins University Press, Baltimore, MD.

Giulotto E, Saito I, Stark GR. 1986. Structure of DNA formed in the first step of CAD gene amplification. *EMBO J* **5:** 2115–2121.

Giulotto E, Raimondi E, Sullivan K. 2017. The unique DNA sequences underlying equine centromeres. *Prog Mol Subcell Biol* **56:** 337–354.

Gong Z, Wu Y, Koblízková A, Torres GA, Wang K, Iovene M, Neumann P, Zhang W, Novák P, Buell CR, et al. 2012. Repeatless and repeat-based centromeres in potato: implications for centromere evolution. *Plant Cell* **24:** 3559–3574.

Han Y, Zhang Z, Liu C, Liu J, Huang S, Jiang J, Jin W. 2009. Centromere repositioning in cucurbit species: implication of the genomic impact from centromere activation and inactivation. *Proc Natl Acad Sci* **106:** 14937–14941.

Hayden KE, Strome ED, Merrett SL, Lee HR, Rudd MK, Willard HF. 2013. Sequences associated with centromere competency in the human genome. *Mol Cell Biol* **33:** 763–772.

Henikoff S, Furuyama T. 2010. Epigenetic inheritance of centromeres. *Cold Spring Harb Symp Quant Biol* **75:** 51–60.

Henikoff S, Ahmad K, Malik HS. 2001. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293:** 1098–1102.

Hori T, Kagawa N, Toyoda A, Fujiyama A, Misu S, Monma N, Makino F, Ikeo K, Fukagawa T. 2017. Constitutive centromere-associated network controls centromere drift in vertebrate cells. *J Cell Biol* **216:** 101–113.

Huang J, Zhao Y, Bai D, Shiraigol W, Li B, Yang L, Wu J, Bao W, Ren X, Jin B, et al. 2015. Donkey genome and insight into the imprinting of fast karyotype evolution. *Sci Rep* **5:** 14106.

Iwata-Otsubo A, Dawicki-McKenna JM, Akera T, Falk SJ, Chmátal L, Yang K, Sullivan BA, Schultz RM, Lampson MA, Black BE. 2017. Expanded satellite repeats amplify a discrete CENP-A nucleosome assembly site on chromosomes that drive in female meiosis. *Curr Biol* **27:** 2365–2373.e8.

Kalitsis P, Choo KA. 2012. The evolutionary life cycle of the resilient centromere. *Chromosoma* **121:** 327–340.

Karpen GH, Allshire RC. 1997. The case for epigenetic effects on centromere identity and function. *Trends Genet* **13:** 489–496.

Kobayashi T, Yamada F, Hashimoto T, Abe S, Matsuda Y, Kuroiwa A. 2008. Centromere repositioning in the X chromosome of XO/XO mammals, Ryukyu spiny rat. *Chromosome Res* **16:** 587–593.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359.

Leeb T, Vogl C, Zhu B, de Jong PJ, Binns MM, Chowdhary BP, Scharfe M, Jarek M, Nordsiek G, Schrader F, et al. 2006. A human-horse comparative map based on equine BAC end sequences. *Genomics* **87:** 772–776.

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

Malik HS, Bayes JJ. 2006. Genetic conflicts during meiosis and the evolutionary origins of centromere complexity. *Biochem Soc Trans* **34:** 569–573.

Maloney KA, Sullivan LL, Matheny JE, Strome ED, Merrett SL, Ferris A, Sullivan BA. 2012. Functional epialleles at an endogenous human centromere. *Proc Natl Acad Sci* **109:** 13704–13709.

Marshall OJ, Chueh AC, Wong LH, Choo KH. 2008. Neocentromeres: new insights into centromere structure, disease development, and karyotype evolution. *Am J Hum Genet* **82:** 261–282.
Martins NM, Bergmann JH, Shono N, Kimura H, Larionov V, Masumoto H, Earnshaw WC. 2016. Epigenetic engineering shows that a human centromere resists silencing mediated by H3K27me3/K9me3. *Mol Biol Cell* **27:** 177–196.
Mayer W, Niveleau A, Walter J, Fundele R, Haaf T. 2000. Embryogenesis: demethylation of the zygotic paternal genome. *Nature* **403:** 501–502.
McKinley KL, Cheeseman IM. 2015. The molecular basis for centromere identity and function. *Nat Rev Mol Cell Biol* **17:** 16–29.
Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D, et al. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol* **14:** R10.
Mendiburo MJ, Padeken J, Fülöp S, Schepers A, Heun P. 2011. *Drosophila* CENH3 is sufficient for centromere formation. *Science* **334:** 686–690.
Mondello C, Smirnova A, Giulotto E. 2010. Gene amplification, radiation sensitivity and DNA double-strand breaks. *Mutat Res* **704:** 29–37.
Montefalcone G, Tempesta S, Rocchi M, Archidiacono N. 1999. Centromere repositioning. *Genome Res* **9:** 1184–1188.
Musilova P, Kubickova S, Vahala J, Rubes J. 2013. Subchromosomal karyotype evolution in Equidae. *Chromosome Res* **21:** 175–187.
Nergadze SG, Belloni E, Piras FM, Khoriauli L, Mazzagatti A, Vella F, Bensi M, Vitelli V, Giulotto E, Raimondi E. 2014. Discovery and comparative analysis of a novel satellite, EC137, in horses and other equids. *Cytogenet Genome Res* **144:** 114–123.
Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, Schubert M, Cappellini E, Petersen B, Moltke I, et al. 2013. Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499:** 74–78.
Palmer DK, O'Day K, Margolis RL. 1990. The centromere specific histone CENP-A is selectively retained in discrete foci in mammalian sperm nuclei. *Chromosoma* **100:** 32–36.
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* **88:** 3734–3738.
Panchenko T, Black BE. 2009. The epigenetic basis for centromere identity. *Prog Mol Subcell Biol* **48:** 1–32.
Phanstiel DH, Boyle AP, Araya CL, Snyder MP. 2014. Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. *Bioinformatics* **30:** 2808–2810.
Piras FM, Nergadze SG, Poletto V, Cerutti F, Ryder OA, Leeb T, Raimondi E, Giulotto E. 2009. Phylogeny of horse chromosome 5q in the genus *Equus* and centromere repositioning. *Cytogenet Genome Res* **126:** 165–172.
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.
Plohl M, Meštrović N, Mravinac B. 2014. Centromere identity from the DNA point of view. *Chromosoma* **123:** 313–325.
Probst AV, Almouzni G. 2011. Heterochromatin establishment in the context of genome-wide epigenetic reprogramming. *Trends Genet* **27:** 177–185.
Purgato S, Belloni E, Piras FM, Zoli M, Badiale C, Cerutti F, Mazzagatti A, Perini G, Della Valle G, Nergadze SG, et al. 2015. Centromere sliding on a mammalian chromosome. *Chromosoma* **124:** 277–287.
Raimondi E, Piras FM, Nergadze SG, Di Meo GP, Ruiz-Herrera A, Ponsà M, Ianuzzi L, Giulotto E. 2011. Polymorphic organization of constitutive heterochromatin in *Equus asinus* ($2n = 62$) chromosome 1. *Hereditas* **148:** 110–113.
Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. 2014. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42:** W187–W191.
Raudsepp T, Lear TL, Chowdhary BP. 2002. Comparative mapping in equids: The asine X chromosome is rearranged compared to horse and Hartmann's mountain zebra. *Cytogenet Genome Res* **96:** 206–209.
Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29:** 24–26.
Ross JE, Woodlief KS, Sullivan BA. 2016. Inheritance of the CENP-A chromatin domain is spatially and temporally constrained at human centromeres. *Epigenetics Chromatin* **9:** 20.
Saito I, Groves R, Giulotto E, Rolfe M, Stark GR. 1989. Evolution and stability of chromosomal DNA coamplified with the CAD gene. *Mol Cell Biol* **9:** 2445–2452.
Santos F, Peters AH, Otte AP, Reik W, Dean W. 2005. Dynamic chromatin modifications characterise the first cell cycle in mouse embryos. *Dev Biol* **280:** 225–236.
Schueler MG, Sullivan BA. 2006. Structural and functional dynamics of human centromeric chromatin. *Annu Rev Genomics Hum Genet* **7:** 301–313.
Shang WH, Hori T, Toyoda A, Kato J, Popendorf K, Sakakibara Y, Fujiyama A, Fukagawa T. 2010. Chickens possess centromeres with both extended tandem repeats and short non-tandem-repetitive sequences. *Genome Res* **20:** 1219–1228.
Steiner CC, Mitelberg A, Tursi R, Ryder OA. 2012. Molecular phylogeny of extant equids and effects of ancestral polymorphism in resolving species-level phylogenies. *Mol Phylogenet Evol* **65:** 573–581.
Stoler S, Keith KC, Curnick KE, Fitzgerald-Hayes M. 1995. A mutation in *CSE4*, an essential gene encoding a novel chromatin-associated protein in yeast, causes chromosome nondisjunction and cell cycle arrest at mitosis. *Genes Dev* **9:** 573–586.
Sullivan BA, Karpen GH. 2004. Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. *Nat Struct Mol Biol* **11:** 1076–1083.
Tolomeo D, Capozzi O, Stanyon RR, Archidiacono N, D'Addabbo P, Catacchio CR, Purgato S, Perini G, Schempp W, Huddleston J, et al. 2017. Epigenetic origin of evolutionary novel centromeres. *Sci Rep* **7:** 41980.
van de Werken HJ, Haan JC, Feodorova Y, Bijos D, Weuts A, Theunis K, Holwerda SJ, Meuleman W, Pagie L, Thanisch K, et al. 2017. Small chromosomal regions position themselves autonomously according to their chromatin class. *Genome Res* **27:** 922–933.
Ventura M, Weigl S, Carbone L, Cardone MF, Misceo D, Teti M, D'Addabbo P, Wandall A, Björck E, de Jong PJ, et al. 2004. Recurrent sites for new centromere seeding. *Genome Res* **14:** 1696–1703.
Ventura M, Antonacci F, Cardone MF, Stanyon R, D'Addabbo P, Cellamare A, Sprague LJ, Eichler EE, Archidiacono N, Rocchi M. 2007. Evolutionary formation of new centromeres in macaque. *Science* **316:** 243–246.
Vidale P, Magnani E, Nergadze SG, Santagostino M, Cristofari G, Smirnova A, Mondello C, Giulotto E. 2012. The catalytic and the RNA subunits of human telomerase are required to immortalize equid primary fibroblasts. *Chromosoma* **121:** 475–488.
Voullaire LE, Slater HR, Petrovic V, Choo KH. 1993. A functional marker centromere with no detectable alpha-satellite, satellite III, or CENP-B protein: activation of a latent centromere? *Am J Hum Genet* **52:** 1153–1163.
Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, Lear TL, Adelson DL, Bailey E, Bellone RR, et al. 2009. Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* **326:** 865–867.
Yang F, Fu B, O'Brien PC, Nie W, Ryder OA, Ferguson-Smith MA. 2004. Refined genome-wide comparative map of the domestic horse, donkey and human based on cross-species chromosome painting: insight into the occasional fertility of mules. *Chromosome Res* **12:** 65–76.
Zeitlin SG, Baker NM, Chapados BR, Soutoglou E, Wang JY, Berns MW, Cleveland DW. 2009. Double-strand DNA breaks recruit the centromeric histone CENP-A. *Proc Natl Acad Sci* **106:** 15762–15767.
Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9:** R137.

Received October 11, 2017; accepted in revised form April 13, 2018.

Birth, evolution, and transmission of satellite-free mammalian centromeric domains

Solomon G. Nergadze, Francesca M. Piras, Riccardo Gamba, et al.

Genome Res. published online April 30, 2018
Access the most recent version at doi:10.1101/gr.231159.117