UNIVERSITA' DEGLI STUDI DI PAVIA

Dipartimento di Biologia e Biotecnologie “Lazzaro Spallanzani”

High resolution molecular cytogenetic approaches for the analysis of the architectural and epigenetic landscape of mammalian centromeres

Alice Mazzagatti

Dottorato di Ricerca in
Genetica, Biologia Molecolare e Cellulare
XXIX Ciclo – A.A. 2013-2016

UNIVERSITA' DEGLI STUDI DI PAVIA

Dipartimento di Biologia e Biotecnologie "Lazzaro Spallanzani"

High resolution molecular cytogenetic approaches for the analysis of the architectural and epigenetic landscape of mammalian centromeres

Alice Mazzagatti

Supervised by Prof. Elena Raimondi

Dottorato di Ricerca in
Genetica, Biologia Molecolare e Cellulare
XXIX Ciclo – A.A. 2013-2016

ABSTRACT

The centromere is the *locus* that drives proper chromosome segregation at meiosis and mitosis, providing the platform for kinetochore assembly. While centromere/kinetochore proteins are conserved across all eukaryotes, the underlying DNA sequence differs among species both in size and in complexity. This paradox suggests that the centromere is an epigenetic structure, which does not depend strictly on the primary DNA sequence.

In the recent years, we demonstrated that in the species belonging to the genus *Equus*, the centromere function and the position of satellite DNA are often uncoupled (Wade *et al.*, 2009; Piras *et al.*, 2009; Piras *et al.*, 2010; Raimondi *et al.*, 2011; Nergadze *et al.*, 2014, Purgato *et al.*, 2015). Moreover, satellite-less centromeres, originated by evolutionary centromere repositioning, are unexpectedly frequent in *Equus* species. As a consequence, satellite based and satellite-less centromeres coexist in single karyotypes: one centromere in the horse (on chromosome 11), sixteen centromeres in the donkey, seventeen in the Grevyi's zebra and seven in the Burchelli's zebra. Thus, we used equid species as a model system especially suitable for the dissection of centromere function.

In a synergical work between the Laboratory of Molecular Cytogenetics and the Laboratory of Cellular and Molecular Biology of the University of Pavia, we have analyzed the overall organization of the different classes of horse satellite DNA (37cen, 2PI and EC137) and the sequence associated with the centromere function. In the horse, the organization of the different satellite DNA families appears to be a mosaic where the three DNA families display an interspersed association of sequence blocks widely variable in size. The molecular organization of the horse centromeres is similar to that of others species and it is composed of CENP-A blocks of variable length immersed in long satellite DNA stretches (Blower et al., 2002) and the major horse satellite DNA family, 37cen, is related to the centromere function in the satellite-based centromeres. Moreover, it has been demonstrated that 37cen is transcriptionally competent.

We also deeply analyzed the centromeric domain of the first natural satellite-less centromere described in literature, the centromere of horse chromosome 11 (ECA11). Analyzing 5 unrelated horses we demonstrated that the centromeric domain of ECA11 is characterized by positional variation, and that in a native mammalian centromere the centromere position can be flexible across a relatively wide (500kb) single-copy genomic region. Our results demonstrated that the positioning of CENP-A binding domains is unrelated to the underlying DNA sequence.

A crucial issue in centromere biology concerns the contribution of satellite DNA to chromosome segregation fidelity. To our knowledge, systematic analyses of the mitotic stability of satellite-less centromeres do not exist. Data from the analysis of pathologic satellite-less centromeres indicate that these marker chromosomes are often present in the individual in mosaic form; this mosaicism may be due to some

intrinsic mitotic instability, but also the selective disadvantage of partial aneuploidy must be considered (Marshall *et al.*, 2008). Human artificial chromosomes, with a conditional centromere, have been used to manipulate the epigenetic state of chromatin and to elucidate the requirements for proper centromere function; it was shown that a dynamic balance between centromeric euchromatin and heterochromatin is essential for kinetochore activity, alphoid satellite DNA stretches having a central role (Nakano *et al.*, 2008; Ohzeki *et al.*, 2015). We decided to analyze the *in vitro* mitotic stability of horse chromosome 11, whose centromere is completely satellite-free, and compare it with that of chromosome 13 (ECA13), which has similar size and a centromere containing long stretches of the canonical horse centromeric satellite DNA families. Our results demonstrated that the segregation accuracy of these two chromosomes is similar, thus suggesting that satellite DNA is dispensable for transmission fidelity.

In view of the absence of repetitive DNA arrays, the only elements that can specify the centromeric function are epigenetic factors. We studied the epigenetic landscape of horse and donkey centromeres through a molecular cytogenetic analysis of the main histone modifications characterizing the centrochromatin. We demonstrated that the satellite-less centromeres, as well as the satellite based-centromeres, are immersed in a heterochromatic environment, even if they contain small amounts of constitutive heterochromatin. This constitutive hyper-condensed heterochromatin defines the borders of the functional centromere domain preventing centrochromatin diffusion. Satellite-less centromeres do also contain facultative heterochromatin since this heterochromatin is prone to be opened and is needed for CENP-A loading. Finally, satellite-less centromeres do contain transcriptionally competent heterochromatin, presumably to interact with *trans* acting lncRNAs transcribed from satellite based centromeres.

ACKNOWLEDGMENTS

I would like to thank my supervisor Prof. Elena Raimondi for the scientific support provided and for the opportunity of improvement, not only under the scientific point of view.

I would like to thank also Prof. Elena Giulotto and her laboratory, for precious collaboration.

CONTENTS

ABBREVIATIONS

BAC	Bacterial Artificial Chromosome
CENP-	CENtromere Protein –
ChIP-on-chip	Chromatin ImmunoPrecipitation – on – chip
ChIP-seq	Chromatin ImmunoPrecipitation – Sequencing
CREST	Calcinosis, Raynaud’s syndrome, Esophageal dysmotility, Sclerodactyly and Telangiectasia
EAS	*Equus asinus*
EBU	*Equus burchelli*
ECA	*Equus caballus*
EGR	*Equus grevyi*
FISH	Fluorescence *In Situ* Hybridization

1. REVIEW OF THE LITERATURE

1.1 - The centromere and its paradox

The centromere is the *locus* that allows the proper chromosomes segregation at meiosis and mitosis, providing the platform for kinetochore assembly (**Figure 1**). The centromere is composed by a functional centromeric domain, named core, which is the site wherein spindle microtubules attach to the kinetochore multi-protein structure. The kinetochore is responsible for the movement of chromosomes to the opposite poles of the cell, ensuring the proper distribution of the genetic material. The centromere core is surrounded by pericentromeric regions, which are the site of sister chromatids cohesion.

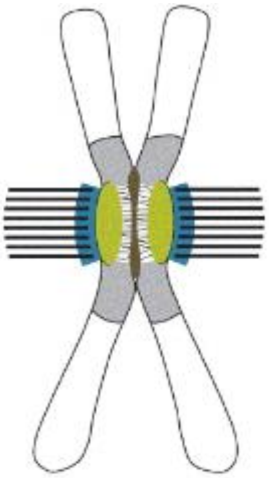

Figure 1 – Schematic representation of a centromere structure. Centromeric chromatin underlies the kinetochore, which contains inner and outer plates (in green and in blue, respectively) that form microtubule-attachment sites (in black). Pericentromeric heterochromatin (in grey) flanks centromeric chromatin (zebra-striped) and contains a high density of cohesins (in brown) which mediates sister-chromatid cohesion (adapted from Bailey *et al.*, 2016).

Generally speaking, the centromere structure becomes more complex while moving along the evolutionary scale (**Figure 2**) (Kalitsis and Choo, 2012).

In *Saccharomyces cerevisiae* (**Figure 2a**), the centromeric region is defined by a sequence of 125 bp, which contains three different DNA sequence elements: CDEI, CDEII and CDEIII. Two of them, CDEI and CDEIII, are extremely conserved at each chromosome and point mutations inhibit their activity. CDEI is involved in maintaining the cohesion and allowing the separation of sister chromatids, while CDEII binds the CBF3 (Centromere Binding Factor 3 - a multisubunit protein complex that binds to centromeric DNA and initiates kinetochore assembly complex) and is responsible of a proper kinetochore assembly (Cleveland *et al.*, 2003). Unlike the previous ones, CDEII is not conserved in terms of sequence, but in terms of minimal sequence length (76 to 84 bp) and A+T rich content (90%). CDEII interacts with Cse4, a modified centromere-specific histone H3variant, which is highly conserved protein motif present at all eukaryotic centromeres (Cheeseman *et al.*, 2002).

The fission yeast *Saccharomyces pombe* (**Figure 2c**) has regional centromeres, which range in size from 40 to 100 kb, composed of an unconserved central core region (cnt) surrounded by centromere specific inner repeats (imr) and long inverted repeats (otr). Both the central core and the inner repeats are bound by Cnp1, the modified centromere-specific histone H3 variant (Cleveland *et al.*, 2003). The same structural organization is presented in the centromeres of *Candida albicans* (**Figure 2b**), in which the core sequence is flanked by small inverted repeats (Henikoff *et al.*, 2001).

In higher eukaryotes, centromeres acquire greater complexity and they are usually localized within tandemly repeated DNA sequences that form long arrays (satellite DNA) in tightly packed heterochromatic regions (Plohl *et al.*, 2014).

A well characterized example is the centromere of *Drosophila melanogaster* (**Figure 2d**), which extends for about 420 kb and is composed of tandemly repeated arrays in which various transposons are interspersed (Sun *et al.*, 1997). Human centromeres (**Figure 2e**) mainly consist of alpha-satellite DNA, composed of monomer units of 171 bp arranged in tandem head-to-tail arrays. Monomers are organized into chromosome-specific higher-order repeats (HORs) that are reiterated thousands of times (Schueler and Sullivan, 2006). In the mouse, centromeric and pericentromeric regions are composed of two highly conserved satellite DNA families named minor and major satellite. The minor satellite is composed of 120-bp AT-rich monomers that occupy 300-600 kb and act as the site of kinetochore assembly and spindle microtubules attachment. The major satellite is more abundant and consists of 234 bp monomers flanking the minor satellite. The major satellite is implicated in heterochromatin formation and sister chromatids cohesion (Komissarov *et al.*, 2011).

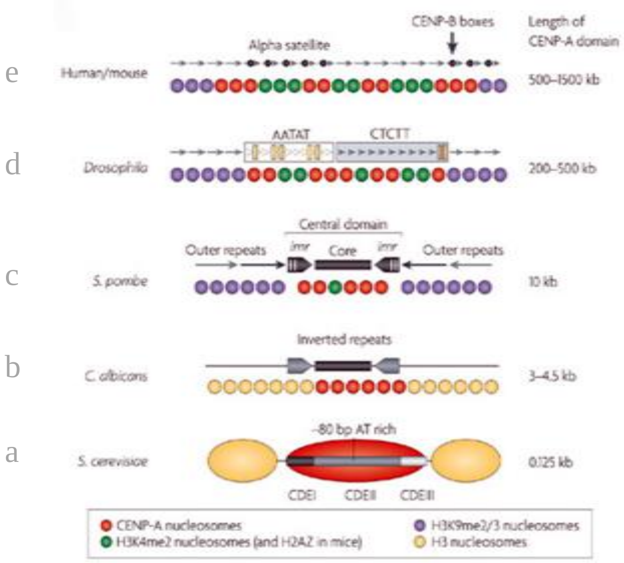

Figure 2 - Schematic depiction of centromeric DNA structure which becomes more complex while moving along the evolutionary scale (from bottom to the top). A schematic representation of the centromeric nucleosomal organization of *C. albicans*, *S. pombe*, *Drosophila* and. humans is also reported (Allshire and Karpen, 2008).

The comparison between the various sequence motifs that make up fully functional centromeres in higher eukaryotes has revealed a total absence of sequence conservation. However, satellite DNA monomers share, among species, a common unit length which nearly corresponds to the nucleosomal DNA unit length (the alpha-satellite unit in primates is 171 bp; in the fish *Sparus aurata*, the centromeric repeat is 186 bp; in the insect *Chironomus pallidivittatus*, it is 155 bp; in both *Arabidopsis* and maize, it is 180 bp; in rice it is 168 bp) (Dawe *et al.*, 1996; Choo, 1997; Shelby *et al.*, 1997; Dong *et al.*, 1998). It can be speculated that the “selection” for nucleosomal unit DNA length would be a driving force in the evolution of the centromeric satellite DNA sequences, in respect of their structural (non-coding) role in the genome. Moreover, a minimal size of satellite DNA arrays seems to exist. For example, in *Drosophila melanogaster*, 420 kb of primary tandem repeats are required for a fully functional centromere, and length reduction leads to chromosome malsegregation (Murphy and Karpen, 1995). In man, the smallest natural supernumerary mini-chromosomes retain at least 100 kb of alpha-satellite DNA and human artificial chromosomes need megabases long stretches of alpha-satellite DNA to be mitotically stable (Ikeno *et al.*, 1998; Yang *et al.*, 2000).

All these evidences led to the idea that specific repeated sequence elements specify centromere location when they are present in a sufficient number of copies (Henikoff *et al.*, 2001). However, this hypothesis was rejected in the light of discovery of neocentromeres and of pseudo-dicentric chromosomes. Neocentromeres are ectopic centromeres lacking any common repetitive element observed occasionally in humans (Depinet *et al.*, 1997; Tyler-Smith *et al.*, 1999; Warburton *et al.*, 2000; Lo *et al.*, 2001). Pseudo-dicentric chromosomes are structural dicentric chromosomes in which one centromere was functionally inactivated without changes in the underlying DNA sequence (McKinley and Cheeseman, 2016).

The failure to detect a universal sequence or a common motif distinguishing centromere function prompted researchers to look at non-DNA sequence determinants able to maintain centromeres (Choo, 2000).

A number of proteins were found only at centromeres. Many of them are present at centromeres only at mitosis and meiosis contributing to the assembly of kinetochore on centromeric chromatin and connecting the centromere to spindle microtubules (Perpelescu and Fukagawa, 2011). Among constitutive centromeric proteins the most important is the marker of centromere function: CENP-A (Earnshaw and Ruthfield, 1985). CENP-A is a centromere-specific histone H3 variant (Sullivan *et al.*, 1994) which is present at constitutive centromeres as well as at neocentromeres (Saffery *et al.*, 2000) but is absent from inactivated centromeres (Sullivan and Willard, 1998). Although histone H3 is evolutionarily conserved, the centromeric histone H3 variants diverge among species. Probably this is due to the need to interact with centromeric DNA (Malik and Henikoff, 2001) which is one of the most rapidly evolving components of eukaryotic genomes (Csink and Henikoff, 1998).

In conclusion, even if the process of chromosome inheritance is highly conserved across all eukaryotes, the DNA and protein components specific to the centromeric *locus* differ among species (See paragraphs 1.2 and 1.3). The discrepancy between the need to preserve the function and the lack of conservation of the actors that are being actively involved in this process is known as “centromere paradox” (Henikoff *et al.*, 2001).

1.2 - The centromere from the DNA point of view

The centromeric region of most eukaryotes is composed of two major repetitive DNA components: transposable elements and satellite DNA sequences.

Transposable elements (TEs) are DNA sequences that can move to new genomic locations and form interspersed repeats if replicated during the process of movement (Kazazian, 2004; Tollis and Boissinot, 2012). They are classified, according to the mechanisms of transposition, as RNA-mediated transposable elements or DNA-mediated transposable elements. Among TEs, LTR-retrotransposons (*i.e.* mobile genetic elements that can replicate themselves through reverse transcription of their

RNA and integrate the resulting cDNA into another *locus*) are highly abundant at centromeric and pericentromeric regions both in plants and in animals (Pimpinelli *et al.,* 1995; Copenhaver *et al.,* 1999; Schueler *et al.,* 2001; Cheng *et al.,* 2002).

The other major class of centromeric repetitive elements, satellite DNA, is defined as a “class of diverse tandemly repeated DNA sequences that comprise long arrays localized in a tightly packed heterochromatin” (Plohl *et al.,* 2014). These sequences undergo rapid evolution according to the principles of concerted evolution (Melters *et al.,* 2013). Centromeric DNA evolution is driven by several mechanisms of nonreciprocal sequence transfer, such as unequal crossing-over, gene conversion, rolling circle replication and transposition-related mechanisms (Dover, 1986). In particular, unequal crossing-over and gene conversion are the elective mechanisms involved in satellite DNA dynamics (Talbert and Henikoff, 2010). It has been also demonstrated that segmental duplication is an important evolutionary force giving rise to the amplification of satellite DNA arrays (Horvath *et al.,* 2005; Ma and Jackson, 2006). Satellite DNA evolution is linked to reproductive isolation and speciation since differences among individuals in the centromere region cause centromere drive, leading to incompatibility of homologous chromosomes in hybrids and ultimately to postzygotic isolation, thus triggering speciation (Bachmann *et al.,* 1989; Henikoff *et al.,* 2001).

Despite the extreme diversity of satellite DNA sequences at eukaryotic centromeres, some sequences seem to be shared. For example, in human alpha-satellite DNA, as well as in various classes of repetitive elements in several mammalian species, a conserved 17 bp sequence has been found (Ohzeki *et al.,* 2002; Alkan *et al.,* 2011). This motif, named CENP-B box, is a binding site for the protein CENP-B whose role is poorly understood (Masumoto *et al.,* 2004). CENP-B box-like motifs were also found in unrelated satDNAs of some distant invertebrates and plants, suggesting a potential functional role of this protein (Canapa *et al.,* 2000; Gindullis *et al.,* 2001; Mravinac *et al.,* 2005; Meštrović *et al.,* 2013).

The diversity of DNA sequences localized in functional centromeres and/or pericentromeres has been evidenced not only in terms of different satDNAs and their organization, but also in terms of other sequences contribution (**Figure 3**).

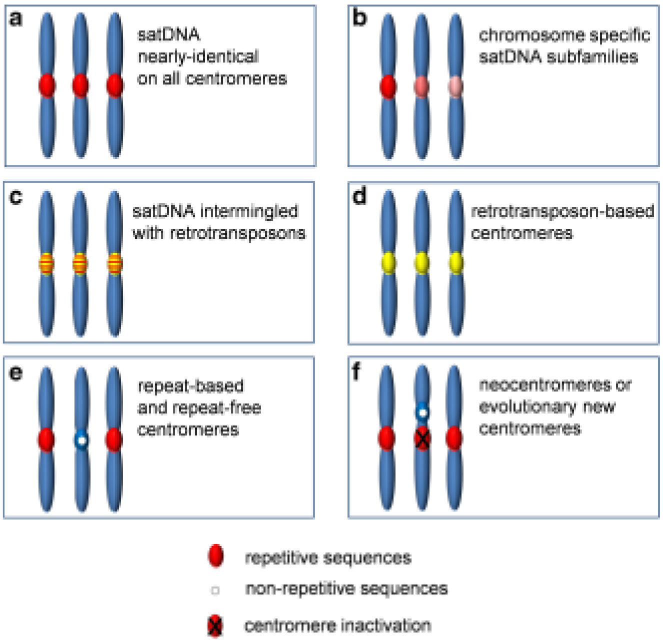
a
satDNA nearly-identical on all centromeres
b
chromosome specific satDNA subfamilies
c
satDNA intermingled with retrotransposons
d
retrotransposon-based centromeres
e
repeat-based and repeat-free centromeres
f
neocentromeres or evolutionary new centromeres
repetitive sequences
non-repetitive sequences
centromere inactivation

Figure 3 – Schematic representation of different DNA sequences in different centromere types (adapted from Plohl *et al.*, 2014).

A global sequence characterization of rice centromeric satellite DNA by sequencing and ChIP experiments, did not reveal any difference between monomers included in the functional centromere and pericentromeric arrays (**Figure 3a**) (Macas *et al.*, 2010). It has been suggested that the absence of chromosome-specific satellite DNA families is related to a high sequence homogenization in the meiotic prophase stage (Durajlija Žinić *et al.*, 2000; Mravinac and Plohl, 2010).

Functional DNA sequences in different centromere types exist. For example, in human a defined number of monomers are organized into chromosome-specific higher-order repeats (HORs) that are reiterated thousands of times creating chromosome specific satellite DNA families (**Figure 3b**) (Schueler and Sullivan, 2006).

In rice, substantial portions of centromere-specific retrotransposons are present. These retrotransposons are intermingled with the satellite DNA and both types of repeated sequences can bind the centromere-specific histone H3 variant (CEN-H3), which confers the centromere identity (**Figure 3c**) (Ma *et al.*, 2007).

In some chromosomes of maize (Wolfgruber *et al.,* 2009) and wheat (Li *et al.,* 2013) species-specific centromeric retrotransposons are the predominant DNA sequences associated with CEN-H3 (**Figure 3d**).

Even if the majority of eukaryotes display satellite-based centromeres, some exceptions exist. Satellite-less centromeres, discussed in detail in the next paragraphs, have been found occasionally in human pathology (neocentromeres) (Vouillaire *et al.*, 1993; Marshall *et al.*, 2008) and in extant species (evolutionary new centromeres) (Wade *et al.*, 2009; Locke *et al.*, 2011; Shang *et al.*, 2010) (**Figure 3e-f**).

1.2.1 - The role of centromeric satellite DNA

The functional role of centromeric satellite DNA has been long debated and a number of hypotheses have been proposed to explain the recruitment, by the majority of eukaryotic centromeres, of large stretches of satellite DNA. In the recent past several evidences assigned to centromeric satellites specific role even if, as a matter of fact, it is dispensable and completely satellite-free functional chromosomes exist (Plohl *et al.,* 2012).

Centromeric repetitive DNA is typically devoid of active genes, thus it may aid the formation of a heterochromatic environment which would favour the stability of the chromosome during mitosis and meiosis (Marshall *et al.*, 2008; Plohl *et al.*, 2008; Plohl *et al.*, 2014). Pericentromeric repetitive DNA might inhibit spreading of the centromere over neighboring genic regions (Sullivan, 2002). It has also been proposed that the satellite DNA may improve the cohesion and the separation of sister chromatids. For sure, it had been demonstrated that centromeric satellite DNA is transcribed and that the transcription of the centromeric regions is essential for centromere maintenance (Steiner and Henikoff, 2015). Transcripts from satellite DNA seem 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 (Rošić *et al.*, 2014; Biscotti *et al.*, 2015; Rošić and Erhardt, 2016).

Transcripts homologous to centromeric and pericentromeric repetitive sequences have been identified in several organisms such as yeast (Ohkuni and Kitagawa, 2011; Choi *et al.,* 2012), mouse (Ferri *et al.,* 2009), and humans (Saffery *et al.,* 2003; Wong *et al.,* 2007). The transcription of satellite DNA seems to be strictly related to kinetochore assembly; indeed, defects in transcriptional competence lead to chromosome malsegregation (Ohkuni and Kitagawa, 2011; Chan *et al.,* 2012).

In *S. pombe*, transcripts derived from the pericentromeric repetitive elements have been proposed to be involved in heterochromatin formation and maintenance by the RNA interference (RNAi) machinery (Volpe *et al.,* 2002; Motamedi *et al.,* 2004; Verdel *et al.,* 2004). RNA polymerase II (RNA Pol II) transcribes centromeric and pericentromeric satellite DNA sequences in long noncoding RNAs (lncRNAs). These lncRNAs are turned in double strand by an RNA-directed RNA polymerase (Rdp1). Subsequently, they are cleaved by Dicer in order to produce siRNAs which in turn recruit factors involved in heterochromatin assembly. At this stage, siRNAs are loaded first by the ARC (Activator Recruited Cofactor) complex and then by the RNA-induced initiation of transcriptional gene silencing complex (RITS). RITS complex uses single-stranded siRNAs to recognize and to target specific chromosome regions by a mechanism that involves either siRNA-DNA or siRNA-nascent transcript base pairing interactions. The localization of RITS at centromeric DNA repeats and its association to centromeric transcripts is Clr4 methyltransferase dependent. Clr4 methylates the histone H3 at lysine 9, providing a binding site for Chp1, a protein belonging to the RITS complex, and stabilizing the tethering of RITS itself. Moreover, Clr4 modifies adjacent histones promoting heterochromatin spreading (Biscotti *et al.,* 2015).

The involvement of RNAi in heterochromatin formation in other eukaryotic organisms is still debated. In chicken, the accumulation of pericentromeric transcripts after downregulation of Dicer seems to indicate a similar mechanism (Fukagawa *et al.,* 2004).

In human, another RNAi-dependent mechanism, leading to the establishment of the heterochromatic state at centromere, has been recently proposed (Maida *et al.,* 2014). It has been hypothesized the involvement of a telomerase reverse transcriptase (TERT) that might act in non-telomeric regions. However, evidence of defective heterochromatin in Dicer-deficient human and murine cells was found (Fukagawa *et al.,* 2004; Kanellopoulou *et al.,* 2005). This suggests that in man there might be two independent mechanisms regulating heterochromatin maintenance at the centromere.

The transcripts of satellite DNA seem to have a role also in maintaining centromere identity. The transcription of centromeric tandemly repeated sequences with a specific size is related to CENP-A loading (Okada *et al.*, 2009). At human centromeres, 1.3 kb lncRNAs act as for the targeting and the loading of CENP-A onto the centromeric DNA (Quénet and Dalal, 2014).

The juxtaposition of other important kinetochore proteins, as CENP-C or the kinase Aurora B, depends on transcripts originated from the centromere (Wong *et al.,* 2007; Ferri *et al.,* 2009). It has been demonstrated that, in the mouse, transcription of centromeric minor satellite and Aurora B kinase activity are mutually dependent (Ferri *et al.,* 2009). It has recently been shown that also in HeLa cells satellite I transcripts are associated with Aurora B and INCENP activity (INner CENtromeric Protein) (Ideue *et al.,* 2014). Moreover, the absence of SATIII pericentromeric transcripts determines defects in chromosome segregation and partial loss of kinetochore components in *Drosophila.*

The expression levels of centromeric satellite DNA is influenced by various cellular stresses, such as heat shock, exposure to hazardous chemicals and ultraviolet radiation, as well as hyperosmotic and oxidative conditions (Jolly *et al.,* 2004; Bouzinba-Segard *et al.,* 2006; Valgardsdottir *et al.,* 2008; Eymery *et al.,* 2009; Hsieh *et al.,* 2011; Hall *et al.,* 2012; Enukashvily and Ponomartsev, 2013). For example, in response to heat shock, nuclear stress bodies originate at pericentromeric regions in human cells (Denegri *et al.,* 2002; Jolly *et al.,* 2002). Under this condition, the epigenetic status of pericentromeric DNA changes and specific euchromatic histone modifications occur. In addition, the transcripts of specific pericentromeric satellites become highly polyadenylated and the transcription of these satellite DNAs is necessary for the recruitment of heat shock factor 1 (HSF1) and RNA Pol II (Jolly *et al.,* 2004; Rizzi *et al.,* 2004).

In mouse, the transcription of the centromeric minor satellite is induced by chemical exposure. The high levels of transcripts impair centromere function, affecting centromere chromatin condensation and sister chromatids cohesion leading to aneuploidy (Bouzinba- Segard *et al.,* 2006).

Also in several types of cancer differences in the expression level of satellite repeats, correlated with the decondensation of pericentromeric heterochromatin, have been reported. (Shumaker *et al.,* 2006; Alexiadis *et al.,* 2007; Enukashvily *et al.,* 2007; Ehrlich *et al.,* 2008; Eymery *et al.,* 2009; Ting *et al.,* 2011; Zhu *et al.,* 2011). The absence of tumor suppressor proteins which have the centromeric domain as target, leads to a considerable increase in pericentromeric satellite transcripts, and consequently, cells undergo segregation defects and an overall genomic instability (Frescas *et al.,* 2008; Zhu *et al.,* 2011).

Even if the precise function of the centromeric satellite transcription process remains unclear, the abnormal variation of centromeric satellite DNA transcription in stress conditions and in cancer suggests that these transcripts may play a crucial role in genome stability. Finally, since regional centromere position is not strictly specified by the DNA sequence, it is possible that the kinetochore position on the underlying DNA might drift slightly. In this case, repetitive arrays could provide a safety buffer within which such drift would be harmless (Fukagawa and Earnshaw, 2014).

1.3 - The centromere from the kinetochore point of view

The chromatin organization consists of individual DNA molecules wrapped around histone proteins (Kornberg, 1974; Olins and Olins, 1974). Together, they form the nucleosome particle, whose core contains one $(H3\text{-}H4)_2$ tetramer and two H2A-H2B dimers (Luger *et al.,* 1997).

One exception is represented by centromeric chromatin, in which domains of nucleosomes containing the canonical histone H3 are intermingled with domains of nucleosomes that contain the histone H3 variant CEN-H3 (also known as CENP-A) (Earnshaw and Rothfield, 1985; Talbert and Henikoff, 2013). This arrangement contributes to the three-dimensional organization of centromeric chromatin (Blower *et al.,* 2002; Ribeiro *et al.,* 2010) and the presence of this centromere-specific histone variant is necessary and sufficient to confer the centromere function to any genomic region (Saffery *et al.*, 2000; Heun *et al.,* 2006; Mendiburo *et al.,* 2011).

CENP-A mediates the specific recruitment of centromere and kinetochore proteins. Furthermore, the properties of CENP-A nucleosomes are critical for the exclusive deposition of CENP-A at the centromere and to prevent its aberrant assembly at non-centromeric locations.

1.3.1 - CENP-A

The discovery of antibodies against the centromere region in the serum from patients affected by the autoimmune CREST syndrome (calcinosis, Raynaud's syndrome, esophageal dysmotility, Sclerodactyly and Telangiectasia), led to the identification of the first set of three canonical human centromeric proteins: CENP-A, CENP-B, and CENP-C (Moroi *et al.,* 1980; Earnshaw and Rothfield, 1985).

As already mentioned before, CENP-A is the centromere-specific histone H3 variant, as well as the epigenetic mark of centromere identity (Warburton *et al.,* 1997; Vafa and Sullivan, 1997). Its recruitment at the centromere is necessary for centromere function and is also essential for the assembly of all known kinetochore components (Regnier *et al.*, 2005; Liu *et al.*, 2006; Fachinetti *et al.*, 2013). CENP-A was found at the active centromeres of dicentric chromosomes as well as at the all identified satellite-less neocentromeres (See paragraph 1.5) (Earnshaw and Migeon, 1985; Marshall *et al.*, 2008). Moreover, the artificial targeting of CENP-A to an ectopic chromosomal *locus* is sufficient to establish a structure able to mediate the microtubules attachment and to ensure chromosome segregation (Heun *et al.,* 2006; Barnhart *et al.,* 2011; Logsdon *et al.,* 2015).

Specific aminoacidic sequences within CENP-A nucleosomes also confer centromere specific functions through the direct binding of the core kinetochore proteins CENP-N and CENP-C. CENP-N binds directly to the CATD (CENP-A targeting domain) of CENP-A (Carroll *et al.,* 2009; Carroll *et al.,* 2010). CENP-C engages extensive contacts with the CENP-A nucleosome and with other histones within the CENP-A nucleosome (Kato *et al.,* 2013). The structural properties of CATD make the tetramers that contain CENP-A more rigid than the tetramers which contain the canonical histone H3 (Sekulic *et al.,* 2010).

The deposition of new CENP‑A is uncoupled to DNA replication and the new CENP‑A molecules are deposited only during the subsequent G1 phase (Black *et al.,* 2007). This different timing between replication and new CENP-A molecules deposition, opens to the question of CENP-A dilution during the S phase. During the

G1 phase, CENP-A deposition is strictly coordinated by the activity of several assembly factors that ensure the faithful deposition of new CENP-A containing nucleosomes exclusively at centromeres (Dunleavy *et al.*, 2009; Foltz *et al.*, 2009; Bassett *et al.*, 2012). This regulated deposition of CENP-A ensures the epigenetic propagation of the centromere at a preexisting location on each chromosome. Many organisms have strategies to prevent the ectopic deposition of CENP-A which could determine an inappropriate attachment to the spindle. However, a proofreading mechanism to remove ectopic CENP-A has not yet been identified in vertebrates (Bodor *et al.*, 2014).

Although CENP-A is an essential component of most centromeres, it is not the only driver of centromere specification. Additional molecular features contribute to defining an active centromere, including the properties of the underlying DNA sequence (see previous paragraph 1.2), the composition of the surrounding chromatin and post-translational modifications of CENP-A itself (see paragraph 1.4).

1.3.2 - CENP-B

Another important protein found at the centromere domain is CENP-B. It is a DNA-binding protein that recognizes a 17 bp sequence, named "CENP-B box", through its amino-terminal region and dimerizes through its carboxy-terminal region (Earnshaw *et al.*, 1987). The CENP-box sequence motif, firstly identified in the human alpha-satellite by Masumoto and colleagues in 1989 (Masumoto *et al.*, 1989), is also conserved in the mouse minor satellite DNA (Okada *et al.*, 2007).

CENP-B derives from transposases mobilizing DNA transposons of the pogo family (Smit and Riggs, 1996). In the CENP-B aminoacidic sequence, three domains involved in the exonuclease activity were substituted, inhibiting CENP-B ability to promote transposition (Marshall and Choo, 2012).

The exact role of this protein is controversial. *De novo* centromere formation in human/mammalian artificial chromosomes requires the presence of alpha-satellite DNA containing the binding motifs for centromeric CENP-B protein (Okada *et al.*, 2007). On the other hand, CENP-B knockout mice are viable (Hudson *et al.*, 1998). The nonlethal CENP-B knockout mouse phenotype could be explained by the functional redundancy of this protein (*i.e.* different proteins exert the same function), whatever its function (Toth *et al.*, 1995; Smit and Riggs, 1996; Kipling and Warburton, 1997; Hudson *et al.*, 1998; Kapoor *et al.*, 1998; Casola *et al.*, 2008). However, recent studies have highlighted that the CENP-B protein works alone without functionally redundant partners. This means that CENP-B is not involved in the formation of an active kinetochore during mitosis (Hudson *et al.*, 1998; Kapoor *et al.*, 1998; Perez-Castro *et al.*, 1998).

In addition, functional human neocentromeres and the human Y chromosome lack the CENP-B box, thus the bound protein (Choo, 2000; Amor and Choo, 2002; Okada *et al.*, 2007). Interestingly, both types of centromeres shown a lower ability to bind CENP- A if compared with the other centromeres (less than 50% of the amount of CENP-A) (Irvine *et al.*, 2004).

1.3.3 - CCAN complex

The purification of the proteins of the centromere/kinetochore interface, collectively referred to as the constitutive centromere-associated network (CCAN; also known as the interphase centromere complex - ICEN) was obtained by immunoprecipitation of the centromeric domain with monoclonal anti-human CENP-A antibodies. CCAN is formed of 39 proteins associated with CENP-A nucleosomes including canonical centromeric proteins, chromatin remodeling complexes, heterochromatin-related proteins, polycomb group proteins, motor proteins, and proteins with unknown functions. However, the number of proteins that are constitutively localized to kinetochores throughout the cell cycle is 16, since some of the CCAN proteins are recruited only during cell division (Sugata *et al.*, 1999; Nishihashi *et al.*, 2002; Foltz *et al.*, 2006; Izuta *et al.*, 2006; Okada *et al.*, 2006; Hori *et al.*, 2008; Amano *et al.*, 2009). CCAN purification gave the most comprehensive view of the composition of centromere/kinetochore, but did not yield any clues concerning the hierarchy of interactions among CCAN components (Peperlescu and Fukagawa, 2011).

CENP-A and all CCAN proteins are in the inner kinetochore plate (Kingwell and Rattner, 1987; Cooke *et al.*, 1990; Saitoh *et al.*, 1992; Warburton *et al.*, 1997; Wan *et al.*, 2009; Suzuki *et al.*, 2011). Based on genetic and biochemical analyses, the CCAN proteins were classified into different subgroups: the CENP-C group, the CENP-T/W group, the CENP-H/I/K(/L/M/N) group, the CENP-O/P/Q/R/U group, and the CENP-S/X group.

CENP-C is a constitutive centromere protein which interacts with CENP-A nucleosomes (Carroll *et al.*, 2010). Deletions and point mutations in the N-terminal region of CENP-C revealed that this protein is important for the localization of other centromeric proteins including the Mis12 complex which is required for normal chromosome alignment and segregation and for kinetochore formation during mitosis (Liu *et al.*, 2006; Kwon *et al.*, 2007; Milks *et al.*, 2009).

Like CENP-C, the CENP-T/W complex bridges interactions between the centromeric chromatin platform and outer kinetochore components and recent studies demonstrated that CENP-T molecule has a motile structure that can stretch between the inner and outer kinetochore when tension is applied (Suzuki *et al.*, 2011).

The CENP-H/I/K group connects CENP-A nucleosomes and microtubules. The depletion of any component of this group induces defects in the chromosome alignment and in kinetochore assembly, determining chromosome mal-segregation (Fukagawa *et al.,* 2001; Nishihashi *et al.,* 2002; Okada *et al.,* 2006).

Knockout cells for CENP-L, CENP-M, and CENP-N show strong mitotic defects (Okada *et al.,* 2006). CENP-L depletion induces monopolar spindles in most mitotic cells (McClelland *et al.,* 2007). CENP-M-deficient cells exhibit mitotic aberrations and aneuploidy (Foltz *et al.,* 2006; Izuta *et al.,* 2006; Okada *et al.,* 2006). Depletion of CENP-N reduces deposition of newly synthesized CENP-A into centromere, leading to a decrease of levels of the CENPs at kinetochores. (Carroll *et al.,* 2009). For this reason, it has been proposed that CENP-N works as a decoder of information carried by CENP-A nucleosomes, which is necessary to recruit the CCAN.

The depletion of CENP-S/X group induces mitotic abnormalities in human and chicken cells and its presence is required for the assembly of outer kinetochore proteins (Amano *et al.,* 2009).

The CENP-O/P/Q/R/U group forms a heterogeneous complex with an important role in the recovery from spindle damage (Hori *et al.,* 2008).

Cells depleted of any centromeric protein from this group are viable, but in conditions needing recovery from spindle damage, this protein is required for adhesion and prevention of premature sister chromatid separation (Foltz *et al.,* 2006; Hori *et al.,* 2008).

1.4 - Centromere specific histone modifications

Post-translational histone modifications regulate functional interchanges between different chromatin environments; specific patterns of histone modifications are involved in assembly, maintenance, and modification of chromatin three-dimensional structure (Jenuwein and Allis, 2001).

A wide range of post-translational modifications, including methylation, acetylation, phosphorylation and ubiquitination can occur at the N-terminal tails of all histones (Jenuwein and Allis, 2001; Fischle *et al.*, 2003). Centromeric chromatin (or centrochromatin) consists of a mixture of both CENP-A-containing and H3-containing nucleosomes, immerses in a heterochromatic environment and is peculiar since shows both euchromatic and heterochromatic features (Bergmann *et al.*, 2012). A centromere specific ratio between typical euchromatic and typical heterochromatic histone modifications is crucial for centromere identity, creating a ‘permissive’ chromatin structure needed for CENP-A recruitment (Mellone and Allshire, 2003; Quénet and Dalal, 2014).

The pericentric regions contain histone modifications that typically mark the heterochromatin, such as the di- and tri-methylation of lysine 9 of histone H3 (H3K9me2 and H3K9me3) (Peters *et al.,* 2003; Rice *et al.,* 2003).

H3K9me2 is a marker of facultative heterochromatin and is involved in gene silencing. This kind of modification is present in the regions flanking the centromere core both in man and in *Drosophila*. A semi quantitative analysis carried out on extended chromatin fibers revealed that H3K9me2 does not overlap, or overlaps minimally, with the edges of the CENP-A-containing domain. However, this is not a rule since ChIP analysis of rice centromeric regions indicated that H3K9me2 is present within the core of the centromere (Nagaki *et al.*, 2004; Sullivan and Karpen, 2004; Bailey *et al.*, 2016).

H3K9me3 is a constitutive heterochromatin marker which was found in the pericentromeric region of *Drosophila*, mouse and human chromosomes (Peters *et al.*, 2003; Rice *et al.*, 2003) while was completely absent in the centromere core. In man, H3K9me3 is concentrated in the pericentromeric region of chromosomes containing large blocks of satellite DNA and is also located in sequences which are far away from CENP-A domains (Sullivan and Karpen, 2004).

This distinctive centromeric chromatin state contributes to maintain the centromeric size, counteracting the spreading of the functional centromeric chromatin and defining the borders of CENP-A binding domains (Martins *et al.*, 2016). At metaphase, when chromosomes are condensed, this centromeric chromatin environment drives the arrangement of CENP-A and H3 containing nucleosomes. Blocks of CENP-A nucleosomes are pushed to the external face of the centrochromatin, to interact with the kinetochore proteins, while H3-containing nucleosomes lie between the sister chromatids (Blower *et al.*, 2002).

As expected, the heterochromatic modifications are absent from the centromere core as well as in the pericentromeric regions (Taddei *et al.*, 2001; Sullivan and Karpen, 2004) but, surprisingly, at the centromere is present a marker of transcriptionally competent heterochromatin, such as the dimethylation of histone H3 at lysine 4 (H3K4me2) (Lehnertz *et al.*, 2003; Guenatri *et al.*, 2004; Smith *et al.*, 2011). Obviously, the amount of H3K4me2 at the centromere is less than the other euchromatic regions on chromosome arms since in centromere *locus*, this histone H3 modification is closely associated to CENP-A-containing region (Martins *et al.*, 2016).

1.5 - Human neocentromeres

Discovered for the first time in 1993 (Voullaire *et al.*, 1993), human neocentromeres are usually formed in regions devoid of satellite DNA, after chromosomal rearrangements which remove or disrupt the constitutive centromere.

So far, only about 100 pathological neocentromeres have been described. These neocentromeres show a high degree of heterogeneity both in terms of the sequences which they are associated with, both in terms of the chromosomal position in which they are formed. Nevertheless, some chromosomal regions seem to be "hotspots" (**Figure 4**) of neocentromere seeding, such as the long arm of chromosomes 3, 13, 15 and Y (Marshall *et al.*, 2008). It must be underlined that neocentromeres arising in the same chromosome region, indeed involve different *loci* at the DNA sequence level (Hasson *et al.,* 2011).

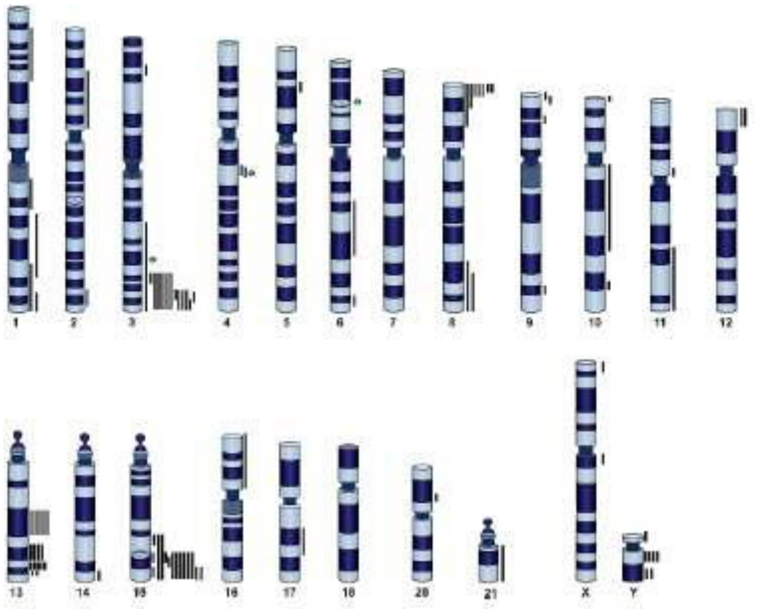

Figure 4 – Graphics representation of pathological neocentromeres. Clinical neocentromeres are indicated with black bars to the right of chromosomes ideograms (Rocchi *et al.,* 2009).

The common causes of pathological neocentromeres formation are chromosome rearrangements after chromatid breaks (**Figure 5**). The resulting acentric fragments can be stabilized by the formation of neocentric supernumerary linear or ring chromosomes, and this situation is often associated to a pathological phenotype (Alonso *et al.*, 2003; Burnside *et al.*, 2011).

After chromatid breakage (**Figure 5I**), the acentric fragment can segregate in two possible ways (**Figure 5II**). After subsequent replication, the broken ends of the acentric fragment rejoin to create an inverted duplication (**Figure 5III**). Neocentromere formation occurs at this stage. If the neocentric fragment segregates with its sister chromatid, the result is partial tetrasomy for the duplicated fragment (**Figure 5IV**, left panel). If the centric fragment segregates with the neocentric fragment, the broken ends of the centric fragment can be stabilized by telomere restitution, and the result is partial trisomy for the duplicated fragment (**Figure 5IV**, right panel). Neocentric chromosomes are often present in the individuals in mosaic form. This mosaicism maybe due to the mechanisms of marker chromosome formation or to some intrinsic mitotic instability of the neocentromere, but the selective disadvantage of partial aneuploidy is likely to be a contributing factor.

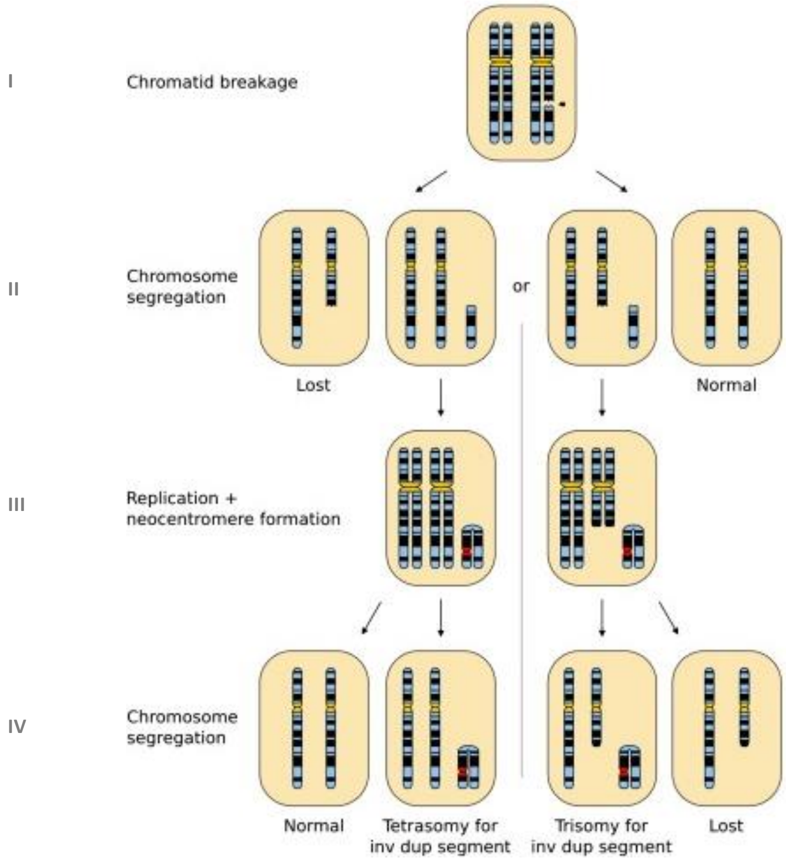
I
Chromatid breakage
II
Chromosome segregation
or
Lost
Normal
III
Replication + neocentromere formation
IV
Chromosome segregation
Normal
Tetrasomy for inv dup segment
Trisomy for inv dup segment
Lost

Figure 5 – Mechanism of pathological neocentromere formation at mitosis. The neocentromeres colored red. The resulting effect on the karyotype is listed underneath each alternative rearrangement (Marshall *et al.*, 2008).

Neocentromeres can arise also on chromosomes in which the constitutive centromere appears unchanged at the DNA sequence level but functionally inactive as demonstrated by the absence of CENP-A that is, conversely, bound to the new centromeric site (Warburton *et al.*, 1997; Voullaire *et al.*, 1999). Since no clinical symptoms are related to this type of neocentromeres, they were discovered by chance through amniocentesis. So far, only eight cases have been described. These human repositioned centromeres do not show mosaicism and are stably transmitted through the generations when the repositioning event occurs in gametes.

Due to genomic instability, also in tumors the presence of neocentromeres has been documented. However, there are few clinical records due to a limit of routine diagnostic analysis rather than to an exceptional nature of the phenomenon (Blom *et al.,* 2010). Among the tumors analyzed, well-differentiated liposarcomas (WDLPS) show the recurrence of neocentromeres formation, which are a pathognomonic characteristic of these tumors. These neocentromeres appear in the form of supernumerary ring or giant chromosomes containing amplified genetic material (Italiano *et al.,* 2009).

Concerning the region on which neocentromeres can arise, it has been observed that neocentromeres are assembled near the break points of chromosome rearrangements. It has been shown that CENP-A transiently binds regions with double-strand breaks and this suggests that these damaged sites may start the neocentromerization (Zeitlin *et al.,* 2009). Human neocentromeres are found both in gene-desert regions and in areas that contain actively transcribed genes. Moreover, they can form in regions containing repeated DNA, although most of them are not associated with such sequences (Alonso *et al.,* 2010; Burrak and Berman, 2012). Furthermore, it seems that LINE-like retrotransposable sequences are important to stabilize these ectopic centromeres: a decrease in their transcription levels compromises the correct functionality during mitosis (Chueh *et al.,* 2009). All the neocentromeric sequences analyzed so far show a high presence of retrotransposable sequences and a significant enrichment in A+T (> 60%); it can be hypothesized that the presence of interspersed repetitive sequences and a high AT content can somehow favor the acquisition of centromeric function and the subsequent assembly of the kinetochore (Mehta *et al.,* 2010).

1.6 - Evolutionary neocentromeres

Evolutionary neocentromeres (ENC) are centromeres that move along the chromosome without structural chromosome rearrangements. Montefalcone and colleagues in 1999 (Montefalcone *et al.,* 1999), unequivocally demonstrated, for the first time, the existence of the centromere repositioning phenomenon. While tracing the phylogeny of chromosome 9 in primates, the authors observed that the position of the centromere changed while the order of molecular markers was conserved. It was therefore hypothesized that the centromere function was shifted along the chromosome without any structural rearrangement (**Figure 6**). In the last decade, a number of ENC were described in primates and other mammals (Ventura *et al.,* 2004; Cardone *et al.,* 2006; Piras *et al.,* 2010). Human pathological neocentromeres have been used as models to study the centromere repositioning phenomenon. These centromere repositioning events "in real time" mimic the events that lead to the formation of evolutionary neocentromeres; for this reason, many of the assumptions formulated on ENC derived from the study of human neocentromeres.

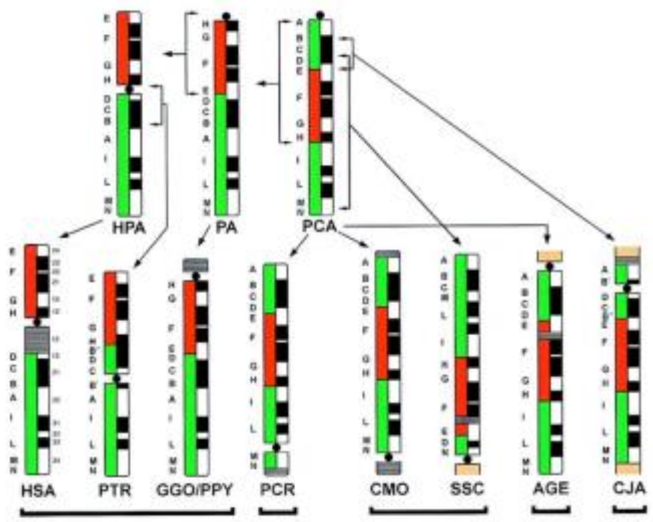

Figure 6 - Schematic representation of evolutionary history of chromosome 9 in primates. Regions orthologous corresponded to the human 9p (red) and 9q (green) are shown on the left of each ideogram (right). The hypothesized pericentric or paracentric inversions are indicated by square parentheses spanning the inverted cytogenetic segment (Montefalcone *et al.*, 1999).

For example, a relationship between ENC and human neocentromeres emerged during the study of chromosome 13. Chromosome 13 is, from an evolutionarily point of view, highly conserved and probably corresponds to the ancestral primate one, which in turn differs from the ancestor of mammals only for a small inversion (Cardone *et al.*, 2006). In old world monkeys, the repositioning of the centromere occurred in the middle of long arm (13q21) and independently, in the same position, an ENC was found also in pigs. In addition, there is a number of clinical human neocentromeres localized in the same band. Molecular cytogenetic studies also indicate that this region is extremely plastic and this leads to the conclusion that there is a non-random pattern of chromosomal evolution that involves specific regions within the mammalian genome in which are recurrent duplications and, on the evolutionary scale, large rearrangements (Rocchi *et al.,* 2009).

Many studies have demonstrated that the repositioning of the centromere during evolution is not a rare phenomenon. Comparing the human and macaque karyotype, fourteen centromere repositioning events were found: nine in the evolutionary line of the macaque, while five in that of man (Ventura *et al.*, 2007). Since the radiation from the common ancestor of the two species took place about 25 million years ago, a repositioning event every 3000 years was estimated. Considering the number of translocations that occurred in the same span of time, only four of these rearrangements were found, which means one in every 12 million of years. It can be concluded that ENCs represent a significant driving force in karyotype evolution.

1.7 - *In vitro*-induction of neocentromere formation

A number of strategies have been adopted to induce *de novo* centromere formation, including the artificial generation of genomic rearrangements, the introduction of centromere-associated DNA into cells and the over-expression of centromeric proteins to trigger the seeding of a neocentromere in a non-centromeric region (Kalitsis and Choo, 2012).

The production of acentric chromosomes has commonly been obtained by irradiation or through recombination systems. The rescue of acentric fragments was successfully obtained in *S. cerevisiae*, in *Drosophila*, and in chicken (Shang *et al.*, 2013). Subsequently, the size of CENP-A binding domains and the DNA sequences associated with the centromere function, were analyzed through chromatin immunoprecipitation-sequencing (ChIP-seq). The CENP-A binding domains span about 40 kb at each neocentromere, without any preference for specific DNA sequences (Shang *et al.*, 2013).

The introduction of putative centromere DNA sequences into cells has been used in several organisms to define the minimal region needed for *de novo* centromere formation (Kalitsis and Choo, 2012). Early experiments have been demonstrated that, in the budding yeast, the centromeric DNA sequence is needed for full chromosome segregation activity (Fitzgerald-Hayes *et al.*, 1982) but, surprisingly, the centromeric sequences are not able to ensure a proper chromosome segregation in *S. cerevisiae* even though the centromeres shared similar sequence structure (Heus *et al.*, 1990; Ohkuma *et al.*, 1995; Kitada *et al.*, 1996; Stoyan and Carbon, 2004).

The transformation of yeast artificial chromosomes (YACs) containing the larger regional centromeres (40 to 100 kb) of the fission yeast *S. pombe*, has been suggest that the flanking repetitive regions are needed for full chromosome stability of the artificial chromosomes (Steiner and Clarke, 1994).

The integration of human alpha satellite DNA into mammalian cell lines showed the capacity to execute the centromere function through the binding the centromeric proteins, but did not provide a stable chromosome segregation (Haaf *et al.*, 1992; Larin *et al.*, 1994). These studies were subsequently followed by the generation of human artificial chromosome constructs complete with centromere, telomere and intervening genomic DNA for transfection into human cells (Harrington *et al.*, 1997). The centromere DNA sequences within such artificial chromosomes were able to bind active centromere proteins and provide stable chromosome inheritance for many cell divisions. Intriguingly, positive results were obtained with human artificial chromosomes containing the CENP-B box motif in the alpha-satellite arrays. This discovery added a further puzzle to centromere biology: since it has been demonstrated that CENP-B is not essential for cell viability (*cenp-b* knockout mice are normal), it plays an important role in *de novo* artificial centromere formation and suppresses the formation of additional centromeres on chromosomes (Hudson *et al.*, 1998; Kapoor *et al.*, 1998; Perez-Castro *et al.*, 1998; Ohzeki *et al.*, 2002: Okada *et al.*, 2007).

The overexpression of centromere proteins, particularly of CENP-A, has been performed in the attempt to induce neocentromere formation in non-centromeric sites. CENP-A overexpression in *Drosophila* produced successful result however, this method does not work in human cultured cells, in which no functional ectopic kinetochores were observed (Van Hooser *et al.*, 2001; Heun *et al.*, 2006). In a different study, CENP-A overexpression and mis-targeting were found to be associated to genome instability in human primary colorectal cancer (Tomonaga *et al.*, 2003).

1.8 - The genus *Equus* as a model system: a paradigm for the study of genome plasticity

The order *Perissodactyla* (i.e. odd-toed, mammals characterized by an odd number of fingers) includes three extant families: *Tapirida*e, *Rhinocerotidae* ed *Equidae*. The *Tapirida*e and the *Rhinocerotidae* belong to the suborder *Ceratomorpha*, while the *Equidae* are in the *Hippomorpha* suborder. This classification, proposed by Wood in 1937, is sustained and confirmed by different studies, including mitochondrial DNA sequence analysis (Pitra and Veits, 2000). Paleonthological and molecular evidences suggest that the divergence of the extant Perossidactyl suborders took place in Laurasia about 56-54 million years ago (Springer *et al.*, 2003).

Phylogenetic analyses, based on interspecific chromosome painting, allowed the reconstruction of the hypothetic Perissodactyl ancestral karyotype (PAK), which comprises 72-76 chromosomes. The ambiguity in the chromosome number is explained by the ancestral polymorphic state of some perissodactyl chromosomes or, alternatively, by breakpoint reuse and fusion/fission events.

Following the radiation from the common ancestor, the karyotypes of species belonging to the *Ceratomorpha* suborder remained quite stable; cytogenetic analyses on living species show a prevalence of acrocentric chromosomes, similarly to the hypothetical ancestral karyotype. On the contrary, the equid karyotypes underwent an evolutionary acceleration after the divergence from the common ancestor 3 million years ago. The karyotypes of the living *Equus* species are predominantly characterized by meta- and submeta-centric chromosomes derived from fusions among ancestral acrocentric elements (Trifonov *et al.,* 2008).

Relying on paleontological and molecular data, it has been possible to date every single divergence node among families and species belonging to the Perissodactyl order (Xu *et al.,* 1996; Tougard *et al.,* 2001; Murphy *et al.,* 2007). *Equus* speciation was accompanied by a huge rate of rearrangements ranging from 2.9 to 22.2 per million years, an 80-fold increase compared to that of ancient *Ceratomorpha* (less than 0.3 rearrangements per million years) (**Figure 7**) (Trifonov *et al.,* 2008).

The Equids evolutionarily radiation dates back to 3 million years ago and the evolutionarily radiation of the extant species belonging to this family dates back to 0.89-1.07 million years ago (Yang *et al.,* 2003).

Nowadays, the genus *Equus* includes: two horse species – the domestic horse (*Equus caballus* - ECA) and the Przewalski horse (*Equus przewalskii* - EPR) – five donkey species – the onagro (*Equus hemionus onager* - EHO), the selvatic african donkey (*Equus africanus* - EAF), the selvatic asian donkey (*Equus hemionus* - EHE), the domestic donkey (*Equus asinus* - EAS), and the Tibetan emione (*Equus kiang* - EKI) – and four zebra species – the Grevy zebra (*Equus grevyi* - EGR), the zebra of the lowlands (*Equus quagga* - EQA), the mountain zebra (*Equus zebra hartmannae* - EZH) and the Burchelli zebra (*Equus burchelli* - EBU).

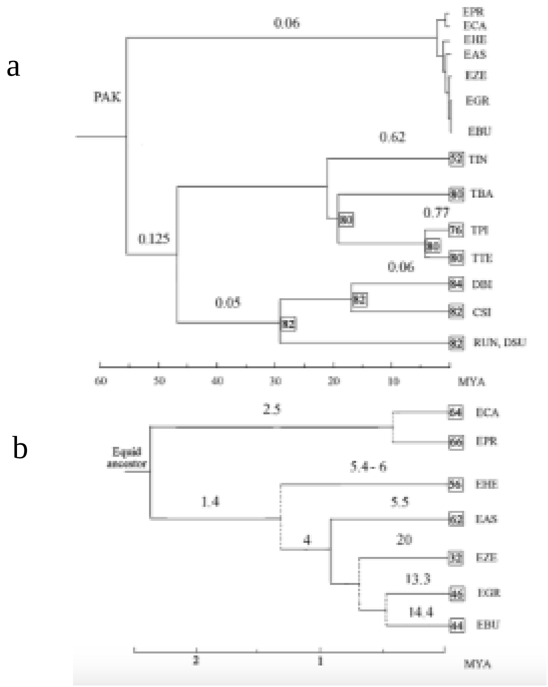

Figure 7 - **Rate of *Ceratomorpha* karyotype evolution. (a)** Perissodactyls. **(b)** Equids. The numbers in the squares indicate the diploid number of chromosomes. The numbers upon the branches represent the average number of rearrangements per millions of years (Trifonov *et al.,* 2008).

The rapid karyotype evolution of these species has been documented also by comparative cytogenetic studies; through comparative chromosome painting and by comparing the banding patterns thanks to digital imaging, it was demonstrated that, despite their recent evolution, the morphological similarity and the possibility to inbreed the karyotypes largely differ. A great variability in the chromosomes number, from a minimum of 32 in *Equus zebra* to a maximum of 66 in *Equus przewalskii*, with a lot of structural differences, was observed (Ryder *et al.,* 1978). Data about the chromosomal architecture in different equid karyotypes and the high variability in terms of chromosomal karyotype number and the rate of shuffling, indicate that the equids evolution is one of the most rapid observed among mammals, comparable only to the one of rodents.

1.9 - The genus *Equus* as a model system: a paradigm for the study of the centromere

As mentioned above, comparative cytogenetic studies demonstrated that during the evolution of eukaryotes, the position of the centromere can change without structural rearrangements, generating Evolutionarily New Centromeres (ENC) (Montefalcone *et al.,* 1999; Ventura *et al.,* 2001). The first event that leads to repositioning probably is the progressive loss of the centromeric function at the level of the constitutive centromere, followed by the acquisition of epigenetic marks in an ectopic position along the chromosome. In evolutionary time scale, at the repositioned centromere sequences of highly repetitive DNA can accumulate presumably conferring a higher stability during chromosome segregation (Marshall *et al.*, 2008). As a result, species that rapidly evolve like equids, are a perfect model system for the study of the dynamics and of the mechanisms at the basis of the evolution of karyotype and, particularly, of the centromeric region. Indeed, studies on equids karyotype revealed a surprisingly high number of centromere repositioning events (one in the horse, sixteen in the donkey, seventeen in the Grevyi's zebra and seven in the Burchelli's zebra). Therefore, these data suggest that centromere repositioning played a driving role in equids evolution (Carbone *et al.*, 2006; Piras *et al.*, 2010).

A peculiar characteristic of this model system emerged during the analysis of the distribution of the two major horse satellite DNA families, 37cen and 2PI, in *E. caballus* (ECA), *E. asinus* (EAS), *E. grevyi* (EGR) and *E. burchelli* (EBU) (Piras *et al.,* 2010). Through FISH experiments it was demonstrated that some chromosomes lack these sequences at the centromere, while these sequences are present in a terminal position (**Figure 8**). These non-centromeric repetitive sequences probably represent the trace of ancestral centromeres on the acrocentric Perossidactyl ancestor. Moreover, several centromeres are completely devoid of satellite DNA (**Figure 8**), such as the evolutionary neocentromere of horse chromosome 11 (ECA11) which is completely devoid of satellite DNA (**Figure 8a**) and was also the first natural vertebrate centromere sequenced and characterized (Wade *et al.,* 2009). As a consequence, the peculiar feature of this model is that satellite based and satellite-less centromeres coexist in single karyotypes.

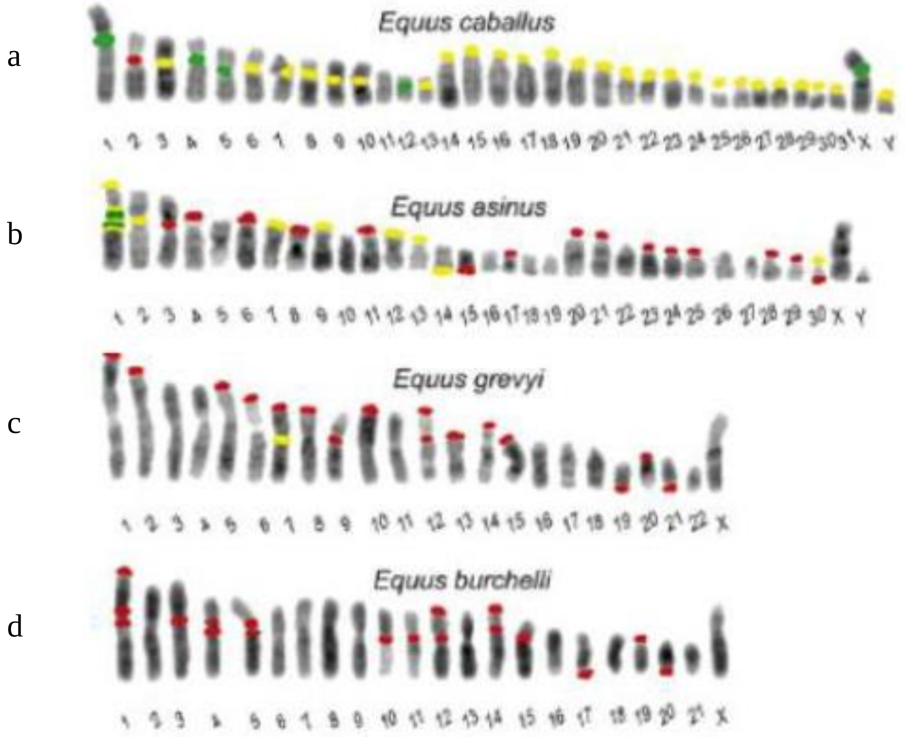

Figure 8 - Schematic representation of the distribution of the major centromeric satellite DNAs of Equids in *Equus caballus* (a), *Equus asinus* (b), *Equus grevyi* (c) and *Equus burchelli* (d). FISH on metaphase chromosomes: in green they are highlighted the *loci* that hybridize only with the 37cen probe, the ones that are positive to the 2PI probe are highlighted in red, and the ones that hybridize with both the probes are highlighted in yellow (Piras *et al.*, 2010).

The analysis of satellite DNA position and of the centromeric functional domain in relation with phylogeny in equid species offers a series of snapshots of the centromere repositioning process (**Figure 9**).

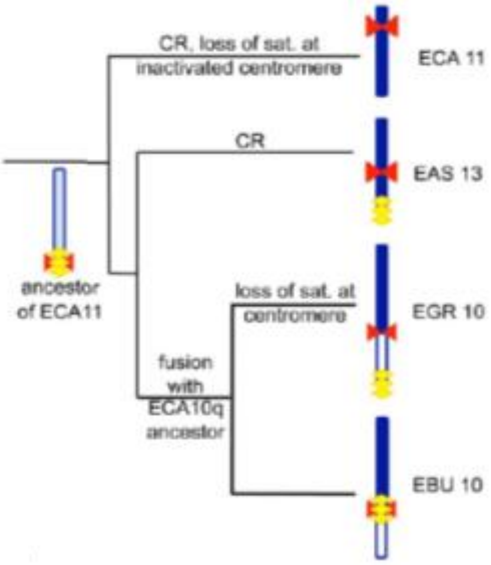
CR, loss of sat. at inactivated centromere
ECA 11
CR
EAS 13
ancestor of ECA11
loss of sat. at centromere
EGR 10
fusion with ECA10q ancestor
EBU 10

Figure 9 - The hypothetical events leading to the formation of four groups of orthologous chromosomes from *E. caballus* (ECA11), *E. asinus* (EAS13), *E. grevyi* (EGR10) and *E. burchelli* (EBU10) (Piras *et al.*, 2010).

The comparison of ECA11 with its orthologous counterparts in *E. asinus* (EAS13), *E. grevyi* (EGR10q) and *E. burchelli* (EBU10q) is an example of different stages of evolutionary neocentromeres formation. It has been hypothesized that the ancestral chromosome from which ECA11, EAS13, EGR10q and EBU10q derive was acrocentric and contained satellite sequences at its centromere. The centromeric location of this hypothetical ancestral chromosome, now corresponds to ECA11q-tel, EAS 13p-tel, EGR10-cen and EBU10-cen. In *E. caballus*, the centromere was shifted in its present position, where no satellite DNA is present. The centromere of EAS13 is also evolutionary new and lacks satellite DNA at the centromere. It can be supposed that, after the fusion that gave rise to EGR10 and EBU10, centromeric satellite DNA was lost in EGR10 and maintained in EBU10. The satellite DNA found on EGR10p-tel might represent the relic of the centromere of the ancestral acrocentric chromosome (Piras *et al.*, 2010).

Based on these results it was formulated a hypothesis, summarized in **Figure 10**, that describes the possible events that led to the formation of evolutionarily neocentromeres in equids, in accordance to the ones previously formulated based on the study of ENC in primates and of pathological neocentromeres in humans.

The initial event of evolutionary repositioning would be the loss of function of the constitutive centromere, followed by the gain of the epigenetic signals in a non-centromeric position (**Figure 10a**). These events would lead to the formation of a centromere in a new chromosome region devoid of satellite DNA, without involvement of DNA sequence alterations (**Figure 10b**). The repetitive arrays present at the level of the ancestral centromere will be maintained in the first step of maturation, but non-reciprocal sequence transfer, unequal crossing-over and transposition-related mechanisms, will lead to the loss of the satellite DNA sequences at the ancestral centromere (**Figure 10c**). The "young" neocentromere can gradually accumulate, during several successive generations, repetitive DNA through various recombination-based mechanisms. Satellite sequences seem to be incorporated at repositioned centromere sites in a subsequent stage (**Figure 10d**), since they probably confer an adaptive advantage, possibly by increasing the accuracy of chromosome segregation. Alternatively, the accumulation of satellite sequences may be a neutral process driven by the presence of heterochromatin in the centromeric DNA (Piras *et al.*, 2010).

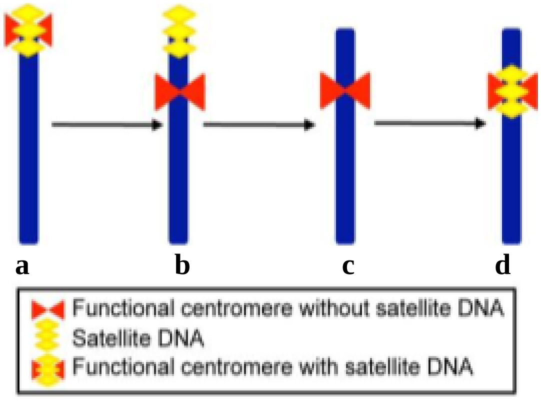
a
b
c
d
Functional centromere without satellite DNA
Satellite DNA
Functional centromere with satellite DNA

Figure 10 - Schematic representation of a four-stage mechanism for the formation of a neocentromeres during the evolution. (a) Ancestral acrocentric chromosome provided with DNA satellite (yellow). **(b)** Submetacentric chromosome derived from the repositioning of the centromere; this chromosome maintains the DNA satellite sequences (yellow) in terminal position, in correspondence of the old centromere, whereas the neocentromere (red) is devoid of repetitive sequences. **(c)** Submetacentric chromosome derived from **(b)** where the terminal satellite sequences are lost. **(d)** Submetacentric chromosome in its stage of full maturity, in which satellite DNA (yellow) is present at the neocentromere (Piras *et al.*, 2010).

2. AIMS OF THE RESEARCH

The work described in this thesis is part of a collaborative project involving the laboratory of Molecular Cytogenetics – directed by professor Elena Raimondi – and the laboratory of Molecular and Cellular Biology – leaded by professor Elena Giulotto – aimed at studying mammalian centromere structure, identity and function. To achieve this target, species belonging to the genus *Equus* are used as a biological model system since satellite-based centromeres and satellite-free centromeres coexist in a single karyotype (Wade *et al.*, 2009; Piras *et al.*, 2010).

In particular, during my PhD program I was involved in the following projects:

- Discovery and comparative analysis of a novel satellite DNA family in the horse, in the donkey and in two zebras. The physical relationships among the new satellite DNA family and the two major horse satellite sequences were investigated in the horse by two color-FISH on metaphase chromosomes, mechanically stretched chromosomes and combed DNA.
- Analysis of the functional organization of satellite-based centromeres in the horse.
  The relation among the three major classes of satellite DNA and CENP-A, which identifies the functional centromeric domains, was analyzed through immuno-FISH on mechanically stretched chromosomes and extended chromatin fibers.
- Deep analysis of the satellite-free centromeric domain of horse chromosome 11 (ECA11) and of donkey satellite-free centromeres.
  The functional centromeric domains were examined at the single molecule level by means of immuno-FISH on extended chromatin fibers.
- *In vitro* analysis of the mitotic stability of horse chromosome 11 whose centromere is satellite-free.
  The mitotic behavior of ECA 11 was compared with that of horse chromosome 13, with a satellite-based centromere, under different experimental conditions by FISH on interphase nuclei and micronuclei.
- Analysis of the centromeric histone modifications in the horse and in the domestic donkey.
  To study the epigenetic state of the centromeric chromatin in satellite-based and satellite-free centromeres, four post-translational histone modifications were analyzed by two color immunofluorescence on metaphase chromosomes, mechanically stretched chromosomes and extended chromatin fibers.

3. MATERIALS AND METHODS

3.1 - Cell lines

Primary fibroblast cell lines from the horse (HSF-B; HSF-C; HSF-D; HSF-E, HSF-G), the domestic donkey (EASn), the Grevy's zebra (EGR) and the Burchelli's zebra (EBU), previously isolated and established in the laboratory of Molecular and Cellular Biology (Prof. Elena Giulotto), were used. Fibroblasts were cultured in Dulbecco's modified Eagle's medium (Euroclone), supplemented with 20 % foetal bovine serum (Euroclone), 2 mM glutamine, 2 % non-essential amino acids and 1 % penicillin/streptomycin. Cells were maintained at 37°C in a humidified atmosphere of 5% CO_2.

3.2 - Metaphase spreads preparation

Mitotically active cells were collected flushing the medium on the cell monolayer and then were centrifuged at 1200 rpm (Z380 centrifuge, Hermle) for 10 minutes. The pellet was resuspended in 75 mM KCl hypotonic solution then incubated at 37°C for 15 minutes. Cold fixative (acetic acid:methanol - 1:3) was added and centrifuged at 1200 rpm (Z380 centrifuge, Hermle) for 30 minutes. The fixation was repeated twice. Then, slides were prepared by dropping the cell suspension perpendicularly to the slides and then air-dried. The preparation was stored, in an appropriate volume of cold fixative, at -20°C.

3.3 - Stretched chromosomes preparation

Mitotically active cells were collected as before. They were centrifuged at 1400 rpm (Z380 centrifuge, Hermle) for 8 minutes and resuspended in 75 mM KCl, 0.8 % Na-citrate, H_2Obd (1:1:1) hypotonic solution for 10 minutes. The cell density in the hypotonic mixture was adjusted to 10^5 cells/ml. After the hypotonic treatment, the cell suspension was cytocentrifuged onto silanized glass slides at 750 rpm (Z300 centrifuge, Hermle) for 4 minutes and fixed in -20°C methanol for 30 min.

3.4 - Chromatin fibers preparation

The cells were treated with trypsin, collected and then centrifuged at 1400 rpm for 8 minutes (Z380 centrifuge, Hermle). The supernatant was removed and the pellet resuspended in PBS, then centrifuged again at 1400 rpm for 8 minutes (Z380 centrifuge, Hermle). The supernatant was removed and the pellet was resuspended in an appropriate volume of 75 mM KCl, 0.8% Na-citrate, H_2Obd (1:1:1) hypotonic solution to obtain a final concentration of $7x10^4$ cells/ml. The treatment with the hypotonic solution was carried out at 37 ° C for 15 minutes. Slides were cytocentrifuged at 1500 rpm for 4 minutes (Z300 cytocentrifuge, Hermle), then placed vertically and air dried. Slides were treated with a lysis buffer (2.5 mM TRIS pH 8, 500 mM NaCl, 0.2 M urea, 1% Triton X-100) for 20 minutes. The slides were then removed from the solution with a constant speed of 300 µm/sec using an apparatus equipped with an electric pulley**.** The constant speed allows the unidirectional distension of the fibers.

3.5 - Genomic DNA extraction

Whole high molecular weight genomic DNA was extracted from fibroblast cells in culture. Phenol-chloroform extractions were used to purify DNA. Cells were collected in a test tube and centrifuged at 1200 rpm (Z380 centrifuge, Hermle) for 10 minutes. The pellet was resuspended in EDTA (10 mM, pH 7.9) and proteinase K (100 µg/ml) and SDS (sodium dodecyl sulfate, 0.5%) were added; the solutions incubated at 37°C overnight. NaCl (0.15 M) additoned and an equal volume of a solution of phenol:chloroform:isoamyl alcohol (25:24:1) were added. The test tubes were then centrifuged at 6000 rpm (MIKRO120 microfuge, Hettich Sentrifugen) for 10 minutes. The upper phase was gently transferred to another tube. The phenol/chloroform extraction was repeated twice. Then, a chloroform:isoamyl alcohol (24:1) extraction was performed, using an equal volume of that of the upper phase recovered in the previous step. Sodium acetate (0.3 M) and 2,5 volumes of absolute ethanol were additioned to promote DNA precipitation. The DNA was recovered by a hook, left to air dry and resuspended in 10 mM Tris, 10 mM EDTA. RNase (20 µg/ml) was added and the solution was incubated at 37°C for 4 hours. The treatment with proteinase K was repeated, as well as the phenol/chloroform extraction and the ethanol precipitation, as previously described. DNA was then rehydrated in 75 % ethanol and resuspended in an appropriate volume of double-distilled and sterile H_2O to optimize the storage at -20°C. The concentration and the degree of purity of the extracted DNA was analyzed with a spectrophotometer and finally the DNA was analyzed by electrophoresis on 0.3 % agarose gel to evaluate the molecular weight (more than 48 kb).

3.6 - Combed DNA preparation

The slides were silanized, by immersion in a solution of 2 % of 3-Aminopropyltriethoxysilane (APTES) in acetone for 40 seconds, then washed in acetone 3 times for 2 minutes. Purified high molecular weight genomic DNA was resuspended at a concentration of 2 µg/ml in a solution of 2-morpholinoethanesulfonic acid (MES) (150 mM pH 5.5) and transferred in a reservoir. Silanized slides were introduced into the reservoir and incubated in the solution for 5 min. Then each slide was vertically raised from the solution, at a constant speed (~ 300 µm / sec) using an apparatus, equipped with a pulley driven by an electric motor. The force generated by the meniscus of the solution and the gravity force promote the stretching of the DNA molecules on the slide. To promote a greater adhesion of the DNA molecules to the surface of the slides, combed DNA slides were dried at 60°C for 12 hours.

3.7 - Interphase nuclei preparation

The cells were seeded on slides and cultured in Dulbecco's modified Eagle's medium (Euroclone), supplemented with 20 % foetal bovine serum (Euroclone), 2 mM glutamine, 2 % non-essential amino acids, 1 % penicillin/streptomycin. After a pre-culture period of 48 hours, a treatment period of 18 hours was performed. Griseofulvin (10 µg/ml – Sigma) or nocodazole (200 mM – Sigma) were added. Cells were maintained at 37°C in a humidified atmosphere of 5% CO_2. Then, slides were treated with 75 mM KCl hypotonic solution and incubated at 37°C for 15 minutes. Cold (-20°C) fixative (acetic acid:methanol - 1:3) was added for 30 minutes. The fixation was repeated twice. A recovery period was also performed. After the treatment period, cells were washed with fresh, drug-free, medium and grown for an 18 hours recovery period. Then, hypotonic treatment and fixation were performed, as described above.

3.8 - Micronuclei preparation

The cells were seeded on slides and cultured in Dulbecco's modified Eagle's medium (Euroclone), supplemented with 20 % foetal bovine serum (Euroclone), 2 mM glutamine, 2 % non-essential amino acids, 1 % penicillin/streptomycin. After a pre-culture period of 48 hours, a treated period of 18 hours was performed. Cytochalasin (5 µg/ml – Sigma) and griseofulvin (10 µg/ml – Sigma) or nocodazole (200 mM – Sigma) were added for 18 hours. Cells were maintained at 37°C in a humidified atmosphere of 5% CO2. Then, slides were treated with 75 mM KCl hypotonic solution then incubated at 37°C for 15 minutes. Cold fixative (acetic acid:methanol - 1:3) was added and centrifuged at 1200 rpm (Z380 centrifuge, Hermle) for 30 minutes. The fixation was repeated twice.

3.9 - DNA probes

Lambda phage 37cen and 2PI DNA clones previously isolated from a horse genomic library in lambda phage (Anglana *et al.*, 1996) were used. pSval_137sat was identified and cloned in the laboratory of Molecular and Cellular Biology of University of Pavia (E. Giulotto).

- 37cen - 221 bp repeat (Accession number: AY029358)
- 2PI - consisting of a 23 bp repeat (Accession numbers: AY029359S1 and AY029359S2)
- pSval_137sat (Accession numbers: JX026961)

Bacterial artificial chromosome (BAC) derived from horse CHORI-241 BAC library were used (Leeb *et al.*, 2006). Their cytogenetic position was validated by fluorescent in situ hybridization (FISH) on metaphase chromosomes.

- CHORI241-402C18 (chr8: 41,977,313 - 42,179,897)
- CHORI241-69K23 (chr11: 27,462,459 - 27,665,182)
- CHORI241-230N11 (chr11: 27,672,994 - 27,826,423)
- CHORI241-33J10 (chr11: 27,532,226 -, 27,754,520)
- CHORI241-389H6 (chr11: 27,430,438 – 27,600,382)
- CHORI241-6F13 (chr11: 27,603,018 – 27,797,999)
- CHORI241-21D14 (chr11: 27,639,936 – 27,829,952)
- CHORI241-316B3 (chr11: 27,868,099 - 278,069,904)
- CHORI241-22C1 (chr13: 7,346,775 - 7,544,907)
- CHORI241-377E16 (chr14: 29,599,697 - 29,806,985)
- CHORI241-428I12 (chr28: 12,869,899 - 13,052,689)

3.9.1 - Plasmid DNA purification

Satellite DNA containing recombinant plasmids was extracted from 10 ml of bacterial cultures. Bacterial clones were plated with 20 ml of LB medium (agar 1.8 % and ampicillin 100 µg/ml) and incubated at 37°C overnight. An isolated colony was taken from each plate and placed in 5 ml of liquid culture (LB medium and ampicillin 100 µg/ml). The tubes were then incubated under constant stirring at 37°C overnight. Aliquots were centrifuged at 13000 rpm (Mini-spin microcentrifuge, Eppendorf) for 1 minute. The supernatant was removed and the pellet was resuspended in an appropriate volume of GTE buffer (50 mM glucose, 25 mM Tris-HCl pH 8, 10 mM EDTA pH 8). Then were added 0.2 N NaOH, 1% SDS and the test tubes were incubated for 5 minutes on ice. The lysis reaction was stopped adding potassium acetate 3 M pH 4.8 and incubating the solution on ice for 5 minutes. The tubes were then centrifuged at 13000 rpm (Mini-spin microcentrifuge, Eppendorf) for 10 minutes and the supernatant was transferred to a new tube. RNase (20 µg/ml) was added to eliminate the RNA and the solution was incubated at 37°C for 20 minutes. Then, a volume of chloroform equivalent to that already contained in the tubes were added and centrifuged at 13000 rpm (Mini-spin microcentrifuge, Eppendorf) for 1 minute. The DNA remains in the upper phase was collected and gently transferred to another tube. The chloroform extraction was repeated twice.

An equal volume of isopropanol was added. The tubes were then centrifuged at 13000 rpm (Mini-spin microcentrifuge, Eppendorf) for 10 minutes. Finally, the pellet was washed with 70 % Et-OH and resuspended in an appropriate volume of water.

3.9.2 - BAC DNA purification

Bacterial clones were plated with 20 ml of LB medium (agar 1.8% and chloramphenicol 12.5 µg/ml) and incubated at 37°C overnight. An isolated colony was taken from each plate and placed in 100 ml of liquid culture (LB medium and chloramphenicol 100 µg/ml). The extraction was carried out with Qiagen Plasmid purification kit®, according to supplier instructions.

3.9.3 - Labeling and precipitation of probes

Probes were labeled by nick translation with Cy3-dUTP (Perkin Elmer), Alexa488-dUTP (Invitrogen), Cy5-dUTP (Perkin Elmer), digoxigenin-11-dUTP or biotin-16-dUTP (Roche). The nick translation reaction was performed at 15°C for 90 minutes (plasmids) or for 180 minutes (BACs) and then, blocked with 0.5 mM EDTA. Precipitation of the probes was obtained by adding 7.5 M ammonium acetate and absolute ethanol to the solution. Probes were then resuspended to a final concentration of 30 ng/µl or 20 ng/µl, depending on the type of hybridization, in a hybridization solution (50 % formaldehyde, 10 % dextran sulphate, 1X Denhart solution, 0.1 % SDS, 40 mM Na_2HPO4 pH 6.8 in 2XSSC).

3.10 - FISH – Fluorescence *In Sit*u Hybridization

3.10.1 - Slide aging

Slides were aged at 90°C for 1 hour and 30 minutes. Then they were treated at 37°C for 30 minutes in a solution consisting of 0.005% pepsin/0.01 M HCl. Slides were washed for 3 times of 5 minutes at room temperature using PBS, a buffer consisting of PBS, 1M MgCl2, and a buffer of PBS, 1M MgCl2, 4% paraformaldehyde. Then slides were dehydrated for 5 minutes in the ethanol series (70 % EtOH; 90 % EtOH; and 100 % EtOH).

3.10.2 - *In situ* hybridization on metaphase and stretched chromosomes, interphase nuclei and micronuclei

Slides were denaturated at 72°C for 4 minutes. Then the probes previously denaturated for 8 minutes at 80°C, were put on the slides and the slides were incubated at 37°C overnight in a moist chamber.

3.10.3 - *In situ* hybridization on combed-DNA fibers

Biotinylated (BIO) and digoxigenin-labeled (DIG) probes were denatured at 80°C for 8 minutes. The slides were denatured with a solution of 70% formamide 2XSSC at 72°C for 2 minutes. The probes were then placed on the slides with "combed" DNA. The hybridization reaction was conducted at 37°C for 12 hours in a moist chamber.

3.10.4 - Post hybridization washes and probes detection

Post-hybridization washes were performed in 50 % formamide, 2XSSC at 42°C. Then the slides were washed three times in 4XSSC, 0.1% Tween20 at 42°C temperature for 5 minutes. Slides hybridized with probes in which fluorescent dye molecules are directly bound to the dUTPs (Cy3-dUTP, Alexa488-dUTP, Cy5-dUTP) are counterstained with DAPI (4 ', 6'-Diamidino-2-phenylindole hydrochloride) (1 µg/ml) and then mounted with DAKO. Slides hybridized with BIO and DIG-labelled probes were treated with 100 µl of "blocking solution" (4XSSC, 0.1 % Tween20, 3 % BSA) and incubated at 37°C for 30 minutes in a moist chamber. DIG and BIO -labelled probes were detected with rhodamine and FITC respectively, using five successive layers of antibodies as follows: (i) anti-DIG-Rhod (sheep) 1/200 (Roche) (ii) anti-sheep Rhod 1/100 (CHEMICON) (iii) ExtrAvidin-FITC 1/100 (Sigma-Aldrich) (iv) biotinylated anti-avidin 1/200 (Sigma-Aldrich) (v) ExtrAvidin-FITC 1/100(Sigma-Aldrich). All antibodies were incubated for 30 min at 37°C, then washed for three times of 5 minutes using 4XSSC, 0.1 % Tween20 at 42°C after each step of antibody detection. After the last antibody, the slides were washed two times for 5 minutes using 4XSSC, 0.1% Tween20 at 42°C. An additional wash in 4XSSC at room temperature were performed. The metaphase and the stretched chromosomes were stained with DAPI (4 ', 6'-Diamidino-2-phenylindole hydrochloride) (1 µg/ml) and then mounted with DAKO (with the only exception of "combed" DNA which is not counter-stained).

3.11 - Immuno-FISH

For the slide preparation see previous paragraphs 3.2-3.6. For fixation, the slides were treated with a solution of 4% paraformaldehyde in KCM (0,12 M KCl, 0,08 mM NaCl, 0,01 M Tris-HCl pH 8, 0.5 mM EDTA, 0.1% Triton) for 10 minutes. Finally, the slides were treated with KCM for 5 minutes.

3.11.1 - Immunofluorescence

The slides were treated with the primary antibody (CREST B2 SERUM; CREST B5 SERUM; anti-CENP-B antibody (abcam84489,); anti-H3K4me2 antibody (abcam32356); anti-H3K9me2 antibody (abcam1220); anti-H3K9me3 antibody (abcam8898) and anti-H3K27me3 antibody (abcam6002)). The slides were subsequently incubated at 4°C for 24 hours in moist chamber. Slides were then washed for three times in KB$^-$ (0,01 TRIS-HC l pH 8, 0.15 M NaCl, 0.5% BSA) of 5 minutes each. The slides were subsequently treated with 35µl of secondary antibody (anti-human conjugated with AlexaFluor-488 (Invitrogen); anti-rabbit conjugated with Alexa488 or with Cy2 (Jackson ImmunoResearch); anti-sheep conjugated with Alexa488 or with rhodamine (Jackson ImmunoResearch); anti-mouse conjugated with TexasRed (Jackson ImmunoResearch). The slides were incubated at 37°C for 1 hour in moist chamber and then washed twice in KB$^-$ for 5 minutes. Slides were treated with 4% paraformaldehyde in KCM for 7 minutes, followed by 2 washes in distilled water of 3 minutes each. If after the immunofluorescence an *In Situ* Fluorescence Hybridization were performed, the slides were immersed in cold methanol-acetic acid (3:1) for 15 minutes and washed in 2XSSC at room temperature for 2 minutes. Otherwise, the slides were mounted with 30µl of a solution of DAKO-DAPI (DAKO: 5 µl/ml).

3.11.2 - *In situ* hybridization on extended chromatin fibers

The slides were dehydrated in 3 steps, each of 3 minutes, in the ethanol series (EtOH 75%, 95%, 100%). The slides were then air-dried. Subsequently, the chromatin fibers on the slides were denatured with a solution of 70% formamide in 2XSSC at 80°C for 4 minutes; the probes were instead denatured at 80°C for 8 minutes. After this step, the slides were immersed in a solution of 2XSSC at 4°C for 2 minutes and then subjected to the ethanol series as described previously. The slides were treated with denatured probe and the hybridization reaction was conducted at 37°C in a moist chamber for 12 hours. After the hybridization, the slides were washed 3 times for 5 minutes in 2XxSSC at 42°C, then treated with blocking solution (3% BSA, 4XSSC, 0.1% Tween20), and incubated at 37°C for 30 minutes. Post-hybridization washes were performed in 50% formamide, 2XSSC at 42°C.

Finally, the slides were mounted with 30µl of a solution of DAKO-DAPI (DAKO: 5 µl/ml).

3.12 - Microscopic analysis of the slides

The slides were analyzed with a fluorescence microscope Axioplan (Zeiss) equipped with a cooled charge-coupled device (CCD) camera (Photometrics). The CCD camera is characterized by a photo-sensor made of a matrix of silicon crystals sensitive to light. The photons contact the sensor and are converted into an electric charge proportional to the intensity of the light. After the exposure to light, the electric charge is transferred to the computer and converted into the binary system. The emission of each pixel is converted into a gray scale value which depends on the intensity and wavelength of the incident light. The tones in the scale vary from 256 to 4096.

The images were acquired and pseudo-colored with the IPLab spectrum software (Digital Pixel Advanced Imaging System, Brighton) and then processed using the software Photoshop ®.

4. RESULTS AND DISCUSSION

4.1 - Architectural organization of horse satellite-based centromeres

In a previous work, the distribution of the two major horse satellite DNA families, 37cen and 2PI, was investigated (Piras *et al.*, 2010). Through FISH experiments on metaphase chromosomes from the horse (*Equus caballus*), the donkey (*Equus asinus*), the Grevyi's zebra (*Equus grevyi*) and the Burchelli's zebra (*Equus burchelli*), a complex arrangement of satellite DNA sequences distribution was observed. At the FISH resolution level, several centromeres were found to be devoid of satellite DNA: one centromere in the horse (on chromosome 11), eighteen centromeres in the donkey, seventeen in the Grevyi's zebra and seven in the Burchelli's zebra. These results demonstrated that the centromere function is uncoupled to the satellite DNA. Moreover, the presence of satellite repeats at non centromeric termini, presumably corresponding to relics of ancestral centromeres. was observed.

To verify if other repetitive DNA, belonging to new satellite DNA families, was present at the centromeres lacking 37cen and/or 2PI signals, a FISH analysis was carried out in the four species, using their total genomic DNA as probe (Piras *et al.*, 2010). In the horse, all the centromeres, with the only exception of chromosome 11, were labelled; an interstitial signal was also localized on the long arm of the X chromosome. In the domestic donkey, in the Grevy's zebra and in the Burchelli's zebra, few additional sites of hybridization were detected. These extra hybridization sites were localized in several centromeric regions, in few telomeric regions and on the long arm of donkey and Burchelli's zebra X chromosomes (Piras *et al.*, 2010). These results suggested the presence, in these species, of tandem repeats, other than 37cen and 2PI.

In the laboratory of Molecular and Cellular Biology, a new horse satellite sequence, EC137, was identified from the horse genome database (EquCab2.0,www.ncbi.nlm.nih.gov/projects/mapview/map_search.cgi?taxid=9796), using *in silico* analysis tools (BLAST, BLAT, Tandem Repeat Finder, RepeatMasker, MultAlin) and then cloned it into a plasmid vector.

4.1.1 - Chromosomal distribution of EC137 in four *Equus* species

The cloned EC137 satellite DNA was used as a probe for FISH; its distribution was analyzed on metaphase chromosomes from *E. caballus* (**Figure 11a**), *E. asinus* (**Figure 11b**), *E. grevyi* (**Figure 11c**) and *E. burchelli* (**Figure 11d**).

In the four species analyzed, the EC137 FISH signals did not coincide with those of the two major equid satellite DNA families (37cen and 2PI).

In the horse, the centromeric region of chromosomes 1, 7, 8, 10, 14, 15, 20, 23, 26, 27, 28, 29 and 30 was labelled (**Figure 11a**). Horse chromosome 11, whose centromere was previously demonstrated to be void of the 37cen and 2PI satellite DNA families, was also negative to EC137 hybridization.

A *in silico* analysis highlighted the presence of the EC137 sequence at the centromere of horse chromosomes 1, 2, 15, 20, 25 and 28 and on the long arm of horse chromosome X. FISH analysis did not mark the centromeric region of horse chromosomes 2 and 25 nor the long arm of the X chromosome. This apparent contradiction is due to two opposite factors. The first one is that the 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 that lack chromosome assignment. On the other hand, FISH experiments allow the direct localization of the repetitive sequences on the chromosomes, including those allocated in the unplaced chromosome in the horse genome database.

The inability to mark by FISH some centromeres, which resulted positive *in silico* analysis, is presumably due to the fact stretches of repetitive sequence were under the resolution limit of FISH, this was the case of horse chromosomes 2 and 25 and of the long arm of the X chromosome.

In the donkey (**Figure 11b**), only two chromosomes were FISH labeled by the EC137 sequence, EAS1 and EAS2. This means that all the other donkey centromeres lacking 37cen and 2PI satellites, are also devoid of EC137; whereby these centromeres are, bona fide, new examples of satellite-less centromeres. An interesting aspect is that the EC137 fluorescence signal on donkey chromosome 1 was in an interstitial position on the short arm. In a previous paper the distribution of the two major equine satellite DNA families (37cen and 2PI) on donkey chromosome 1 was investigated in detail (Raimondi *et al.*, 2011). On the short arm of donkey chromosome 1, three hybridization sites were found and no one coincided with the position of the EC137 signal.

In the Grevy’s zebra, FISH signals were detected in the centromeric region of chromosomes 10 and 12 (**Figure 11c**). Among the seventeen Grevy’s zebra centromeres which were negative for 37cen and 2PI, only the centromere of chromosome 10 was positive to EC137. Again, this result suggests that the chromosomes negative for 37cen, 2PI and 37cen, are immature satellite-less centromeres.

In Burchelli's zebra, five centromeres (7, 9, 10, 13, 14) were labelled (**Figure 11d**). In all the analyzed metaphase spreads (total number 25), only one homologous of both chromosomes 14 was FISH positive and only one homologous of both chromosomes 11 had the telomeric end labelled, suggesting a polymorphic variation in the number of EC137 repeats. The polymorphic nature of the FISH signals is not surprising since the intra- and interspecific variability in the amount and distribution of satellite DNA sequences is well documented (Plohl *et al.*, 2008).

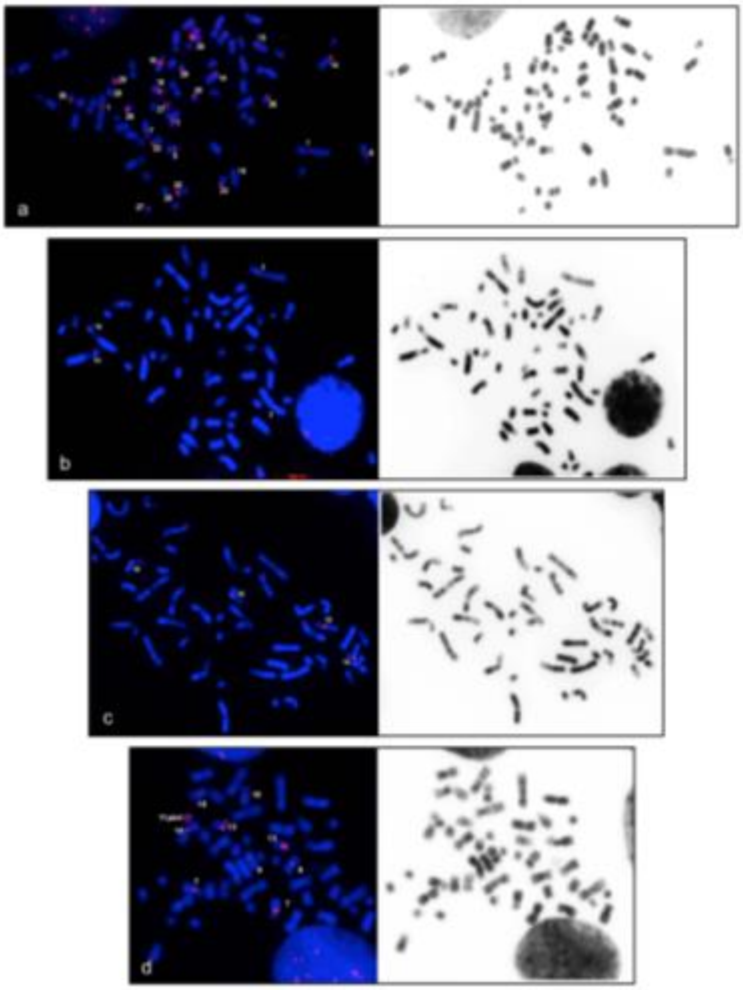

Figure 11 – Localization of the EC137 satellite (red) by FISH on metaphase chromosomes from the horse (a), the domestic donkey (b), the Grevyi's zebra (c) and the Burchelli's zebra (d) (Nergadze *et al.*, 2014).

4.1.2 - High resolution analysis of horse satellites organization

To investigate the physical relations between the different classes of equid satellite DNA sequences, two color FISH experiments on horse metaphase chromosomes were carried out (**Figure 12**).

The results of the co-hybridization of the 37cen and EC137 satellites are shown in **Figure 12a**. The 37cen satellite is present on all the centromeres except those of chromosomes 2 and 11, while the EC137 satellite is located only on 26 out of 64 centromeres. In addition, the two satellite DNA sequences appear to be in a different position: the 37cen signal always coincides with the primary constriction, on the contrary, the EC137 signal is mostly pericentromeric, with no or limited overlap with 37cen (arrows in **Figure 12a**).

The results of the co-hybridization of the 2PI and EC137 satellites are shown in **Figure 12b**. The majority of the chromosomes that share the 2PI and the EC137 sequences are yellow labeled. The yellow signal is due to the overlap of the green (2PI) and red (EC137) fluorescence signals arising from the single probes. This data, demonstrated that, in the horse, twelve chromosomes carried all three satellite DNA sequences (7, 8, 10, 14, 15, 20, 23, 26, 27, 28, 29, 30). By analyzing the positions of the three classes of satellite DNA, it was possible to conclude that, in the horse, the 37cen satellite may be the functional centromeric satellite, since the 37cen signals always coincide with the primary constriction. Instead, the 2PI and the EC137 sequences may represent accessory pericentromeric elements, since their signals are mostly pericentromeric compared to the 37cen ones, except for the horse chromosome 2. At this centromere, the 2PI sequence is the only satellite observed by FISH. This suggests that, at least in this case, 2PI might be able to drive kinetochore assembly.

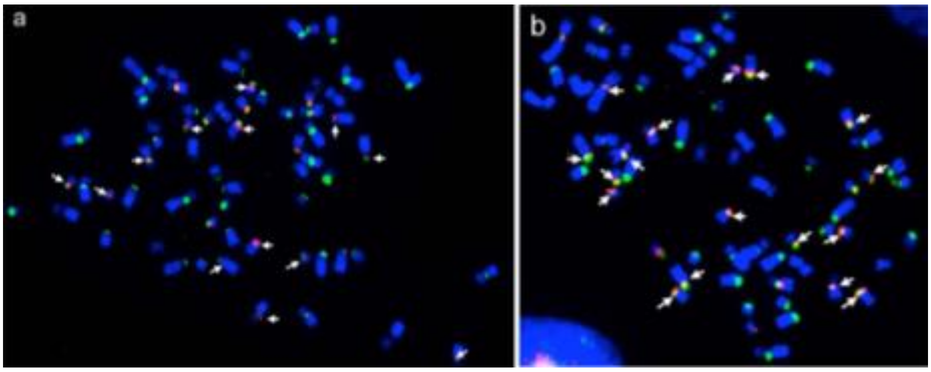

Figure 12 – Two color FISH with EC137 (red) and 37cen (green) satellites (a) and EC137 (red) and 2PI (green) satellites (b) on horse metaphase chromosomes (Nergadze *et al.*, 2014).

To define, at a higher resolution level, the physical relationships among the different satellite DNA families, three color FISH experiments on mechanically stretched horse chromosomes were performed (**Figure 13**). A total number of 89 stretched chromosomes were analyzed but only on 37 chromosomes the three satellite DNA families were present together.

Among these 37 chromosomes, we observed five patterns of physical organization of the satellite sequences (**Figure 13**). In 17 out of 37 chromosomes (46%), the 37cen sequence covered the whole primary constriction while 2PI and the EC137 satellites co-localized in the distal portion of the 37cen positive region (**Figure 13a**). In 7 out of 37 chromosomes (19%), the 37cen sequence again covered the whole primary constriction, while the 2PI sequence was spread along the 37cen positive region and the EC137 satellite was underrepresented and localized in different positions within the 37cen and 2PI positive region (**Figure 13b**). Concerning these centromeres, we can claim that 37cen sequence plays a role in centromere function while 2PI and EC137 may represent accessory elements.

In 7 out of 37 cases (19%), 37cen and the 2PI sequences were both very abundant, while the EC137 sequence was extremely scanty and interspersed within the other satellites (**Figure 13c**). In 4 out of 37 chromosomes (11%), the 2PI sequence spread out in an uncoiled pericentromeric region which was 37cen and EC137 negative. All chromosomes that displayed this type of arrangement were metacentric or sub-metacentric (**Figure 13d**). In 2 out of 37 chromosomes (5%), which were acrocentric, centric chromatin formed uncoiled extensions protruding out of the main chromosome body; these protruding fibers were 2PI positive and 37cen and EC137 negative (**Figure 13e**).

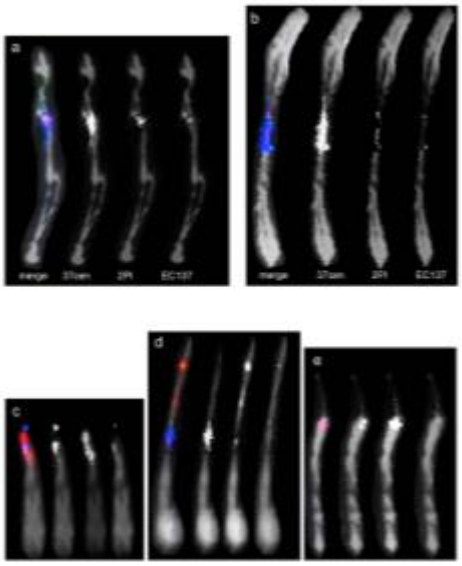

Figure 13 – Three color FISH on horse mechanically stretched chromosomes. In each panel, the merged image is shown on the left; in the other images, the separate color channels are reported, corresponding to the 137cen probe (blue in the left image), to the 2PI probe (red in the left image) and to the EC137 probe (green in the left image), respectively (Nergadze *et al.*, 2014).

To define, at the single molecule level, the physical relations between the three satellite DNA families dual color FISH on horse combed DNA fibers was performed (**Figure 14**). This technique allows to obtain DNA fibers with a uniform degree of elongation, thereafter quantitative estimates can be performed.

To determine the average degree of DNA fiber extension, a molecular ruler was set up. Using a horse BAC clone of known length as a probe, parallel FISH experiments on combed DNA were performed. In this way, it was possible to relate the length of the hybridization signals, measured in centimeters on digital images, with the corresponding length in base pairs of the target sequence.

In the merged images shown in **Figures 14a** and Figure **14c**, the 37cen satellite covers a long continuous region, extending for hundred kilobases (occupying more than one microscope field). When the 2PI (**Figure 14a**) or EC137 (**Figure 14c**) were present, these satellites formed small stretches (2-8 kb) that were strictly interspersed within the 37cen clusters. These regions appear as yellow fluorescence hybridization signals (overlapping green and red fluorescence signals), indicating a high interspersion of satellite sequences on a small scale. The 2PI or EC137 stretches occurred every 6-80 kb into the 37cen blocks. In **Figure 14b**, the results of two color FISH on combed DNA hybridized with 2PI and EC137 is shown. These two satellite DNA sequences are both organized in small stretches (2-8 kb) which are strictly intermingled; the 2PI satellite appears to be more abundant than EC137.

The overall organization of the different classes of horse satellite DNA appears to be a mosaic where the three DNA families display an interspersed association of sequence blocks widely variable in size. This organizational pattern of DNA sequences in heterochromatin might be common in genomes, such as the *Equus* species ones, characterized by a high rate of inter-chromosomal exchange (Zinic *et al.*, 2000).

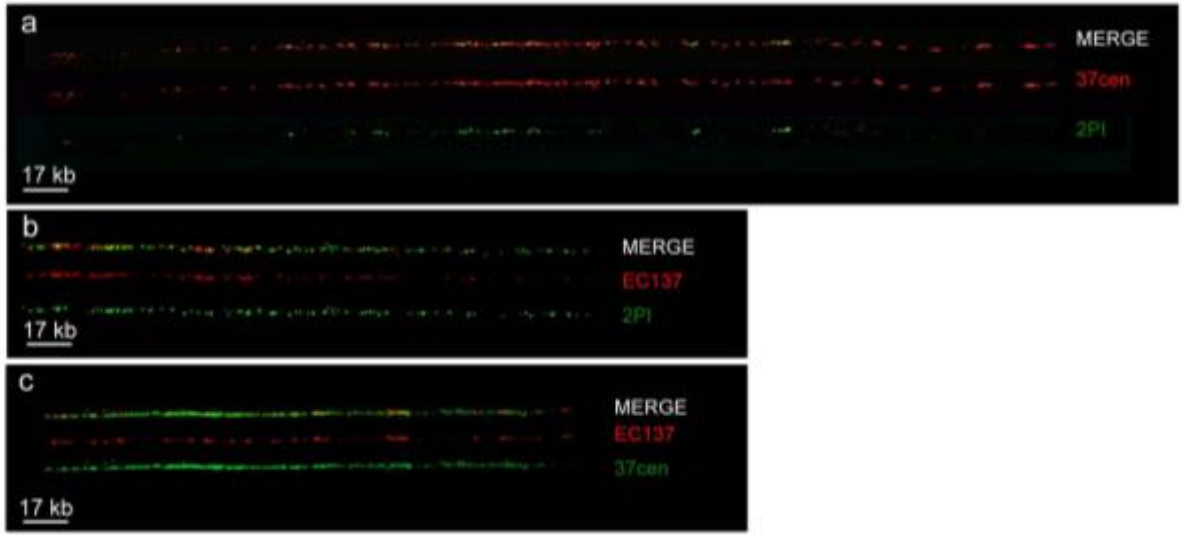

Figure 14 – Two color FISH on horse combed DNA fibers. (**a**) 37cen (red) and 2PI (green) satellites (**b**) EC137 (red) and 2PI (green) satellites (**c**) EC137 (red) and 37cen (green) satellites (Nergadze *et al.*, 2014)

4.2 - Functional organization analysis of satellite-based centromeres in the horse genome

Although the centromeric function is highly conserved through eukaryotes, centromeric satellite DNA is rapidly evolving, often being species specific (Melters *et al.*, 2013; Plohl *et al.*, 2014).

Following our initial description of a centromere completely devoid of satellite DNA in the horse (Wade *et al.*, 2009), other examples of naturally occurring satellite-less centromeres were observed in plants and animals (Piras *et al.*, 2010; Gong *et al.*, 2012). These observations raised 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 the binding of the centromere specific histone CENP-A (the main epigenetic mark of centromere function) to centromeric chromatin (Steiner and Henikoff, 2015). As mentioned in the introduction, in several species, centromeric satellite DNA is transcribed. Transcriptional competence of the centromeric regions seems to be important for chromatin opening and CENP-A loading; centromeric transcripts are believed to provide a flexible scaffold that allows the assembly of the kinetochore proteins. It has also been hypothesized that these transcripts could act *in trans* on all, or on a subset of chromosomes, independently of the primary DNA sequence (Rošić *et al.*, 2014; Biscotti *et al.*, 2015; Rošić and Erhardt, 2016).

To identify the satellite repeat driving the centromeric function in satellite based horse centromeres, we used a high-resolution cytogenetic approach.

4.2.1 - High resolution cytogenetic analysis of the functional organization of horse satellite-based centromeres

Our previous FISH analyses on stretched chromosomes and combed DNA fibers demonstrated that horse centromeric and pericentromeric regions display a mosaic arrangement of different satellite DNA families (Nergadze *et al.*, 2014).

To analyze the physical organization of the horse centromeric functional domains, were carried out immuno-FISH experiments on mechanically stretched chromosomes using 37cen as a FISH probe and a previously tested CREST serum (Purgato *et al.*, 2015) to mark the centromeric domains (**Figure 15**). Ninety-nine stretched chromosomes (46 meta/submeta-centric and 53 acrocentric) were examined. The abundance of the 37cen sequence was variable among chromosomes, extending in some instances over a large pericentromeric region (white arrows) or being apparently confined to the primary constriction. The CREST signals always colocalized with the 37cen fluorescence, however, no clear correlation seemed to exist between the 37cen and the CREST signals, nor in intensity nor in length. These data prove that the GC rich 37cen sequence is associated with the centromeric function. 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 (Blower *et al.*, 2002).

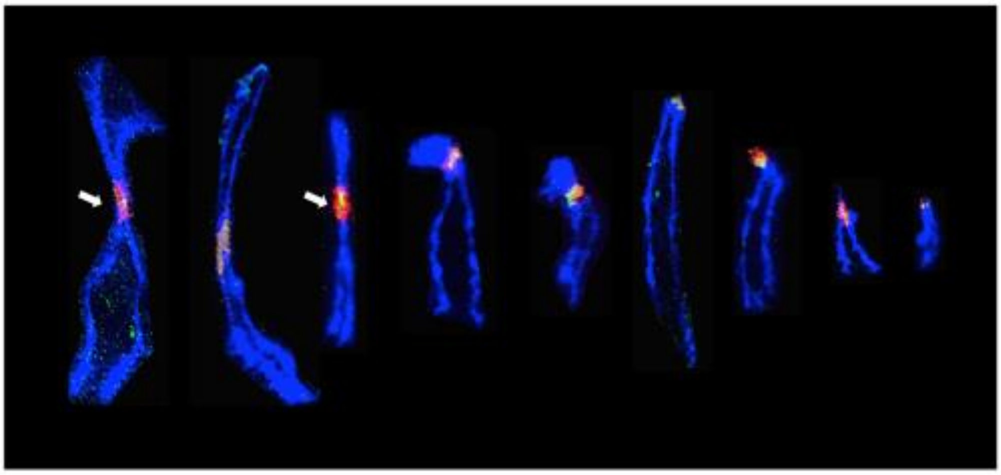

Figure 15 - Immuno-FISH on mechanically stretched chromosomes. 37cen is red labelled while CENP-A, detected by an anti-CENP-A enriched CREST serum, is green labelled. A total number of 99 stretched chromosomes was analyzed. A sample of representative images is reported in the figure (Cerutti *et al.*, 2016).

To more accurately define the relationship between the 37cen satellite and the centromeric function, was performed a higher-resolution immuno-FISH analysis on horse chromatin fibers. On a total number of 25 extended fibers, different arrangements of CENP-A domains were observed (**Figure 16**). Sixty percent of the fibers (15/25) showed CENP-A binding over the whole length of the 37cen positive region (**Figure 16I**), in 28% (7/25) of the cases (**Figure 16II**) CENP-A domains appeared as blocks of variable length intermingled into the 37cen stretches. This discontinuous presence of CENP-A at horse centromeres resembles the chromatin organization observed using the same high resolution morphological approach, in human and in *Drosophila* (Blower *et al.*, 2002).

A 37cen FISH signal with no overlap with or flanking the CENP-A signal was observed in 12% of the fibers (3/25) (**Figure 16III**). These fibers presumably derive from pericentromeric locations that contain the 37cen satellite.

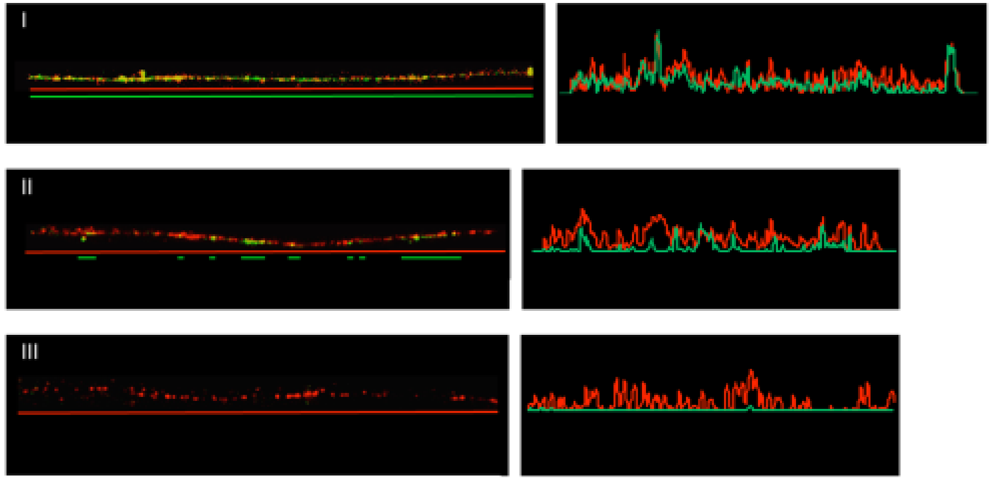

Figure 16 – Immuno-FISH on extended chromatin fibers. The 37cen satellite DNA is labelled in red. CENP-A, identified with an anti-CENP-A enriched CREST serum is green labelled. In each panel, under the microscope image of the fiber, is sketched the CENP-A binding pattern observed. The images on the right show line graphs quantifying the fluorescence staining along the length of each fiber. (**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 fiber with no CENP-A binding is reported (Cerutti *et al.*, 2016)

The primary constriction of mammalian chromosomes is typically embedded in a constitutive heterochromatic repeated satellite DNA. The horse is peculiar among mammalian species because the centromere of chromosome 11 is completely devoid of satellite DNA (Wade *et al.*, 2009; Piras *et al.*, 2010; Purgato *et al.*, 2015). 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 (Nergadze *et al.*, 2014) but only the GC rich 37cen sequence is associated with the centromeric function and is also transcriptionally active (Cerutti *et al.*, 2016). The significance of satellite DNA at centromeres has so far been elusive because satellite-less centromeres are perfectly functional (Purgato *et al.*, 2015). 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.

4.3 - Extensive analysis of the functional centromeric domains of the first natural satellite-free centromere described in the literature (ECA11)

As detailed in the introduction, the centromere of horse chromosome 11 is devoid of any repeated sequence. During the horse genome sequencing, the analysis, by ChIP-on-chip, of the primary constriction of ECA11 revealed two regions (136 kb and 99 kb) bound by CENP-A (Wade *et al.*, 2009). The unexpected observation of two CENP-A binding domains in the centromere of horse chromosome 11 (Wade *et al.*, 2009) prompted us to extend the analysis to new horse individuals.

ChIP-on-chip experiments were performed on fibroblast cell lines from five unrelated horses in the Laboratory of Molecular and Cellular Biology of Pavia (Prof. Elena Giulotto) and in the Laboratory of Functional Genomics and Epigenetics of Bologna (Prof. Giuliano della Valle).

To define the position and the number of CENP-A binding domains, the immunoprecipitated chromatin was labeled and hybridized on an array that contained a region of about 2 Mb corresponding to the centromeric region of horse chromosome 11 (ECA11: 25,566,599-28,305,611). This approach enabled to determine, for each sample, the location of the CENP-A binding domain. The length of each binding domain ranged from 78 to 212 kb in a region covering about 535 kb (ECA11: 27,514,628-28,049,577).

Three individuals clearly showed two distinct and well separate CENP-A binding peaks (HSF-B, -C, -G) while the others showed only one wider peak (HSF-D, -E) (**Figure 17**).

Real-time PCR on the DNA purified from CENP-A immunoprecipitated chromatin confirmed that HSF-B, HSF-C and HSF-G individuals had two regions of CENP-A binding, while a single region was present in HSF-D and HSF-E.

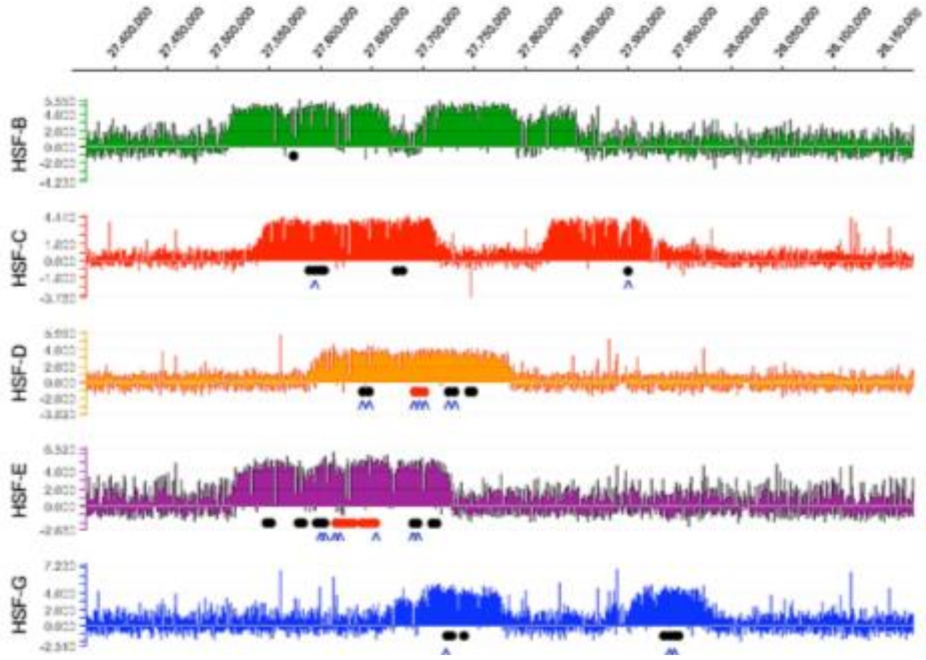

Figure 17 - Variable position of the centromere of horse chromosome 11. DNA obtained by chromatin immunoprecipitation. Using an anti-CENPA antibody, from five different horse fibroblast cultures was hybridized to a tiling array covering the centromere region. Results are presented as the log2 ratio of 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), red dots (both SNP alleles are present in immunoprecipitated DNA) and blue carats (Purgato *et al.*, 2015).

The presence, in some cell lines, of two CENP-A binding domains could reflect, a "multi-domain structure" of the centromere widely described in the literature (**Figure 18a**) (Blower *et al.*, 2002; Alonso *et al.*, 2003; Cleveland *et al.*, 2003; Chueh *et al.*, 2005; Sullivan *et al.*, 2004; Alonso *et al.*, 2010; Schueler *et al.*, 2001). As an example, in a neocentromere derived from human chromosome 10 CENP-A does not bind uniformly the whole centromeric domain (about 330 kb), rather forms discontinuous blocks (Chueh *et al.*, 2005). Moreover, immuno-FISH studies on extended chromatin fibers showed that in man alpha satellite contains discontinuous domains of CENP-A, (15 to 40 kb), interspersed with domains containing the canonical histone H3.

An alternative intriguing interpretation of our results is that the two distinct domains observed in HSF-B, -C, -G could be epialleles, in other words different functional alleles might be present on the same DNA sequence, individuals showing separate peaks being heterozygous for the functional alleles (**Figure 18b**). Thus, a "multi-domain" model reflects the uneven distribution of CENP-A on both homologous, while the epiallele model would be predictive of a different association of CENP-A and on the homologous.

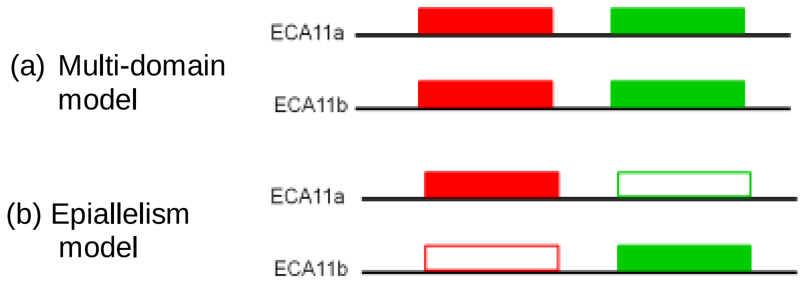

Figure 18 – Schematic representation of "multi domain" (a) and "epiallelism" (b) hypotheses. In the multi-domain model, the centromere function is associated with two sequences (red and green filled rectangles) on both homologous, while in the epiallelism model, the centromere function is related to one sequence on one chromosome (red filled rectangle) and to another sequence on its homologous (green filled rectangle).

4.3.1 - Analysis of ECA11 centromeric domains organization by immuno-FISH on chromatin fibers

To discriminate between the multi-domain model and the epiallele hypothesis, a single molecule analysis of centromeric domains by immuno-FISH on chromatin fibers, was carried out. BACs covering the ECA11 centromeric domain (as determined by ChIP-on-chip) were used as FISH probes and a CREST serum was used to detect the functional centromeric domain.

Samples from HSF-B, HSF-C, HSF-D, HSF-E and HSF-G were analyzed and two different organization patterns of FISH and immuno-staining fluorescent signals were observed. The first individuals analyzed were the horses displaying two clearly separated ChIP-on-chip peaks (HSF-B, HSF-C and HSFG). Two distinct epialleles were distingui zinic d, one of which (epiallele 1 in **Figure 19a**) had the immuno-staining flanking the FISH signal, while in the other one (epiallele 2 in **Figure 19a**), the immuno-staining and FISH signals were superimposed.

Subsequently the horses displaying a single broad ChIP-on-chip peak (HFS-D and HSF-E) were studied. As shown in **Figure 19a**, also in these individuals, two partially overlapping functional alleles were observed. In one epiallele (epiallele 1 in **Figure 19b**), the immuno-staining partially covered the FISH signal and extended in the flanking region, while in the other epiallele (epiallele 2 in **Figure 19b**), the immuno-staining covered the FISH signal. The immuno-labelled regions of epiallele 1 and epiallele 2 were partially overlapping. These results imply that each homologous chromosome 11 has a protein binding region in a defined position, and that the broad peak, found by ChIP-on-chip, is the result of the contribution of the single peaks on each homologous.

In conclusion, at least seven functional epialleles were identified in the five horses and each epiallele occupied about 80–160 kb. These results demonstrate that the centromeric domain of horse chromosome 11 is characterized by positional variation, and that in a native mammalian centromere the centromere position can be flexible across a relatively wide (500kb) single-copy genomic region. Our results definitely demonstrated that the positioning of CENP-A binding domains is unrelated to the underlying DNA sequence.

No functionally homozygous individual was observed; therefore, in spite of the limited sample size, it is possible to infer that this epigenetic *locus* is highly polymorphic. It is possible that the centromere studied here is particularly dynamic because it is evolutionarily young and lacks satellite tandem repeats (Wade *et al.*, 2009; Piras *et al.*, 2010). In our system, the lack of satellite DNA at the centromere of horse chromosome 11 is a stable feature in all individuals of the horse species and was maintained for many generations during evolution; therefore, the mechanism of satellite DNA recruitment and the precise role of repetitive sequences in centromere function and stabilization remain to be established. As mentioned in the introduction, satellite DNA recruitment appears to be a late step in new centromere maturation. Maybe the colonization of a CENPA domain by satellite DNA progressively reduces the positional flexibility of the centromere through a satellite mediated stabilization mechanism.

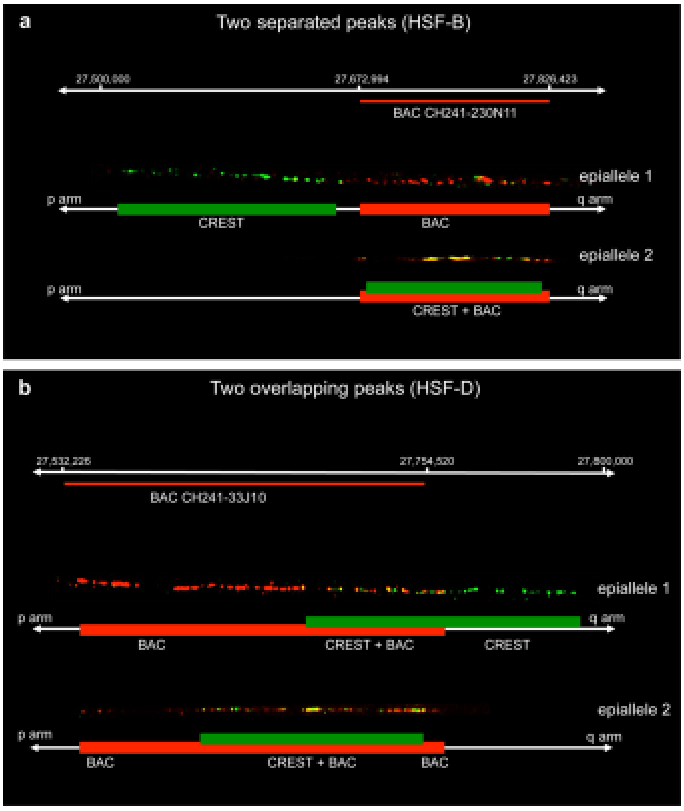

Figure 19 - Single molecule analysis of centromeric epialleles on chromatin fibers by immuno-FISH. (**a**) Organization pattern of functional alleles in horses displaying two separated ChIP-on-chip peaks (HSF-B). (**b**) Pattern of functional alleles organization in horses displaying two overlapping ChIP-on-chip peaks (HSF-D). At the top of each panel are reported the coordinates of the regions occupied by the centromeric domains, and BAC coverage is represented by a red line. CREST immunostaining is green labelled while the BAC FISH signals are red labelled. Under each fiber image, a schematic representation is depicted with green rectangles corresponding to centromeric domains and red rectangles indicating BAC hybridization (Purgato *et al.*, 2015).

4.4 - Analysis of donkey (EAS4, EAS7 and EAS9) satellite-less centromeric domains organization by immuno-FISH on chromatin fibers

Our previous work (Piras *et al.*, 2010) indicated that in the donkey 18 centromeres are devoid of satellite DNA, at the FISH resolution level. To confirm the absence of highly repetitive DNA sequences at some donkey centromeres ChIP-seq experiments on donkey primary skin fibroblasts were carried out in the Laboratory of Molecular and Cellular Biology.

Sixteen donkey chromosomes containing one or two distinct CENP-A binding domains on unique sequence regions homologous were identified (**Figure 20**).

We analyzed, by two color immuno-FISH on chromatin fibers, the centromeric domain of donkey chromosomes 4, 7 and 9. EAS4 and EAS7 show a single broad protein binding domain, while EAS9 exhibited a single spike peak.

Using BACs that covered the EAS4 or EAS7 centromeric domains (as determined by ChIP-seq) as FISH probes and a CREST serum to detect the functional centromeric domain, two-color immuno-FISH was performed. Only one type of arrangement was observed in all the centromeres analyzed (**Figure 21a** and **21b**). In both homologous the immuno signal colocalized with the FISH signal. We can conclude that in these donkey satellite-less centromeres, no functional polymorphism exists and the centromere function is always related to the same sequence.

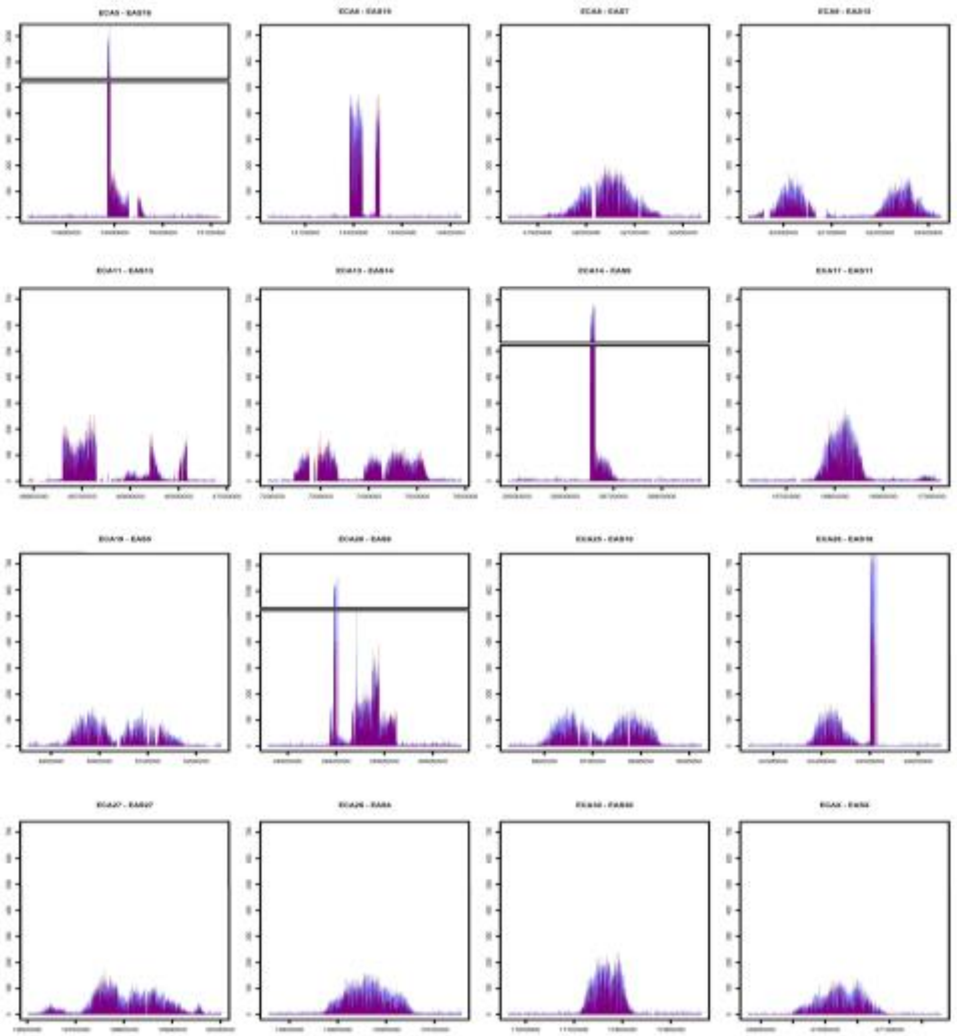

Figure 20 – ChIP seq profile of donkey satellite-less centromeres. DNA was obtained by chromatin immunoprecipitation using an anti-CENPA antibody to identify the centromere functional domains of the donkey satellite-less centromeres.

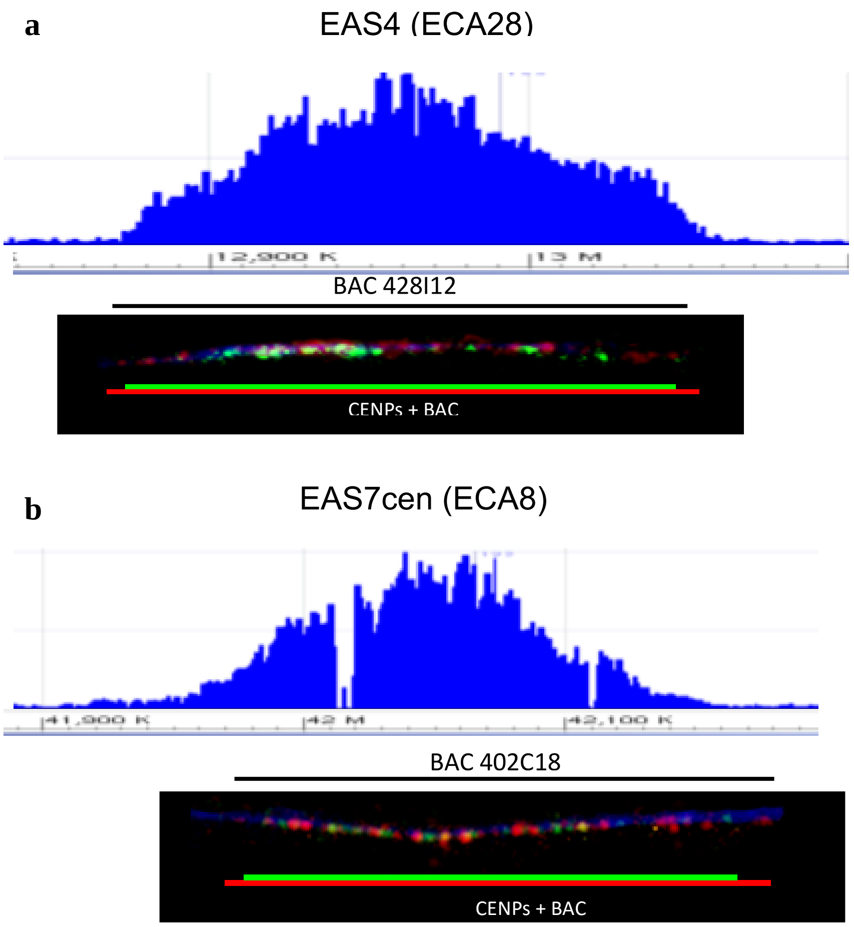

Figure 21 - Single molecule analysis of two donkey centromeric domains on chromatin fibers by immuno-FISH. Organization pattern of the functional centromeric domains of donkey chromosome 4 (**a**) and donkey chromosome 7 (**b**). At the top of each panel the regions occupied by the centromeric domains are reported, and BAC coverage is represented by a black line. CREST immunostaining is green while the BAC FISH signals are red. Under each fiber image, a schematic representation is depicted with green rectangles corresponding to centromeric domains and red rectangles indicating BAC hybridization.

The third donkey centromere analyzed was that of EAS9 whose immunoprecipitation profile showed a tiny and tall peak, completely different from the others (except for EAS8, EAS16 and EAS19). We hypothesized that this spike-like peak might reflected sequence differences between the donkey and the horse genome as the donkey ChIP-seq reads were assembled on the reference horse genome. PCR analysis, carried out in the laboratory of Molecular and Cellular

Biology, indicated that a 10kb long sequence was present in 3 tandem copies in the donkey with respect to the horse orthologous region.

Donkey EAS9 centromere (**Figure 22**) was analyzed by two color immuno-FISH on extended chromatin fibers in the same way of the previous centromeres. Also in this case, only one type of arrangement was observed (**Figure 22**). In both homologous, the immuno signal always colocalized with the BAC signal. Notably, the extension of the CENP-A positive region was comparable to that observed for the centromeres of chromosomes EAS4 and EAS7, thus supporting the hypothesis that the spike peak was actually due to a distortion originated by donkey sequences alignment on the horse reference genome.

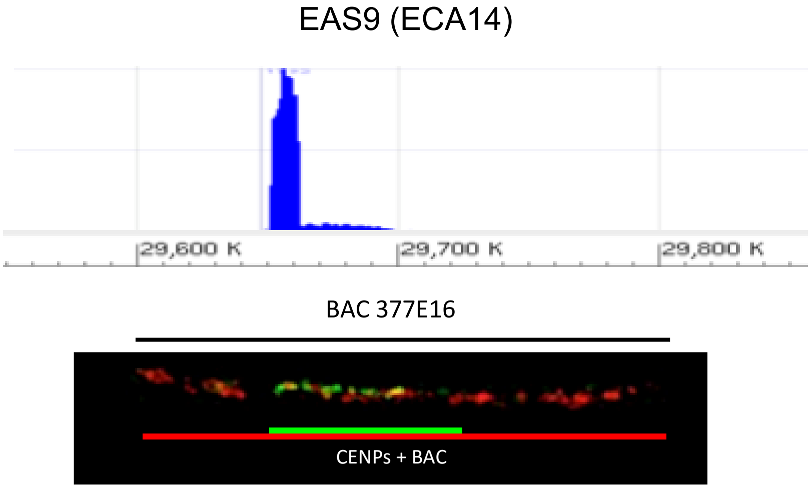

Figure 22 - Single molecule analysis of centromere of donkey chromosome 9 on chromatin fibers by immuno-FISH. At the top of the panel is reported the regions occupied by the centromeric domains. BAC coverage is represented by a black line. CREST immunostaining is green while the BAC FISH signals is red. Under fiber image, a schematic representation is depicted with green rectangles corresponding to centromeric domains and red rectangles indicating BAC hybridization.

4.5 - *In vitro* analysis of the mitotic stability of horse chromosome 11 with a satellite-free centromere

The presence of completely satellite-free and stable natural centromeres opens the question of the functional role played by satellite DNA at the centromere (Marshall *et al.*, 2008; Nakano *et al.*, 2008; Rocchi *et al.*, 2012; Shang *et al.*, 2013).

Concerning the contribution of satellite DNA to chromosome segregation fidelity, some data come from the analysis of pathologic satellite-less centromeres and from human artificial chromosomes.

Pathological neocentromeres are often present as mosaics; this mosaicism might be due to intrinsic mitotic instability (Marshall *et al.*, 2008), however it is more plausible to hypothesize that, since the pathological neocentromeres produce an unbalanced karyotype, neocentromere containing cells are counterselected.

Artificial human chromosomes have been demonstrated to need alphoid DNA for *de novo* centromere formation. It has been suggested that alpha-satellite DNA creates a proper epigenetic environment essential for kinetochore activity (Nakano *et al.*, 2008; Ohzeki *et al.*, 2015). In addition, human artificial centromeres require alpha-satellite arrays with binding sites for the CENP-B protein to be propagated in culture (Masumoto *et al.*, 2004; Henikoff *et al.*, 2015).

To our knowledge, there are no data about the mitotic behavior of natural satellite-free centromeres. To fill this gap, we used as model system the species belonging to the genus *Equus*. In these species, the centromere function and the position of satellite DNA are often uncoupled (Piras *et al.*, 2010). Moreover, satellite-less centromeres, originated by evolutionary centromere repositioning, are unexpectedly frequent; as a consequence, satellite based and satellite-less centromeres coexist in single karyotypes (Wade *et al.*, 2009; Piras *et al.*, 2009; Piras *et al.*, 2010; Raimondi *et al.*, 2011; Nergadze *et al.*, 2014, Purgato *et al.*, 2015).

We analyzed the segregation fidelity of horse chromosome 11 (ECA11), whose centromere is satellite-free, and compared it with that of horse chromosome 13 (ECA13), which is similar in size and has a centromere containing long stretches of the canonical horse centromeric satellite DNA families. Two chromosome stability assays interphase aneuploidy analysis and the cytokinesis-blocked micronucleus assay (CBMN) were combined with FISH with chromosome specific centromeric probes. The two assays were performed on control cells and on cells treated with nocodazole or griseofulvin. These drugs are well known antimitotic agents which interfere with the function of spindle and cytoplasmic microtubules by binding to tubulin; however, while nocodazole is a colchicine competitor which binds beta tubulin, griseofulvin binds both alpha and beta tubulin and does not compete with colchicine for tubulin binding. The use of these drugs was aimed at amplifying the difference, if any, in segregation fidelity between the satellite-less and the normal centromere and also at identifying possible differences in the sensitivity of the two centromeres to conditions perturbing cell division.

We decided to use two different chromosome stability assays. FISH on interphase nuclei since this is a rapid molecular-cytogenetic approach for the targeted detection of aneuploidies (Faas *et al.*, 2011) and the cytokinesis-blocked micronucleus assay (CBMN). The micronucleus assay is a mutagenic test system for detection of the formation of small membrane-bound DNA fragments (*i.e.* micronuclei in the cytoplasm of interphase cells) induced by chemical and physical agents (Fenech, 2000). Centric and acentric chromosome fragments, as well as whole chromosomes unable to migrate to one pole during anaphase, can be included into micronuclei. Two mechanisms, chromosome breakage and disturbance of chromosome segregation, may lead to the formation of micronuclei; in both cases, micronucleus expression requires a mitotic division. The cytokinesis-blocked micronucleus assay (CNBMN) allows to distinguish cells which completed nuclear division during in vitro culture since they are bi-nucleated after cytokinesis inhibition with cytochalasin (Kirsch-Volders *et al.*, 2011).

4.5.1 - Cell viability assay

To select the proper dose of each drug which depresses cell growth, but allows cell recovery after drug release, a cell viability assay was performed. Three doses were tested for both drugs: 5 µg/ml, 10 µg/ml and 20 µg/ml for griseofulvin and 100 nM, 200 nM and 300 nM for nocodazole. After a pre-culture period of 48h, cells were exposed for one cell cycle (18 hours) to the chemicals, then washed with fresh, drug-free, medium and grown for another cell cycle (recovery period).

In **Figure 23** the results of these experiments are reported. For griseofulvin (**Figure 23a**), the lowest dose determined a general decrease in cell growth rate but did not modify the cell growth curve; on the contrary, the highest dose induced a significant cell death, with no recovery of cell growth after drug release. The intermediate dose, which caused a depression of cell growth, but allowed cell recovery after drug release, was chosen.

The results of the cell viability test set up with nocodazole were similar to those observed with griseofulvin and the central dose (200 nM) was chosen also in this case (**Figure 23b**).

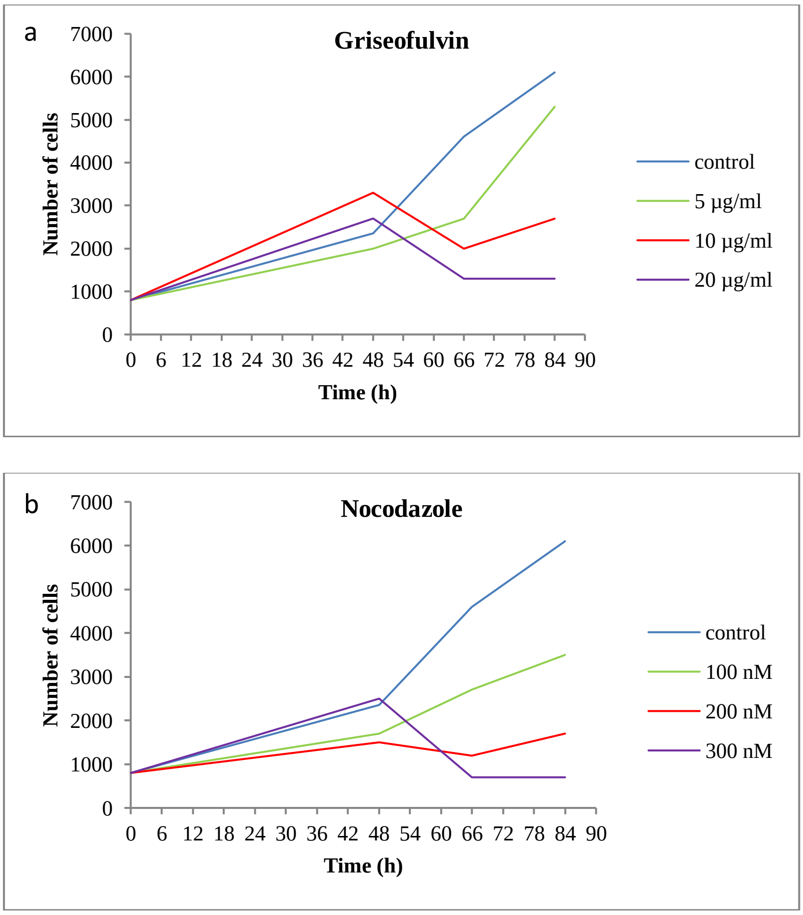

Figure 23 – Cell viability assay. (**a**) Treatment with griseofulvin: blue control, green 5µg/ml, red 10µg/ml, purple 20µg/ml. (**b**) Treatment with nocodazole: blue control, green 100nM, red 200nM, purple 300nM.

4.5.2 - Interphase aneuploidy analysis

To compare the migration fidelity of horse chromosome 11 with that of horse chromosome 13, interphase FISH with centromeric probes, specific for ECA11 and ECA13 was set up. Horse fibroblasts were exposed to the selected doses of griseofulvin and nocodazole and the cells were analyzed both just after the treatment and after a release period (corresponding to one cell cycle). The release period was required to identify segregation errors that need a lapse time to be expressed and errors which persist after *in vitro* selection.

In **Table Ia**, the total number of nuclei, aneuploid for chromosome 11 or for chromosome 13, observed in control and in treated cell cultures, is reported. Both griseofulvin and nocodazole induced a statistically significant increase in aneuploid nuclei. The comparison was performed with the chi-square test on data normalized for cell sample size.

In **Table Ib**, the total number of aneuploid nuclei observed after a recovery period (18 hours) in control and in treated cell cultures, is reported. Again, both griseofulvin and nocodazole induced a statistically significant increase of aneuploid nuclei, indicating that the effect of both drugs is reversible and that aneuploid cells are not counter selected *in vitro*. The comparison was performed with the chi-square test on data normalized for cell sample size.

In **Table IIa** and **IIb**, is reported the comparison of the mitotic behavior of ECA11 and ECA13, without and with release. In control cells, as well as in cells exposed to griseofulvin or nocodazole, no difference was observed between the two chromosomes neither without nor with the release period (chi-square test performed on normalized data for cell sample size). These results indicate that, in the horse model system, the proneness to segregation errors of a chromosome with a satellite-less centromere is comparable to that of a chromosome with a canonical, satellite based, centromere.

Table I – Interphase aneuploidy analysis without and with release. Control and treated cells are compared. Due to the difference in the total number of nuclei analyzed for each treatment, the χ2 test was performed after normalization for sample size. (**a**) Results of the analysis performed immediately after drug treatment. (**b**) Results of the analysis performed after a release period corresponding to one cell cycle.

a

interphase aneuploidy analysis without release			
		treated	
	control	GRF [10μg/ml]	NOC [200nM]
n. dip. nuc. (%)	2082 (95,5%)	1291 (93,4%)	916 (90,4%)
n. aneup. nuc. for ECA11 or ECA13 (%)	97 (4,5%)	91 (6,6%)	97 (9,6%)
total n. (%)	2179 (100%)	1382 (100%)	1013 (100%)

griseofulvin vs control p*=0,0019
nocodazole vs control p**=3,6396 E-11

b

interphase aneuploidy analysis with release			
		treated	
	control	GRF [10μg/ml]	NOC [200nM]
n. dip. nuc. (%)	2124 (96,7%)	1977 (95,4%)	1288 (93,7%)
n. aneup. nuc. for ECA11 or ECA13 (%)	72 (3,3%)	95 (4,6%)	87 (6,3%)
total n. (%)	2196 (100%)	2072 (100%)	1375 (100%)

griseofulvin vs control p*=0,0245
nocodazole vs control p**=2,2738 E-06

Table II – Comparison of the number of nuclei aneuploid for ECA11 and ECA13 without and with release. The χ^2 test was performed on data normalized for cell sample size. (**a**) Results of the analysis performed immediately after drug treatment. (**b**) Results of the analysis performed after a recovery period corresponding to one cell cycle.

a

FISH on interphase nuclei without release								
control			GRF [10μg/ml]			NOC [200nM]		
	ECA11	ECA13		ECA11	ECA13		ECA11	ECA13
n. dip. nuc. (%)	1045 (95,8%)	1037 (95,7%)	n. dip. nuc. (%)	695 (93,3%)	596 (93,6%)	n. dip. nuc.(%)	461 (89,9%)	455 (91%)
n. aneup. nuc. (%)	46 (4,2%)	47 (4,3%)	n. aneup. nuc. (%)	50 (6,7%)	41 (6,4%)	n. aneup. nuc. (%)	52 (10,1%)	45 (9%)
total n. (%)	1091 (100%)	1084 (100%)	total n. (%)	745 (100%)	637 (100%)	total n. (%)	513 (100%)	500 (100%)
		p = 0,5352			*p = 0,7966*			*p = 0,3839*

b

FISH on interphase nuclei with release								
control			GRF [10μg/ml]			NOC [200nM]		
	ECA11	ECA13		ECA11	ECA13		ECA11	ECA13
n. dip. nuc. (%)	1059 (96,4%)	1065 (97%)	n. dip. nuc. (%)	1052 (94,9%)	925 (96%)	n. dip. nuc. (%)	677 (93,4%)	611 (94%)
n. aneup. nuc. (%)	39 (3,6%)	33 (3%)	n. aneup. nuc. (%)	57 (5,1%)	38 (4%)	n. aneup. nuc. (%)	48 (6,6%)	39 (6%)
total number (%)	1098 (100%)	1098 (100%)	total n. (%)	1109 (100%)	963 (100%)	total n. (%)	725 (100%)	650 (100%)
		p = 0,4722			*p = 0,1519*			*p = 0,5399*

4.5.3 - Cytokinesis-blocked micronucleus assay (CNBMN)

Since the CNBMN assay is performed on cytokinesis blocked cells a release period after drug treatment was not performed. Horse fibroblasts were exposed to the selected doses of griseofulvin and nocodazole in the presence of cytochalasin (an inhibitor of actin) and the cells were analyzed just after the treatment. In **Table III** the total number of micronuclei observed in a simple of 1500 binucleated cells in control and in treated cell cultures is reported. Both the drugs induced a statistically significant increase in micronuclei (chi-square test).

The behavior of the two chromosomes at mitosis was compared following the CBMN assay. FISH experiments with the centromeric probes specific for ECA11 and ECA13 were set up. The comparison of the rate of micronuclei containing horse chromosome 11 or horse chromosome 13, as revealed by FISH (**Table IV**), demonstrated that the mitotic behavior of the two chromosomes was comparable in all the conditions tested. These results confirm those obtained by FISH interphase analysis.

Thus, two independent molecular-cytogenetic approaches demonstrated that, in the horse system, the *in vitro* segregation fidelity of a chromosome is not influenced by the presence of highly repetitive DNA sequences at its centromere.

Table III – Cytokinesis-blocked micronucleus test. The number of micronuclei observed in control cells is compared with the one observed after drug treatment. The results were compared with the χ^2 test.

CNBMN			
		treated	
	control	GRF [10μg/ml]	NOC [200nM]
n. binucleated cells	1500	1500	1500
n. of micronuclei (%)	30 (2%)	81 (5,4%)	143 (9,5%)

griseofulvin vs control p*=1,9923 E-06
nocodazole vs control p**=6,3229 E-17

Table IV – Comparison of the number of micronuclei positive for ECA11 and for ECA 13. The data were compared by the χ^2 test.

FISH on MN								
control			GRF [10μg/ml]			NOC [200nM]		
	ECA11	ECA13		ECA11	ECA13		ECA11	ECA13
n. pos.	0	2	n. pos.	6	5	n. pos.	5	3
MN (%)	(0%)	(13,3%)	MN (%)	(16,2%)	(11,4%)	MN (%)	(6,9%)	(4,2%)
n. neg.	15	13	n. neg.	31	39	n. neg.	67	68
MN (%)	(100%)	(86,7%)	MN (%)	(83,8%)	(88,6%)	MN (%)	(93,1%)	(95,8%)
total n.	15	15	total n.	37	44	total n.	72	71
MN (%)	(100%)	(100%)	MN (%)	(100%)	(100%)	MN (%)	(100%)	(100%)
		p = 0,1432			*p = 0,525*			*p = 0,4793*

Here we analysed the *in vitro* mitotic behaviour of the satellite-less centromere of ECA11, and compared it with that of the canonical, satellite-based, ECA13 centromere. Our results demonstrated that the segregation accuracy of these two chromosomes is similar, thus suggesting that satellite DNA is dispensable for transmission fidelity. As mentioned in the introduction (see paragraph 1.2.1), the role played by satellite DNA at the centromere is a matter of debate, literature data strongly suggesting that centromeric and/or pericentromeric repeated DNA sequences create an ideal chromatin environment needed for sister chromatid cohesion and for kinetochore recruitment (Marshall *et al.*, 2008; Nakano *et al.*, 2008; Rocchi *et al.*, 2012; Shang *et al.*, 2013; Rošić *et al.*, 2014; Biscotti *et al.*, 2015; Rošić and Erhardt, 2016). Indeed, the large majority of vertebrate centromeres contain highly repeated DNA sequences (Plohl *et al.*, 2008; Plohl *et al.*, 2014); the biological preference for repeated DNA at centromeres suggests that there is a positive selection for centromeres with this kind of arrangement. It must be also reminded that, karyotypes containing only chromosomes with centromeres completely devoid of repeated DNA sequences have never been described, nor in animals nor in plants (Plohl *et al.*, 2008). In this scenario, it might be hypothesized that centromeres void of highly repeated DNA stretches are somehow defective, this is the reason why they tend to accumulate satellite DNA sequences during their evolutionary maturation (Piras *et al.*, 2010), and that the missing functions may be provided *in trans* by canonical centromeres by means of genetic complementation.

4.6 – Analysis of the centromeric and pericentromeric chromatin histone modifications in the horse and in the donkey

Taking advantage of the equid model system in which satellite-based and satellite free centromeres, at different stage of maturation, coexist, a molecular cytogenetic analysis of the main histone modifications characterizing the centrochromatin of satellite-based and satellite-free centromeres was carried out.

In the horse, one chromosome presents a satellite-free centromere and in the donkey, sixteen chromosomes have centromeres completely devoid of satellite DNA. In view of the absence of repetitive DNA arrays, the only elements which can specify the centromeric competence are epigenetic factors.

As mentioned in the introduction (see paragraph 1.4), at the centromere a specific ratio of typically euchromatic and typically heterochromatic post-translational modification exists. This balance specifies for a peculiar chromatin, named centrochromatin, which is prone to transcription, a prerequisite for CENP-A recruitment.

A molecular cytogenetic approach, based on double immunofluorescence on metaphase chromosomes, mechanically stretched chromosomes and extended chromatin fibers, was used to analyze the arrangement of modified centromeric histones localization in the horse and in the donkey.

We analyzed four histone modifications: the trimethylation of lysine 9 of histone H3 (H3K9me3), the dimethylation of lysine 9 of histone H3 (H3K9me2), the trimethylation of lysine 27 of histone H3 (H3K27me3) and the dimethylation of lysine 4 of histone H3 (H3K4me2).

The trimethylation of lysine 9 of histone H3 (H3K9me3) is the best known constitutive heterochromatin marker. H3K9me3 has been found in the pericentromeric regions of *Drosophila*, mouse and human chromosomes (Peters *et al.,* 2003; Rice *et al.,* 2003) and not at all in the centromere core. On human chromosomes, H3K9me3 is concentrated in the pericentromeric regions of chromosomes that contain large blocks of satellite DNA and is also located in sequences which are far away from CENP-A-containing domains (Sullivan and Karpen, 2004).

The dimethylation of lysine 9 of histone H3 (H3K9me2) is a marker of facultative heterochromatin; this type of histone modification is present in regions actively transcribed that must be silenced if necessary. This modification is present in the pericentromeric regions both in man and in *Drosophila*. (Nagaki *et al.*, 2004; Sullivan and Karpen, 2004; Beiley *et al.*, 2015).

Histone H3 trimethylated at the lysine 27 (H3K27me3) is a marker of facultative heterochromatin, related to the presence of satellite DNA, as well as to regions of single copy DNA (Aldrup-Macdonald and Sullivan, 2014; Miga, 2015). It is present in regions usually silenced, that can be activated if necessary (Mravinac *et al.*, 2009).

Dimethylation of histone H3 at lysine 4 (H3K4me2) is involved in the transition to an active chromatin state. This modification has been found at promoters and transcribed genomic regions and it is related to a not necessarily active euchromatin (Schneider *et al.*, 2004; Lam *et al.*, 2006).

First of all we analyzed the architectural organization of post-translational modification of satellite-less and satellite-based centromeres, by two color immunofluorescence with antibodies against H3K9me3 and CENP-A in the horse and in the donkey at different resolution levels. In **Figure 24** double immunofluorescence analysis of horse and donkey metaphase chromosomes is shown. The H3K9me3 signal localization is pseudo-coloured in green and the CENP-A localization is pseudo-colored in red.

In the horse (**Figure 24a**) we observed a constant CENP-A signal intensity al all centromere. On the contrary, the H3K9me3 signal was variable in size and intensity among centromeres. Notably, no centromeres devoid of H3K9me3 immunostaining were observed. These results demonstrated that also the centromere of horse chromosome 11, while being satellite-free, is immersed in a heterochromatic environment.

In donkey metaphase chromosomes, the distribution of H3K9me3 signals was different. As shown in **Figure 24b**, the distribution and the intensity of heterochromatic signals, in green, were highly variable among chromosomes. In detail, a variable number of centromeres displayed a very faint H3K9me3 signal, notably, most of the chromosomes which showed a weak centromeric heterochromatic signal, exhibited a large H3K9me3 positive region at one telomere. These results are in agreement with our previous data concerning the distribution of satellite DNA families in the Donkey (Piras *et al*, 2010). We can therefore infer that the donkey centromeres showing a faint H3K9me3 signal are indeed those satellite-less centromeres (18 in the donkey) that conserve residual satellite DNA sequence at one telemetric terminus (Piras *et al*, 2010).

Donkey chromosome 1 displayed a peculiar patter of heterochromatin distribution: large positive regions were detected at the p telomeric terminus and, in the sub-centromeric region. This pattern was expected since we previously demonstrated that, donkey chromosome 1 has an abundant and highly polymorphic heterochromatin content (Raimondi *et al.*, 2011). Taking into account of all these evidences, it is possible to hypothesize that, the centromeres that displayed a weak heterochromatic signal are bona fide satellite-less. This hypothesis will be confirmed by immuno-FISH experiments using chromosome-specific probes.

In summary this part of the research indicated that the trimethylation of lysine 9 of histone H3 marks heterochromatic regions that contain large blocks of satellite DNA, as described by Sullivan and Karpen (Sullivan and Karpen, 2004) but this post-translational modification is also present at satellite-less centromeres, although in smaller amount.

A peculiar centromere organization was observed; in the horse and in the donkey: the sister CENP-A spots faced the outside of the centromeric *locus*, towards the plates of the kinetochore, while the heterochromatic core, as identified by the presence of the H3K9me3 modification, was confined in the inner centromere structure (**Figure 24a-b**). This result perfectly reflects the hypothetical three-dimensional arrangement of super coiled centro-chromatin. It has been hypothesized that centromeric DNA may supercoil in a cylindrical structure, leading to the alignment of nucleosomes with the same composition, to promote proper kinetochore assembly, CENP-A nucleosomes being exposed in the outer face of the cylinder to be able to interact with spindle fibers (Blower *et al.*, 2002) (**Figure 25**).

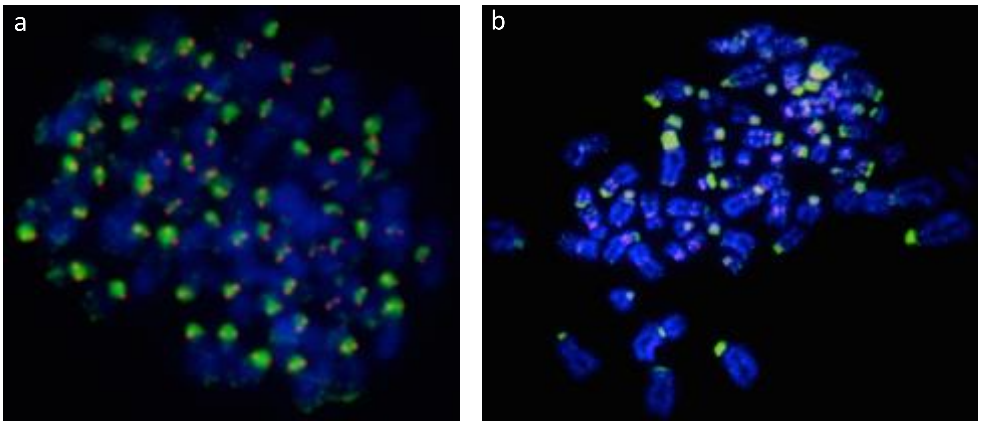

Figure 24 – Double immunofluorescence on horse (a) and donkey (b) metaphase chromosomes. Metaphase chromosomes show the localization of the trimethylation of lysine 9 of histone H3 (H3K9me3) (green) with respect to CENP-A (red).

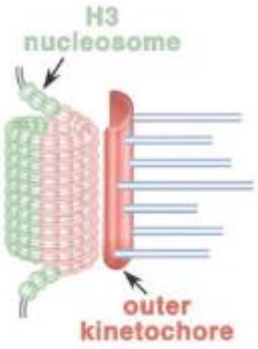

Figure 25 – Three-dimensional arrangement of super coiled centro-chromatin (Blower *et al.*, 2002)**.**

To better investigate the H3K9me3 centromeric localization, the resolution level was improved. Double immunofluorescence on mechanically stretched chromosome was performed. An example of the results is reported in **Figure 26**. In this acrocentric chromosome, the heterochromatic region, in green, is clearly present between the CENP-A spots, in red. This image is the direct morphological evidence of the model depicted in **Figure 25**.

Further increase of the resolution was achieved by experiments set up on extended chromatin fibers (**Figure 27**); as shown in the figure, the H3K9me3 histone modification signal (green) flanks the core centromere functional domain (red).

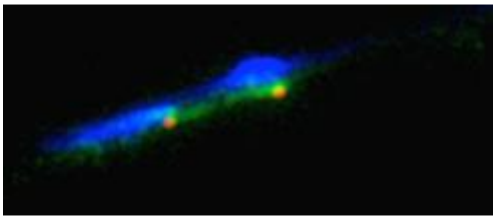

Figure 26 – Double immunofluorescence on horse mechanically stretched chromosome. Localization of the trimethylation of lysine 9 of histone H3 (H3K9me3) (green) with respect to CENP-A (red).

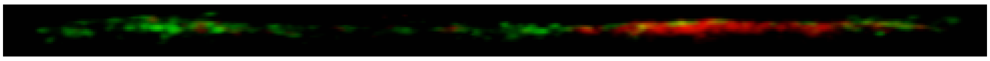

Figure 27 – Double immunofluorescence on horse extended chromatin fibers. Localization of the trimethylation of lysine 9 of histone H3 (H3K9me3) (green) with respect to CENP-A (red).

Afterwards, the distribution of H3K9me2 was analyzed in horse (**Figure 28a**) and in donkey (**Figure 28b**) metaphase chromosomes. No difference was in the two species. All centromeres were positive to the immunostaining indicating that both satellite-less and satellite-containing centromeres contain heterochromatin prone to be opened. This result is not surprising since a relaxed chromatin environment is known to be essential for CENP-A loading.

The analysis of extended chromatin fibers (**Figure 29**) confirmed the data. No difference was observed between horse and donkey and all the analyzed fibers showed the same pattern were CENPA and H3K9me2 signals were superimposed. Two examples are reported in **Figure 29**. Again, the histone modification is green labelled while the centromere functional domain is red colored.

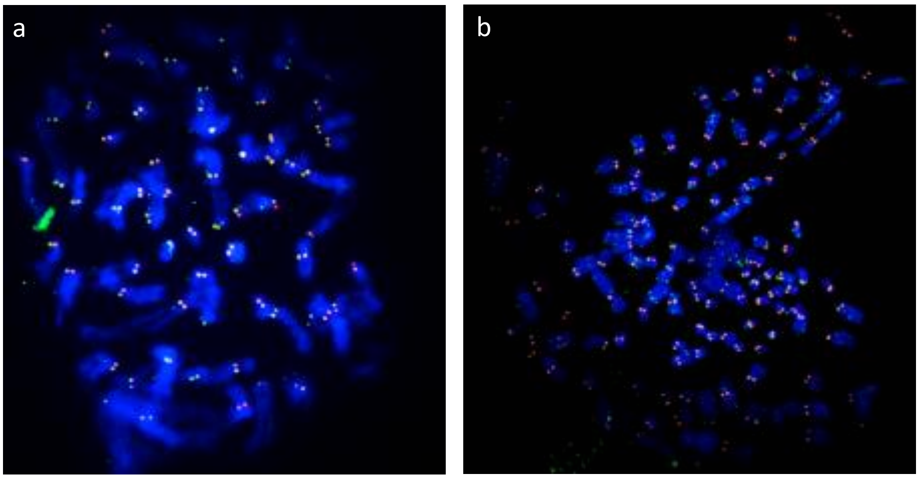

Figure 28 – Double immunofluorescence on horse (a) and donkey (b) metaphase chromosomes. Metaphase chromosomes show the localization of the dimethylation of lysine 9 of histone H3 (H3K9me2) (green) with respect to CENP-A (red).

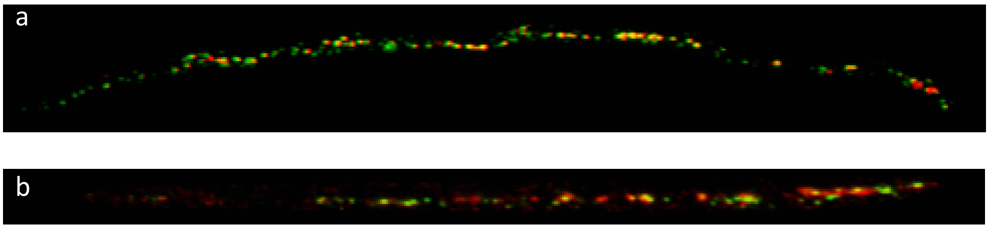

Figure 29 – Double immunofluorescence on horse (a) and donkey (a) extended chromatin fibers. Localization of the dimethylation of lysine 9 of histone H3 (H3K9me2) (green) with respect to CENP-A (red).

The third histone modification investigated was H3K27me3. A double immuno-fluorescence analysis on horse (**Figure 30a**) and donkey (**Figure 30b**) metaphase chromosomes was performed. No difference was observed between horse and donkey, all centromeres being positive to the immunostaining.

No difference has been observed between horses and donkeys. All centromeres were positive to the immunostaining antibody against H3K27me3. Intriguing, H3K27me3 signals show a G-like banding and this is in accordance with the presence of this modification in silenced regions that can be activated if necessary (Mravinac *et al.*, 2009). In fact, the G-banding highlights late replication regions which are typically heterochromatic and gene-poor. This means that at the centromere is present a facultative heterochromatin prone to be opened to allow the CENP-A loading.

The analysis of extended chromatin fibers confirmed these data. Two examples are reported in **Figure 31**. The histone modification is green labelled while the centromere functional domain is red colored. It can be observed that blocks of chromatin containing the dimethylated histone H3 are interspersed within the centromere functional domain, identified by CENP-A.

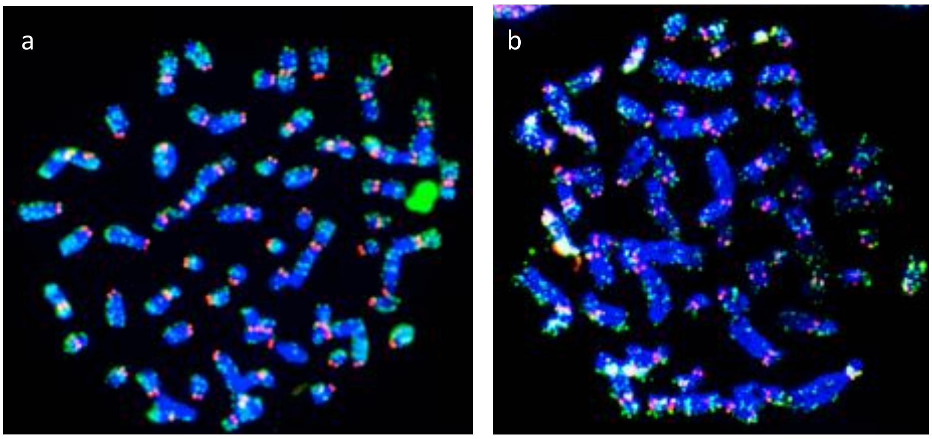

Figure 30 – Double immunofluorescence on horse (a) and donkey (b) metaphase chromosomes. Metaphase chromosomes show the localization of the trimethylation of lysine 27 of histone H3 (H3K27me3) (green) with respect to CENP-A (red).

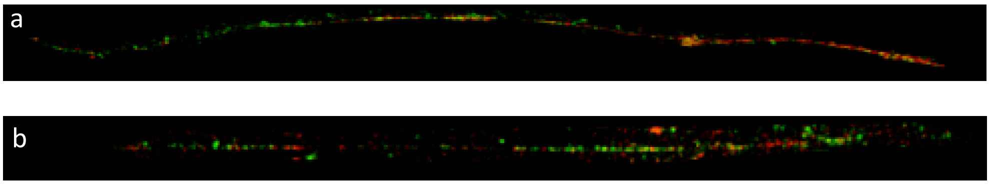

Figure 31 – Double immunofluorescence on horse (a) and donkey (b) extended chromatin fibers. Localization of the trimethylation of lysine 27 of histone H3 (H3K27me3) (green) with respect to CENP-A (red).

The following experiments were aimed at analyzing the distribution of H3K4me2 on horse (**Figure 32a**) and donkey (**Figure 32b**) metaphase chromosomes. No difference has been observed between horses and donkeys. All centromeres were positive to the immunostaining.

H3K4me2 is related to a not necessarily active euchromatin (Schneider *et al.*, 2004; Lam *et al.*, 2006), this means that all the centromeres, including the satellite-less ones, possess permissive centrochromatin which is prone to be eventually transcribed. Since H3K4me3 immuno-staining highlights transcriptionally competent regions, the resulting pattern is the opposite of the one observed using the H3K9me3 immunostaining (**Figure 24**).

The analysis of extended chromatin fibers (**Figure 33**) confirms the presence of the H3K4me3 modification at all the centromeres, both in the horse and in the donkey. Indeed, in all the analyzed fibers, CENP-A signal and H3K4me3 signal always colocalized. These data confirmed the need to possess a transcriptionally competent heterochromatin at the centromere.

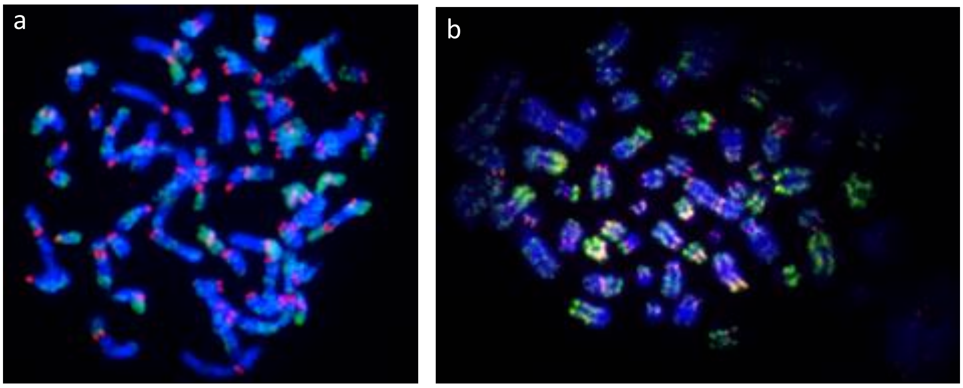

Figure 32 – Double immunofluorescence on horse (a) and donkey (b) metaphase chromosomes. Metaphase chromosomes show the localization of the dimethylation of lysine 4 of histone H3 (H3K4me2) (green) with respect to CENP-A (red).

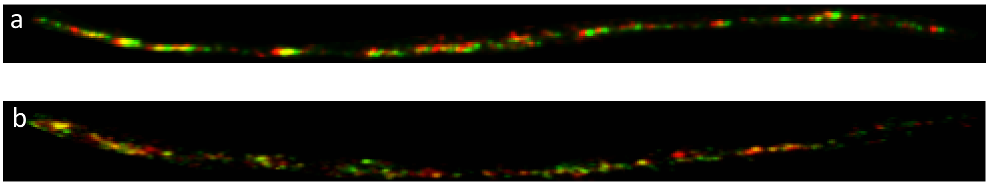

Figure 33 – Double immunofluorescence on horse (a) and donkey (b) extended chromatin fibers. Localization of the dimethylation of lysine 4 of histone H3 (H3K4me2) (green) with respect to CENP-A (red).

The results of histone modifications analysis, strongly suggest that the satellite-less centromeres, as well as the satellite-based ones, are immersed into a heterochromatic environment, even if they contain small amounts of constitutive heterochromatin. This constitutive hyper-condensed heterochromatin is presumably needed to define the borders of the functional centromere domain preventing centrochromatin diffusion. Our observation of the sliding behavior of the centromeric domain of the satellite-free centromere of horse chromosome 11 (see paragraph 4.3) demonstrated that the centromere function can move. However, this positional instability is confined to a region spanning about 500 kb.

Satellite-less centromeres do also contain facultative heterochromatin: those “conditional” heterochromatin is supposed to be a prerequisite for proneness to load the centromere mark CENP-A. Finally, satellite-less centromeres do contain transcriptionally competent heterochromatin. This type of chromatin is not necessarily transcribed but might interact with *trans* acting lncRNAs transcribed from satellite based centromeres.

5. CONCLUSIONS

The centromere of mammalian chromosomes is typically embedded in a heterochromatic environment characterized by long arrays of tandemly repeated satellite DNA. However, centromere DNA array length is highly variable, both among homologous and heterologous centromeres, moreover, centromeric DNA sequences are rapidly evolving among species and within chromosomes of the same species (Plohl *et al.*, 2008). Satellite DNA is not necessary for centromere function since satellite-DNA-free centromeres have been found in human pathology (Voullaire *et al.*, 1993; Marshall *et al.*, 2008) and in extant species (Wade *et al.*, 2009; Shang *et al.*, 2010; Locke *et al.*, 2011).

The work described in this thesis is part of a collaborative project involving the laboratory of Molecular Cytogenetics – directed by professor Elena Raimondi – and the laboratory of Molecular and Cellular Biology – leaded by professor Elena Giulotto – aimed at studying mammalian centromere structure, identity and function. To achieve this target, species belonging to genus *Equus* are used as a biological model system since satellite-based centromeres and satellite-free centromeres coexist in a single karyotype (Wade *et al.*, 2009; Piras *et al.*, 2010).

High resolution FISH analysis on combed DNA fibers demonstrated that satellite DNA clusters at horse centromeres show a peculiar architectural organization, where small arrays of 2PI and EC137 satellites are strictly intermingled and immerged within very large stretches of the 37cen sequence. This observation allows us to hypothesize that, at horse centromere, satellite sequence interchanges are a frequent occurrence; this hypothesis agrees with the highly plastic nature of equid genomes. Centromeric horse satellite EC137 is an accessory DNA element, presumably contributing to the organization of pericentromeric chromatin while the 37cen sequence is associated with the centromeric function and is transcriptionally active. Moreover, 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.

Concerning the satellite-less centromere found in horse chromosome 11, a remarkable plasticity of the centromeric domain has been demonstrated. Out of ten horse chromosomes 11, at least seven distinct CENP-A binding domains were found across a region of about 500 kb. These results demonstrate that, in a native mammalian centromere, the positioning of CENP-A binding domains is unrelated to the sequence of the DNA the centromere is associated with and that centromere position can be flexible across a relatively wide single-copy genomic region.

The *in vitro* mitotic behavior of the satellite-less centromere of horse chromosome 11 is comparable with that of the satellite-based centromere of horse chromosome 13 (which has similar size and a centromere containing long stretches of the canonical horse centromeric satellite DNA families). The segregation accuracy of these two chromosomes is similar, thus suggesting that satellite DNA is dispensable for transmission fidelity even if the biological preference for repeated DNA at centromeres suggests that there is a positive selection for centromeres with this kind of arrangement. In this scenario, it might be hypothesized that centromeres void of highly repeated DNA stretches are somehow defective, this is the reason why they tend to accumulate satellite DNA sequences during their evolutionary maturation (Piras *et al.*, 2010), and that the missing functions may be provided in *trans* by canonical centromeres by means of genetic complementation.

Finally, the analysis of the centromeric histone modifications in the horse and in the domestic donkey revealed that satellite-free centromeres contain small amounts of constitutive heterochromatin and that this constitutive hyper-condensed heterochromatin defines the borders of the functional centromere domain thus preventing centrochromatin diffusion. Satellite-less centromeres do contain heterochromatin prone to be opened which is needed for CENP-A loading and, finally, satellite-less centromeres do contain transcriptionally competent heterochromatin. We presume that the permissive centrochromatin of satellite-less centromeres is prone to interact with *trans* acting lncRNAs transcribed from satellite based centromeres.

6. REFERENCES

Aldrup-Macdonald ME, Sullivan BA. The past, present, and future of human centromere genomics. Genes (Basel). 2014; 5:33-50.

Alexiadis V, Ballestas ME, Sanchez C, Winokur S, Vedanarayanan V, *et al.* RNAPol-ChIP analysis of transcription from FSHD-linked tandem repeats and satellite DNA. Biochim Biophys Acta. 2007; 1769:29-40.

Alkan C, Cardone MF, Catacchio CR, Antonacci F, O'Brien SJ, *et al.* Genome-wide characterization of centromeric satellites from multiple mammalian genomes. Genome Res. 2011; 21:137-145.

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

Alonso A, Mahmood R, Li S, Cheung F, Yoda K, Warburton PE. Genomic microarray analysis reveals distinct locations for the CENP-A binding domains in three human chromosome 13q32 neocentromeres. Hum Mol Genet. 2003; 12:2711-2721.

Alonso A, Hasson D, Cheung F, Warburton PE. A paucity of heterochromatin at functional human neocentromeres. Epigenetics Chromatin. 2010; 3:6.

Amano M, Suzuki A, Hori T, Backer C, Okawa K, *et al.* The CENP-S complex is essential for the stable assembly of outer kinetochore structure. J Cell Biol. 2009; 186:173-182.

Amor DJ, Choo KH. 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.

Bachmann L, Raab M, Sperlic D. Satellite DNA and speciation: a species specific satellite DNA of *Drosophila guanche*. Journal of Zoological Systematics and Evolutionary Research. 1989; 27:84-93.

Bailey AO, Panchenko T, Shabanowitz J, Lehman SM, Bai DL, *et al.* Identification of the post-translational modifications present in centromeric chromatin. Mol Cell Proteomics. 2016; 15:918-931.

Barnhart MC, Kuich PH, Stellfox ME, Ward JA, Bassett EA, *et al.* HJURP is a CENP-A chromatin assembly factor sufficient to form a functional *de novo* kinetochore. J Cell Biol. 2011; 194:229-243.

Bassett EA, DeNizio J, Barnhart-Dailey MC, Panchenko T, Sekulic N, *et al.* HJURP uses distinct CENP-A surfaces to recognize and to stabilize CENP-A/histone H4 for centromere assembly. Dev Cell. 2012; 22:749-762.

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

Bernard P, Maure JF, Partridge JF, Genier S, Javerzat JP, *et al.* Requirement of heterochromatin for cohesion at centromeres. Science. 2001; 294:2539-2542.

Biscotti MA, Canapa A, Forconi M, Olmo E, Barucca M. Transcription of tandemly repetitive DNA: functional roles. Chromosome Res. 2015; 23:463-477.

Black BE, Jansen LE, Maddox PS, Foltz DR, Desai AB, *et al.* Centromere identity maintained by nucleosomes assembled with histone H3 containing the CENP-A targeting domain. Mol Cell. 2007; 25:309-322.

Blom E, Heyning FH, Kroes WG. A case of angioimmunoblastic T-cell non-Hodgkin lymphoma with a neocentric inv dup(1). Cancer Genet Cytogenet. 2010; 202:38-42.

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

Bodor DL, Mata JF, Sergeev M, David AF, Salimian KJ, *et al.* The quantitative architecture of centromeric chromatin. Elife. 2014; 3:e02137.

Bouzinba-Segard H, Guais A, Francastel C. Accumulation of small murine minor satellite transcripts leads to impaired centromeric architecture and function. Proc Natl Acad Sci USA. 2006; 103:8709-8714.

Burnside RD, Ibrahim J, Flora C, Schwartz S, Tepperberg JH, *et al.* Interstitial deletion of proximal 8q including part of the centromere from unbalanced segregation of a paternal deletion/marker karyotype with neocentromere formation at 8p22. Cytogenet Genome Res. 2011; 132:227-232.

Burrack LS, Berman J. Neocentromeres and epigenetically inherited features of centromeres. Chromosome Res. 2012; 20:607-619.

Canapa A, Barucca M, Cerioni PN, Olmo E. A satellite DNA containing CENP-B box-like motifs is present in the antarctic scallop *Adamussium colbecki*. Gene. 2000; 247:175-180.

Carbone L, Nergadze SG, Magnani E, Misceo D, Francesca Cardone M, *et al*. Evolutionary movement of centromeres in horse, donkey, and zebra. Genomics. 2006; 87:777-782.

Cardone MF, Alonso A, Pazienza M, Ventura M, Montemurro G, *et al.* Independent centromere formation in a capricious, gene-free domain of chromosome 13q21 in Old World monkeys and pigs. Genome Biol. 2006; 7:R91.

Carroll CW, Silva MC, Godek KM, Jansen LE, Straight AF. Centromere assembly requires the direct recognition of CENP-A nucleosomes by CENP-N. Nat Cell Biol. 2009; 11:896-902.

Carroll CW, Milks KJ, Straight AF. Dual recognition of CENP-A nucleosomes is required for centromere assembly. J Cell Biol. 2010; 189:1143-1155.

Casola C, Hucks D, Feschotte C. Convergent domestication of pogo-like transposases into centromere-binding proteins in fission yeast and mammals. Mol Biol Evol. 2008; 25:29-41.

Cerutti F, Gamba R, Mazzagatti A, Piras FM, Cappelletti E, *et al.* The major horse satellite DNA family is associated with centromere competence. Mol Cytogenet. 2016; 9:35.

Chan FL, Marshall OJ, Saffery R, Kim BW, Earle E, *et al.* Active transcription and essential role of RNA polymerase II at the centromere during mitosis. Proc Natl Acad Sci USA. 2012; 109:1979-1984.

Cheeseman IM, Drubin DG, Barnes G. Simple centromere, complex kinetochore: linking spindle microtubules and centromeric DNA in budding yeast. J Cell Biol. 2002; 157:199-203.

Cheng Z, Dong F, Langdon T, Ouyang S, Buell CR, *et al.* Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. Plant Cell. 2002; 14:1691-1704.

Choi ES, Strålfors A, Catania S, Castillo AG, Svensson JP, *et al.* Factors that promote H3 chromatin integrity during transcription prevent promiscuous deposition of CENP-A (Cnp1) in fission yeast. PLoS Genet. 2012; 8:e1002985.

Choo KH. Centromere DNA dynamics: latent centromeres and neocentromere formation. Am J Hum Genet. 1997; 61:1225-1233.

Choo KH. Centromerization. Trends Cell Biol. 2000; 10:182-128.

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

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

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

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

Copenhaver GP, Nickel K, Kuromori T, Benito MI, Kaul S, *et al.* Genetic definition and sequence analysis of *Arabidopsis* centromeres. Science. 1999; 286:2468-2474.

Csink AK, Henikoff S. Something from nothing: the evolution and utility of satellite repeats. Trends Genet. 1998; 14:200-204.

Dawe RK, Cande WZ. Induction of centromeric activity in maize by suppressor of meiotic drive 1. Proc Natl Acad Sci USA. 1996; 93:8512-8517.

Denegri M, Moralli D, Rocchi M, Biggiogera M, Raimondi E, *et al.* Human chromosomes 9, 12, and 15 contain the nucleation sites of stress-induced nuclear bodies. Mol Biol Cell. 2002; 13:2069-2079.

Depinet TW, Zackowski JL, Earnshaw WC, Kaffe S, Sekhon GS, *et al.* Characterization of neo-centromeres in marker chromosomes lacking detectable alpha-satellite DNA. Hum Mol Genet. 1997; 6:1195-1204.

Dong F, Miller JT, Jackson SA, Wang GL, Ronald PC, *et al.* Rice (*Oryza sativa*) centromeric regions consist of complex DNA. Proc Natl Acad Sci USA. 1998; 95:8135-8140.

Dover GA. Molecular drive in multigene families: how biological novelties arise, spread and are assimilated. Trends Genet. 1986; 2:159-165.

Dunleavy EM, Roche D, Tagami H, Lacoste N, Ray-Gallet D, *et al.* HJURP is a cell-cycle-dependent maintenance and deposition factor of CENP-A at centromeres. Cell. 2009; 137:485-497.

Durajlija-Žinić S, Ugarković D, Cornudella L, Plohl M. A novel interspersed type of organization of satellite DNAs in *Tribolium madens* heterochromatin. Chromosome Res 2000; 8:201-212

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

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, *et al.* Molecular cloning of cDNA for CENP-B, the major human centromere autoantigen. J Cell Biol. 1987; 104:817-829.

Ehrlich M, Sanchez C, Shao C, Nishiyama R, Kehrl J, *et al.* ICF, an immunodeficiency syndrome: DNA methyltransferase 3B involvement, chromosome anomalies, and gene dysregulation. Autoimmunity. 2008; 41:253-271.

Enukashvily NI, Donev R, Waisertreiger IS, Podgornaya OI. Human chromosome 1 satellite 3 DNA is decondensed, demethylated and transcribed in senescent cells and in A431 epithelial carcinoma cells. Cytogenet Genome Res. 2007; 118:42-54.

Enukashvily NI, Ponomartsev NV. Mammalian satellite DNA: a speaking dumb. Adv Protein Chem Struct Biol. 2013; 90:31-65.

Eymery A, Horard B, El Atifi-Borel M, Fourel G, Berger F, *et al.* A transcriptomic analysis of human centromeric and pericentric sequences in normal and tumor cells. Nucleic Acids Res. 2009; 37:6340-6354.

Faas BH, Cirigliano V, Bui TH. Rapid methods for targeted prenatal diagnosis of common chromosome aneuploidies. Semin Fetal Neonatal Med. 2011; 16:81-87.

Fachinetti D, Folco HD, Nechemia-Arbely Y, Valente LP, Nguyen K, *et al.* A two-step mechanism for epigenetic specification of centromere identity and function. Nat Cell Biol. 2013; 15:1056-1066.

Fenech M. The *in vitro* micronucleus technique. Mutat Res. 2000; 455:81-95.

Ferri F, Bouzinba-Segard H, Velasco G, Hubé F, Francastel C. Non-coding murine centromeric transcripts associate with and potentiate Aurora B kinase. Nucleic Acids Res. 2009; 37:5071-5080.

Fischle W, Wang Y, Allis CD. Histone and chromatin cross-talk. Curr Opin Cell Biol. 2003; 15:172-183.

Fitzgerald-Hayes M, Clarke L, Carbon J. Nucleotide sequence comparisons and functional analysis of yeast centromere DNAs. Cell. 1982; 29:235-244.

Foltz DR, Jansen LE, Black BE, Bailey AO, Yates JR 3rd, *et al.* The human CENP-A centromeric nucleosome-associated complex. Nat Cell Biol. 2006; 8:458-469.

Foltz DR, Jansen LE, Bailey AO, Yates JR 3rd, Bassett EA, *et al.* Centromere-specific assembly of CENP-a nucleosomes is mediated by HJURP. Cell. 2009; 137:472-484.

Frescas D, Guardavaccaro D, Kuchay SM, Kato H, Poleshko A, *et al.* KDM2A represses transcription of centromeric satellite repeats and maintains the heterochromatic state. Cell Cycle. 2008; 7:3539-3547.

Fukagawa T, Mikami Y, Nishihashi A, Regnier V, Haraguchi T, *et al.* CENP-H, a constitutive centromere component, is required for centromere targeting of CENP-C in vertebrate cells. EMBO J. 2001; 20:4603-4617.

Fukagawa T, Nogami M, Yoshikawa M, Ikeno M, Okazaki T, *et al.* Dicer is essential for formation of the heterochromatin structure in vertebrate cells. Nat Cell Biol. 2004; 6:784-791.

Fukagawa T, Earnshaw WC. Neocentromeres. Curr Biol. 2014; 24:R946-947.

Gindullis F, Dechyeva D, Schmidt T. Construction and characterization of a BAC library for the molecular dissection of a single wild beet centromere and sugar beet (*Beta vulgaris*) genome analysis. Genome. 2001; 44:846-855.

Gong Z, Wu Y, Koblízková A, Torres GA, Wang K, *et al.* Repeatless and repeat-based centromeres in potato: implications for centromere evolution. Plant Cell. 2012; 24:3559-3574.

Guenatri M, Bailly D, Maison C, Almouzni G. Mouse centric and pericentric satellite repeats form distinct functional heterochromatin. J Cell Biol. 2004; 166:493-505.

Haaf T, Warburton PE, Willard HF. Integration of human alpha-satellite DNA into simian chromosomes: centromere protein binding and disruption of normal chromosome segregation. Cell. 1992; 70:681-696.

Hall LE, Mitchell SE, O'Neill RJ. Pericentric and centromeric transcription: a perfect balance required. Chromosome Res. 2012; 20:535-546.

Harrington JJ, Van Bokkelen G, Mays RW, Gustashaw K, Willard HF. Formation of *de novo* centromeres and construction of first-generation human artificial microchromosomes. Nat Genet. 1997; 15:345-355.

Hasson D, Alonso A, Cheung F, Tepperberg JH, Papenhausen PR, *et al.* Formation of novel CENP-A domains on tandem repetitive DNA and across chromosome breakpoints on human chromosome 8q21 neocentromeres. Chromosoma. 2011; 120:621-632.

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

Henikoff JG, Thakur J, Kasinathan S, Henikoff S. A unique chromatin complex occupies young α-satellite arrays of human centromeres. Sci Adv. 2015; 1.pii:e1400234.

Heun P, Erhardt S, Blower MD, Weiss S, Skora AD, *et al.* Mislocalization of the *Drosophila* centromere-specific histone CID promotes formation of functional ectopic kinetochores. Dev Cell. 2006; 10:303-315.

Heus JJ, Zonneveld BJ, Steensma HY, Van den Berg JA. Centromeric DNA of Kluyveromyces lactis. Curr Genet. 1990; 18:517-522.

Hori T, Amano M, Suzuki A, Backer CB, Welburn JP, *et al.* CCAN makes multiple contacts with centromeric DNA to provide distinct pathways to the outer kinetochore. Cell. 2008; 135:1039-1052.

Horvath JE, Gulden CL, Vallente RU, Eichler MY, Ventura M, *et al.* Punctuated duplication seeding events during the evolution of human chromosome 2p11. Genome Res. 2005; 15:914-927.

Hsieh CL, Lin CL, Liu H, Chang YJ, Shih CJ, *et al.* WDHD1 modulates the post-transcriptional step of the centromeric silencing pathway. Nucleic Acids Res. 2011; 39:4048-4062.

Hudson DF, Fowler KJ, Earle E, Saffery R, Kalitsis P, *et al.* Centromere protein B null mice are mitotically and meiotically normal but have lower body and testis weights. J Cell Biol. 1998; 141:309-319.

Ideue T, Cho Y, Nishimura K, Tani T. Involvement of satellite I noncoding RNA in regulation of chromosome segregation. Genes Cells. 2014; 19:528-538.

Ikeno M, Grimes B, Okazaki T, Nakano M, Saitoh K, *et al.* Construction of YAC-based mammalian artificial chromosomes. Nat Biotechnol. 1998; 16:431-439.

Irvine DV, Amor DJ, Perry J, Sirvent N, Pedeutour F, *et al.* Chromosome size and origin as determinants of the level of CENP-A incorporation into human centromeres. Chromosome Res. 2004; 12:805-815.

Italiano A, Maire G, Sirvent N, Nuin PA, Keslair F, *et al.* Variability of origin for the neocentromeric sequences in analphoid supernumerary marker chromosomes of well-differentiated liposarcomas. Cancer Lett. 2009; 273:323-330.

Izuta H, Ikeno M, Suzuki N, Tomonaga T, Nozaki N, *et al.* Comprehensive analysis of the ICEN (Interphase Centromere Complex) components enriched in the CENP-A chromatin of human cells. Genes Cells. 2006 11:673-684.

Jenuwein T, Allis CD. Translating the histone code. Science. 2001; 293:1074-1080.

Jolly C, Konecny L, Grady DL, Kutskova YA, Cotto JJ, *et al.* *In vivo* binding of active heat shock transcription factor 1 to human chromosome 9 heterochromatin during stress. J Cell Biol. 2002; 156:775-781.

Jolly C, Metz A, Govin J, Vigneron M, Turner BM, *et al.* Stress-induced transcription of satellite III repeats. J Cell Biol. 2004; 164:25-33.

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

Kanellopoulou C, Muljo SA, Kung AL, Ganesan S, Drapkin R, *et al.* *dicer*-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. Genes Dev. 2005; 19:489-501.

Kapoor M, Montes de Oca Luna R, Liu G, Lozano G, *et al.* The *cenpB* gene is not essential in mice. Chromosoma. 1998; 107:570-556.

Kato H, Jiang J, Zhou BR, Rozendaal M, Feng H, *et al.* A conserved mechanism for centromeric nucleosome recognition by centromere protein CENP-C. Science. 2013; 340:1110-1113.

Kazazian HH Jr. Mobile elements: drivers of genome evolution. Science. 2004; 303:1626-1632.

Kingwell B, Rattner JB. Mammalian kinetochore/centromere composition: a 50 kDa antigen is present in the mammalian kinetochore/centromere. Chromosoma. 1987; 95:403-407.

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

Kirsch-Volders M, Plas G, Elhajouji A, Lukamowicz M, Gonzalez L, *et al*. The in vitro MN assay in 2011: origin and fate, biological significance, protocols, high throughput methodologies and toxicological relevance. Arch Toxicol. 2011; 85:873-899.

Kitada K, Yamaguchi E, Arisawa M. Isolation of a *Candida glabrata* centromere and its use in construction of plasmid vectors. Gene. 1996; 175:105-108.

Komissarov AS, Gavrilova EV, Demin SJ, Ishov AM, Podgornaya OI. Tandemly repeated DNA families in the mouse genome. BMC Genomics. 2011; 12:531.

Kornberg RD. Chromatin structure: a repeating unit of histones and DNA. Science. 1974; 184:868-871.

Kwon MS, Hori T, Okada M, Fukagawa T. CENP-C is involved in chromosome segregation, mitotic checkpoint function, and kinetochore assembly. Mol Biol Cell. 2007; 18:2155-2168.

Lam AL, Boivin CD, Bonney CF, Rudd MK, Sullivan BA. Human centromeric chromatin is a dynamic chromosomal domain that can spread over noncentromeric DNA. Proc Natl Acad Sci USA. 2006; 103:4186-4191.

Larin Z, Fricker MD, Tyler-Smith C. *De novo* formation of several features of a centromere following introduction of a Y alphoid YAC into mammalian cells. Hum Mol Genet. 1994; 3:689-695.

Lehnertz B, Ueda Y, Derijck AA, Braunschweig U, Perez-Burgos L, *et al.* Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. Curr Biol. 2003; 13:1192-1200.

Li B, Choulet F, Heng Y, Hao W, Paux E, *et al.* Wheat centromeric retrotransposons: the new ones take a major role in centromeric structure. Plant J. 2013; 73:952-965.

Liu ST, Rattner JB, Jablonski SA, Yen TJ. Mapping the assembly pathways that specify formation of the trilaminar kinetochore plates in human cells. J Cell Biol. 2006; 175:41-53.

Lo AW, Craig JM, Saffery R, Kalitsis P, Irvine DV, *et al*. A 330 kb CENP-A binding domain and altered replication timing at a human neocentromere. EMBO J. 2001; 20:2087-2096.

Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, *et al.* Comparative and demographic analysis of orangutan genomes. Nature. 2011; 469:529-533.

Logsdon GA, Barrey EJ, Bassett EA, DeNizio JE, Guo LY, *et al.* Both tails and the centromere targeting domain of CENP-A are required for centromere establishment. J Cell Biol. 2015; 208:521-531.

Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature. 1997; 389:251-260.

Ma J, Jackson SA. Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. Genome Res. 2006; 16:251-259.

Ma J, Wing RA, Bennetzen JL, Jackson SA. Evolutionary history and positional shift of a rice centromere. Genetics. 2007; 177:1217-1220.

Macas J, Neumann P, Novák P, Jiang J. Global sequence characterization of rice centromeric satellite based on oligomer frequency analysis in large-scale sequencing data. Bioinformatics. 2010; 26:2101-2108.

Maida Y, Yasukawa M, Okamoto N, Ohka S, Kinoshita K, *et al.* Involvement of telomerase reverse transcriptase in heterochromatin maintenance. Mol Cell Biol. 2014; 34:1576-1593.

Malik HS, Henikoff S. Adaptive evolution of Cid, a centromere-specific histone in *Drosophila*. Genetics. 2001; 157:1293-1298.

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

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

Martins NM, Bergmann JH, Shono N, Kimura H, Larionov V, *et al.* Epigenetic engineering shows that a human centromere resists silencing mediated by 3K27me3/K9me3. Mol Biol Cell. 2016; 27:177-196.

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, Nakano M, Ohzeki J. The role of CENP-B and alpha-satellite DNA: de novo assembly and epigenetic maintenance of human centromeres. Chromosome Res. 2004; 12:543-556.

McClelland SE, Borusu S, Amaro AC, Winter JR, Belwal M, *et al.* The CENP-A NAC/CAD kinetochore complex controls chromosome congression and spindle bipolarity. EMBO J. 2007; 26:5033-5047.

McKinley KL, Cheeseman IM. The molecular basis for centromere identity and function. Nat Rev Mol Cell Biol. 2016; 17:16-29.

Mehta GD, Agarwal MP, Ghosh SK. Centromere identity: a challenge to be faced. Mol Genet Genomics. 2010; 284:75-94.

Mellone BG, Allshire RC. Stretching it: putting the CENP-A in centromere. Curr Opin Genet Dev. 2003; 13:191-198.

Melters DP, Bradnam KR, Young HA, Telis N, May MR, *et al.* Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol. 2013; 14:R10.

Mendiburo MJ, Padeken J, Fülöp S, Schepers A, Heun P. *Drosophila* CENH3 is sufficient for centromere formation. Science. 2011; 334:686-690.

Meštrović N, Pavlek M, Car A, Castagnone-Sereno P, Abad P, *et al.* Conserved DNA motifs, including the CENP-B Box-like, are possible promoters of satellite DNA array rearrangements in *Nematodes*. PLoS One. 2013; 8:e67328.

Miga KH. Completing the human genome: the progress and challenge of satellite DNA assembly. Chromosome Res. 2015; 23:421-426.

Milks KJ, Moree B, Straight AF. Dissection of CENP-C-directed centromere and kinetochore assembly. Mol Biol Cell. 2009; 20:4246-4255.

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.

Motamedi MR, Verdel A, Colmenares SU, Gerber SA, Gygi SP, *et al.* Two RNAi complexes, RITS and RDRC, physically interact and localize to noncoding centromeric RNAs. Cell. 2004; 119:789-802.

Mravinac B, Plohl M, Ugarković D. Preservation and high sequence conservation of satellite DNAs suggest functional constraints. J Mol Evol. 2005; 61:542-550.

Mravinac B, Plohl M. Parallelism in evolution of highly repetitive DNAs in sibling species. Mol Biol Evol. 2010; 27:1857-1867.

Murphy J, Armour J, Blais BW. Cloth-based hybridization array system for expanded identification of the animal species origin of derived materials in feeds. J Food Prot. 2007; 70:2900-2905.

Murphy TD, Karpen GH. Localization of centromere function in a *Drosophila* minichromosome. Cell. 1995; 82:599-609.

Nagaki K, Cheng Z, Ouyang S, Talbert PB, Kim M, *et al.* Sequencing of a rice centromere uncovers active genes. Nat. Genet. 2004; 36 138–145.

Nakano M, Cardinale S, Noskov VN, Gassmann R, Vagnarelli P, *et al.* Inactivation of a human kinetochore by specific targeting of chromatin modifiers. Dev Cell. 2008; 14:507-522.

Nergadze SG, Belloni E, Piras FM, Khoriauli L, Mazzagatti A, *et al.* Discovery and comparative analysis of a novel satellite, EC137, in horses and other equids. Cytogenet Genome Res. 2014; 144:114-123.

Nishihashi A, Haraguchi T, Hiraoka Y, Ikemura T, Regnier V, *et al.* CENP-I is essential for centromere function in vertebrate cells. Dev Cell. 2002; 2:463-476.

Ohkuma M, Kobayashi K, Kawai S, Hwang CW, Ohta A *et al.* Identification of a centromeric activity in the autonomously replicating TRA region allows improvement of the host-vector system for *Candida maltosa*. Mol Gen Genet. 1995; 249:447-455.

Ohkuni K, Kitagawa K. Endogenous transcription at the centromere facilitates centromere activity in budding yeast. Curr Biol. 2011; 21:1695-1703.

Ohzeki J, 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.

Ohzeki J, Larionov V, Earnshaw WC, Masumoto H. Genetic and epigenetic regulation of centromeres: a look at HAC formation. Chromosome Res. 2015; 23:87-103.

Okada M, Cheeseman IM, Hori T, Okawa K, McLeod IX, *et al.* The CENP-H-I complex is required for the efficient incorporation of newly synthesized CENP-A into centromeres. Nat Cell Biol. 2006; 8:446-457.

Okada M, Okawa K, Isobe T, Fukagawa T. CENP-H-containing complex facilitates centromere deposition of CENP-A in cooperation with FACT and CHD1. Mol Biol Cell. 2009; 20:3986-3995.

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

Olins AL, Olins DE. Spheroid chromatin units (v bodies). Science 1974; 183: 330–332.

Perez-Castro AV, Shamanski FL, Meneses JJ, Lovato TL, Vogel KG, *et al.* Centromeric protein B null mice are viable with no apparent abnormalities. Dev Biol. 1998; 201:135-143.

Perpelescu M, Fukagawa T. The ABCs of CENPs. Chromosoma. 2011; 120:425-446.

Peters AH, Kubicek S, Mechtler K, O'Sullivan RJ, Derijck AA, *et al.* Partitioning and plasticity of repressive histone methylation states in mammalian chromatin. Mol Cell. 2003; 12:1577-1589

Pimpinelli S, Berloco M, Fanti L, Dimitri P, Bonaccorsi S, *et al.* Transposable elements are stable structural components of *Drosophila melanogaster* heterochromatin. Proc Natl Acad Sci USA. 1995; 92:3804-3808.

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

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

Pitra C., Veits J. Use of mitochondrial DNA sequences to test the *Ceratomorpha* (*Perissodactyla*:Mammalia) hypothesis. J Zool Syst Evolv Res. 2000; 38:65Y72.

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.

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

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

Purgato S, Belloni E, Piras FM, Zoli M, Badiale C, *et al.* Centromere sliding on a mammalian chromosome. Chromosoma. 2015; 124:277-287.

Quénet D, Dalal Y. A long non-coding RNA is required for targeting centromeric protein A to the human centromere. Elife. 2014; 3:e03254.

Raimondi E, Piras FM, Nergadze SG, Di Meo GP, Ruiz-Herrera A, *et al.* Polymorphic organization of constitutive heterochromatin in *Equus asinus* (2n = 62) chromosome 1. Hereditas. 2011; 148:110-113.

Régnier V, Vagnarelli P, Fukagawa T, Zerjal T, Burns E, *et al.* CENP-A is required for accurate chromosome segregation and sustained kinetochore association of BubR1. Mol Cell Biol. 2005; 25:3967-3981.

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

Rice JC, Briggs SD, Ueberheide B, Barber CM, Shabanowitz J, *et al.* Histone methyltransferases direct different degrees of methylation to define distinct chromatin domains. Mol Cell. 2003; 12:1591-1598.

Rizzi N, Denegri M, Chiodi I, Corioni M, Valgardsdottir R, *et al.* Transcriptional activation of a constitutive heterochromatic domain of the human genome in response to heat shock. Mol Biol Cell. 2004; 15:543-551.

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

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

Rošić S, Köhler F, Erhardt S. Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division. J Cell Biol. 2014; 207:335-349.

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.

Ryder OA, Epel NC, Benirschke K. Chromosome banding studies of the *Equidae*. Cytogenet Cell Genet. 1978; 20:332-350.

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

Saffery R, Sumer H, Hassan S, Wong LH, Craig JM, *et al.* Transcription within a functional human centromere. Mol Cell. 2003; 12:509-516.

Saitoh H, Tomkiel J, Cooke CA, Ratrie H 3rd, Maurer M, *et al.* CENP-C, an autoantigen in scleroderma, is a component of the human inner kinetochore plate. Cell. 1992; 70:115-125.

Schneider R, Bannister AJ, Myers FA, Thorne AW, Crane-Robinson C, *et al.* Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. Nat Cell Biol. 2004; 6:73-77.

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

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

Sekulic N, Bassett EA, Rogers DJ, Black BE. The structure of (CENP-A-H4)(2) reveals physical features that mark centromeres. Nature. 2010; 467:347-351.

Shang WH, Hori T, Toyoda A, Kato J, Popendorf K, *et al.* Chickens possess centromeres with both extended tandem repeats and short non-tandem-repetitive sequences. Genome Res. 2010; 20:1219-1228.

Shang WH, Hori T, Martins NM, Toyoda A, Misu S, *et al.* Chromosome engineering allows the efficient isolation of vertebrate neocentromeres. Dev Cell. 2013; 24:635-648.

Shelby RD, Vafa O, Sullivan KF. Assembly of CENP-A into centromeric chromatin requires a cooperative array of nucleosomal DNA contact sites. J Cell Biol. 1997; 136:501-513.

Shumaker DK, Dechat T, Kohlmaier A, Adam SA, Bozovsky MR, *et al.* Mutant nuclear lamin A leads to progressive alterations of epigenetic control in premature aging. Proc Natl Acad Sci USA. 2006; 103:8703-8708.

Smit AF, Riggs AD. *Tiggers* and DNA transposon fossils in the human genome. Proc Natl Acad Sci USA. 1996; 93:1443-1448.

Smith KM, Phatale PA, Sullivan CM, Pomraning KR, Freitag M. Heterochromatin is required for normal distribution of *Neurospora crassa* CenH3. Mol Cell Biol. 2011; 31:2528-2542.

Springer MS, Murphy WJ, Eizirik E, O'Brien SJ. Placental mammal diversification and the Cretaceous-Tertiary boundary. Proc Natl Acad Sci USA. 2003; 100:1056-1061.

Steiner NC, Clarke L. A novel epigenetic effect can alter centromere function in fission yeast. Cell. 1994; 79:865-874.

Steiner FA, Henikoff S. Diversity in the organization of centromeric chromatin. Curr Opin Genet Dev. 2015; 31:28-35.

Stoyan T, Carbon J. Inner kinetochore of the pathogenic yeast *Candida glabrata*. Eukaryot Cell. 2004; 3:1154-1163.

Sugata N, Munekata E, Todokoro K. Characterization of a novel kinetochore protein, CENP-H. J Biol Chem. 1999; 274:27343-27346.

Sullivan BA, Willard HF. Stable dicentric X chromosomes with two functional centromeres. Nat Genet. 1998; 20:227-228.

Sullivan BA. Centromere round-up at the heterochromatin corral. Trends Biotechnol. 2002; 20:89-92.

Sullivan BA, Karpen GH. Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. Nat Struct Mol Biol. 2004; 11:1076-1083.

Sullivan KF, Hechenberger M, Masri K. Human CENP-A contains a histone H3 related histone fold domain that is required for targeting to the centromere. J Cell Biol. 1994; 127:581-592.

Sun X, Wahlstrom J, Karpen G. Molecular structure of a functional *Drosophila* centromere. Cell. 1997; 91:1007-1019.

Suzuki A, Hori T, Nishino T, Usukura J, Miyagi A, *et al*. Spindle microtubules generate tension-dependent changes in the distribution of inner kinetochore proteins. J Cell Biol. 2011; 193:125-140.

Taddei A, Maison C, Roche D, Almouzni G. Reversible disruption of pericentric heterochromatin and centromere function by inhibiting deacetylases. Nat Cell Biol. 2001; 3:114-120.

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

Talbert PB, Henikoff S. Phylogeny as the basis for naming histones. Trends Genet. 2013; 29:499-500.

Ting DT, Lipson D, Paul S, Brannigan BW, Akhavanfard S, *et al.* Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. Science. 2011; 331:593-596.

Tollis M, Boissinot S. The evolutionary dynamics of transposable elements in eukaryote genomes. Genome Dyn. 2012; 7:68-91.

Tomonaga T, Matsushita K, Yamaguchi S, Oohashi T, Shimada H, *et al.* Overexpression and mistargeting of centromere protein-A in human primary colorectal cancer. Cancer Res. 2003; 63:3511-3516.

Toth M, Grimsby J, Buzsaki G, Donovan GP. Epileptic seizures caused by inactivation of a novel gene, jerky, related to centromere binding protein-B in transgenic mice. Nat Genet. 1995; 11:71-75.

Tougard C, Delefosse T, Hänni C, Montgelard C. Phylogenetic relationships of the five extant Rhinoceros species (*Rhinocerotidae*, *Perissodactyla*) based on mitochondrial cytochrome b and 12S rRNA genes. Mol Phylogenet Evol. 2001; 19:34-44.

Trifonov VA, Stanyon R, Nesterenko AI, Fu B, Perelman PL, *et al.* Multidirectional cross-species painting illuminates the history of karyotypic evolution in *Perissodactyla*. Chromosome Res. 2008; 16:89-107.

Tyler-Smith C, Gimelli G, Giglio S, Floridia G, Pandya A, *et al.* Transmission of a fully functional human neocentromere through three generations. Am J Hum Genet. 1999; 64:1440-1444.

Vafa O, Sullivan KF. Chromatin containing CENP-A and alpha-satellite DNA is a major component of the inner kinetochore plate. Curr Biol. 1997; 7:897-900.

Valgardsdottir R, Chiodi I, Giordano M, Rossi A, Bazzini S, *et al.* Transcription of Satellite III non-coding RNAs is a general stress response in human cells. Nucleic Acids Res. 2008; 36:423-434.

Van Hooser AA, Ouspenski II, Gregson HC, Starr DA, Yen TJ, *et al.* Specification of kinetochore-forming chromatin by the histone H3 variant CENP-A. J Cell Sci. 2001; 114:3529-3542.

Ventura M, Archidiacono N, Rocchi M. Centromere emergence in evolution. Genome Res. 2001; 11:595-599.

Ventura M, Weigl S, Carbone L, Cardone MF, Misceo D, *et al.* Recurrent sites for new centromere seeding. Genome Res. 2004; 14:1696-1703.

Ventura M, Antonacci F, Cardone MF, Stanyon R, D'Addabbo P, *et al.* Evolutionary formation of new centromeres in macaque. Science. 2007; 316:243-246.

Verdel A, Jia S, Gerber S, Sugiyama T, Gygi S, *et al.* RNAi-mediated targeting of heterochromatin by the RITS complex. Science. 2004; 303:672-676.

Volpe TA, Kidner C, Hall IM, Teng G, Grewal SI, Martienssen RA. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. Science. 2002; 297:1833-1837.

Voullaire L, Slater HR, Petrovic V, Choo KH. 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.

Voullaire L, Saffery R, Davies J, Earle E, Kalitsis P, *et al.* Trisomy 20p resulting from inverted duplication and neocentromere formation. Am J Med Genet. 1999; 85:403-408.

Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, *et al.* Genome sequence, comparative analysis, and population genetics of the domestic horse. Science. 2009; 326:865-867.

Wan X, O'Quinn RP, Pierce HL, Joglekar AP, Gall WE, *et al.* Protein architecture of the human kinetochore microtubule attachment site. Cell. 2009; 137:672-684.

Warburton PE, Cooke CA, Bourassa S, Vafa O, Sullivan BA, *et al.* Immunolocalization of CENP-A suggests a distinct nucleosome structure at the inner kinetochore plate of active centromeres. Curr Biol. 1997; 7:901-904.

Warburton PE, Dolled M, Mahmood R, Alonso A, Li S, *et al.* Molecular cytogenetic analysis of eight inversion duplications of human chromosome 13q that each contain a neocentromere. Am J Hum Genet. 2000; 66:1794-1806.

Wolfgruber TK, Sharma A, Schneider KL, Albert PS, Koo DH, *et al.* Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic *loci* shaped primarily by retrotransposons. PLoS Genet. 2009; 5:e1000743.

Wong LH, Brettingham-Moore KH, Chan L, Quach JM, Anderson MA, *et al.* Centromere RNA is a key component for the assembly of nucleoproteins at the nucleolus and centromere. Genome Res. 2007; 17:1146-1160.

Xu X, Janke A, Arnason U. The complete mitochondrial DNA sequence of the greater *Indian rhinoceros*, *Rhinoceros unicornis*, and the Phylogenetic relationship among *Carnivora*, *Perissodactyla*, and *Artiodactyla* (+ *Cetacea*). Mol Biol Evol. 1996; 13:1167-1173.

Yang F, Fu B, O'Brien PC, Robinson TJ, Ryder OA, *et al.* Karyotypic relationships of horses and zebras: results of cross-species chromosome painting. Cytogenet Genome Res. 2003; 102:235-243.

Yang JW, Pendon C, Yang J, Haywood N, Chand A, *et al.* Human mini-chromosomes with minimal centromeres. Hum Mol Genet. 2000; 9:1891-1902.

Zeitlin SG, Baker NM, Chapados BR, Soutoglou E, Wang JY, *et al.* Double-strand DNA breaks recruit the centromeric histone CENP-A. Proc Natl Acad Sci USA. 2009; 106:15762-15767.

Zhu J, Tan Z, Hollis-Hansen K, Zhang Y, Yu C, *et al.* Epidemiological Trends in Colorectal Cancer in China: An Ecological Study. Dig Dis Sci. 2016; Epub ahead of print.

Zinić SD, Ugarković D, Cornudella L, Plohl M. A novel interspersed type of organization of satellite DNAs in *Tribolium madens* heterochromatin. Chromosome Res. 2000; 8:201-212.

LIST OF ORIGINAL MANUSCRIPTS

PEER REVIEW PAPERS

- 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.
  * Equal contributors
- Santagostino M, Khoriauli L, Gamba R, Bonuglia M, Klipstein O, Piras FM, Vella F, Russo A, Badiale C, **Mazzagatti A**, Raimondi E, Nergadze SG, Giulotto E. 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.
- 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.
- 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.

MEETING ABSTRACTS AND ORAL COMMUNICATIONS

- **Mazzagatti A**, Langella A, Roberti A, Bensi M, Piras FM, Cappelletti E, Gamba R, Giulotto E, Raimondi E. The epigenetic landscape of equid centromeres: a cytogenetic approach. FISV XIV Congress. Roma 2016 (oral communication).
- **Mazzagatti A**, Bensi M, Giulotto E, Raimondi E. Mitotic stability of a horse satellite-free centromere. Convegno AGI-SIMA. Cortona 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. SIBBM. Torino 2015.
- **Mazzagatti A**, Piras FM, Nergadze SG, Belloni E, Badiale C, Giulotto E, Raimondi E. The sliding behaviour of horse chromosome 11 centromere. FISV XIII Congress. Pisa 2014 (oral communication).
- Belloni E, Piras F, **Mazzagatti A**, Badiale C, Meinardi B, Castro A, Savini G, Cerutti F, Bensi M, Nergadze S, Raimondi E and Giulotto E. Analysis of the physical organization and of the CENP binding ability of horse centromeric satellite DNA families. Convegno AGI - Associazione Genetica Italiana. Cortona 2013.
- Santagostino M, Nergadze SG, Vella F, Klipstein O, Gamba R, Khoriauli L, **Mazzagatti A**, Raimondi E, Giulotto E. Genome wide analysis of ERE1 transposable elements in the horse and in other equids: evolutionary and functional aspects. 10th Dorothy Russel Havemeyer Foundation. Furnas, S. Miguel, Azores (Portugal) 2013.
- Raimondi E, Belloni E, Piras F, **Mazzagatti A**, Badiale C, Meinardi B, Castro A, Savini G, Bensi M, Nergadze S, Giulotto E. Physical organisation and CENP binding ability of horse centromeric satellite DNA families. 19th International Chromosome Conference, Bologna, 2013.

Original Article

Cytogenetic and Genome Research

Cytogenet Genome Res
DOI: 10.1159/000368138

Accepted: June 12, 2014
by M. Schmid
Published online: October 21, 2014

Discovery and Comparative Analysis of a Novel Satellite, EC137, in Horses and Other Equids

Solomon G. Nergadze Elisa Belloni Francesca M. Piras Lela Khoriauli Alice Mazzagatti Francesco Vella Mirella Bensi Valerio Vitelli Elena Giulotto Elena Raimondi

Department of Biology and Biotechnology 'L. Spallanzani', University of Pavia, Pavia, Italy

Key Words

Centromere · DNA combing · Equid species · Fiber FISH · Horse · Satellite DNA

Abstract

Centromeres are the sites of kinetochore assembly and spindle fiber attachment and consist of protein-DNA complexes in which the DNA component is typically characterized by the presence of extended arrays of tandem repeats called satellite DNA. Here, we describe the isolation and characterization of a 137-bp-long new satellite DNA sequence from the horse genome (EC137), which is also present, even if less abundant, in the domestic donkey, the Grevy's zebra and the Burchelli's zebra. We investigated the chromosomal distribution of the EC137 sequence in these 4 species. Moreover, we analyzed its architectural organization by high-resolution FISH. The position of this sequence with respect to the primary constriction and in relation to the 2 major horse satellite tandem repeats (37cen and 2PI) on horse chromosomes suggests that the new centromeric equine satellite is an accessory DNA element, presumably contributing to the organization of pericentromeric chromatin. FISH on combed DNA fibers reveals that the EC137 satellite is organized in relatively short stretches (2–8 kb) which are strictly intermingled within 37cen or 2PI arrays. This arrangement suggests that interchanges between satellite families are a frequent occurrence in the horse genome.

Centromeres are the functional elements controlling chromosome segregation during cell division. Vertebrate centromeres, which typically contain large amounts of tandem repeats (satellite DNA), are highly conserved for function, but not for DNA sequence. This observation, known as the 'centromere paradox', pointed to epigenetic factors as being responsible for centromere function through binding of the DNA to kinetochore proteins [Allshire and Karpen, 2008].

We previously isolated 2 centromeric satellite DNA sequences, 37cen and 2PI [Piras et al., 2010], from a horse genomic library in lambda phage [Anglana et al., 1996]. The 37cen sequence (GenBank: AY029358) is 93% identical to the horse major satellite family [Wijers et al., 1993; Sakagami et al., 1994], while the 2PI sequence (GenBank: AY029359S1 and AY029359S2) belongs to the e4/1 satellite family [Broad et al., 1995a, b] and shares 83% identity with it. We investigated the

KARGER

E-Mail karger@karger.com
www.karger.com/cgr

© 2014 S. Karger AG, Basel
1424–8581/14/0000–0000$39.50/0

Elena Raimondi
Department of Biology and Biotechnology "L. Spallanzani"
University of Pavia, Via Ferrata 9
IT–27100 Pavia (Italy)
E-Mail elena.raimondi@unipv.it

chromosomal distribution of these satellite tandem repeats in the horse (*Equus caballus*, ECA), the domestic donkey (*E. asinus*, EAS), the Grevy's zebra (*E. grevyi*, EGR), and the Burchelli's zebra (*E. burchelli*, EBU) and demonstrated that several centromeres lack satellite DNA at the FISH resolution level [Piras et al., 2010]. Moreover, we observed that satellite repeats are often present at non-centromeric termini, probably corresponding to relics of ancestral, now inactive, centromeres. It is worth noting that the centromere of horse chromosome 11 lacks any satellite DNA, as it was demonstrated in our previous horse genome sequencing work [Wade et al., 2009]. The absence of FISH-detectable 37cen and 2PI signals from the centromere of several domestic donkey, Grevy's zebra and Burchelli's zebra chromosomes [Piras et al., 2009, 2010] raises the question whether satellite DNA, belonging to other families, might be present at such centromeres. To investigate this possibility, we performed FISH analyses on the chromosomes of the 4 species, using their total genomic DNA as probe [Piras et al., 2010]. This is a procedure that allows the identification of regions containing very abundant tandem repeats due to the different hybridization kinetics of highly reiterated sequences versus single-copy DNA and revealed to be especially effective for the identification of satellite DNA in the *Equus* species, providing a resolution comparable to that of FISH performed with cloned satellite probes. As expected, in the horse, all the centromeres, except the one of chromosome 11, were labeled with specific signals. A faint interstitial signal, detectable only by hybridization with genomic DNA, was also present on the long arm of the X chromosome. Also in the domestic donkey, Grevy's zebra and Burchelli's zebra, the distribution of the FISH signals obtained using their total genomic DNA was not exactly comparable with that observed using the cloned horse satellite DNA probes. Actually, a few additional sites of hybridization were detected that were not visible with the 37cen and 2PI probes. These extra hybridization sites were localized in several centromeric regions, in a few telomeric regions and on the long arm of the donkey and Burchelli's zebra X chromosome. These results suggested the presence of tandem repeats other than the 2 major ones in these species.

Here, we describe the isolation and characterization of a new centromeric satellite DNA sequence from the horse genome, which is also present, even if less abundant, in the genomes of the domestic donkey, the Grevy's zebra and the Burchelli's zebra.

Materials and Methods

Cell Lines and Chromosome Preparation

Fibroblast cell lines from horse, domestic donkey, Grevy's zebra, and Burchelli's zebra were previously established [Piras et al., 2010]. Fibroblasts were cultured in Dulbecco's modified Eagle's medium (Euroclone), supplemented with 20% fetal calf serum (Euroclone), 2 mM glutamine, 2% non-essential amino acids, 1× penicillin/streptomycin. Cells were maintained at 37 °C in a humidified atmosphere of 5% CO_2. Mitoses were collected by directly blowing the medium on the dish surface, and metaphase spreads were prepared following the standard air-drying procedure.

EC137 Satellite Plasmid Vector Construction

Arrays of the EC137 repeat were amplified using oligonucleotides designed on the consensus sequence deduced from the comparison of 120 EC137 units. *Eco*RI and *Sal*I adapters were added for cloning purposes (5′-AAGAATTCTTGTGATGGAGGATGCAGTG-3′ and 5′-ATGGTCGACTGTGACACTGCATCCACTG-3′). PCR reactions were carried out with horse genomic DNA. PCR products were digested with *Eco*RI/*Sal*I and cloned into the pSVal plasmid vector [Nergadze et al., 2009]. A pSVal_137sat clone was sequenced and analyzed. It contains a 562-bp insert (about 4 copies of the EC137 repeat unit). To perform subsequent FISH analysis, we obtained a longer insert, comparable to those of the 2 other plasmids used in this work (37cen and 2PI, see below). Briefly, the *Bam*HI/*Xho*I-digested insert was extracted (QIAquick® Gel Extraction Kit; QIAGEN) and re-inserted next to the original copy using *Bam*HI/*Sal*I sites. This procedure was repeated 3 more times to obtain a new plasmid, pSVal_137sat16, containing 64 copies of the EC137 repeat unit, which was used for FISH and Southern blot analysis.

Southern Blot

For Southern blot, 7.5 μg of genomic DNA from each equid species was digested with *Bam*HI and *Eco*RI and separated on a 0.8% agarose gel. The gel was blotted onto a nylon membrane (Hybond N+, Amersham).

The membrane was then hybridized with the [α-^{32}P]dCTP-labeled EC137 probe. Hybridization was carried out overnight at 64 °C, and the final washing was performed in 0.1× SSC, 0.5% SDS.

FISH on Metaphase Chromosomes, Mechanically Stretched Chromosomes and Combed DNA Fibers

pSVal_137sat16, 37cen and 2PI plasmid DNAs were prepared using the Quantum Prep Plasmid miniprep kit (BioRad), according to the supplier's instructions. DNAs were labeled by nick translation with Cy3-dUTP (Perkin Elmer), Alexa488-dUTP (Invitrogen) or Cy5-dUTP (Perkin Elmer) and hybridized to metaphase spreads of primary fibroblasts from the 4 equid species and to horse mechanically stretched chromosomes. Alternatively, for hybridization to horse combed DNA, they were labeled with digoxigenin (DIG)-dUTP (Roche) or biotin-dUTP (Invitrogen).

Mechanically stretched chromosomes were prepared as previously described [Haaf and Ward, 1994]. Briefly, mitotically active cells were washed in PBS and resuspended in a hypotonic solution consisting of equal volumes of 75 mM KCl, 0.8% Na citrate and H_2O. The cell density in the hypotonic mixture was adjusted to 10^5 cells/ml. After 10 min of hypotonic treatment at room temperature, the cell suspension was cytocentrifuged onto glass slides at 80 *g* for 4 min and fixed in −20 °C methanol for 30 min.

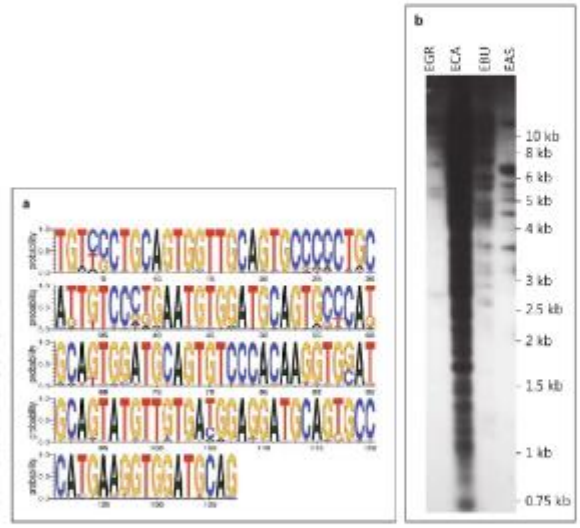

Fig. 1. a Consensus sequence of the EC137 satellite. The height of the letters at each position is proportional to the frequency of the corresponding nucleotide. The overall height of the stack is measured in bits (y axis), a numerical value proportional to the degree of conservation of the sequence. For each position (x axis), the nucleotides are stacked starting from the most frequent at the top. **b** Southern blot analysis of the EC137 sequence in Grevy's zebra (EGR), horse (ECA), Burchelli's zebra (EBU), and donkey (EAS).

The protocol used to prepare combed DNA was slightly modified from Michalet et al. [1997]. High molecular weight total horse genomic DNA was diluted in 150 mM MES at a concentration of 2 μg/ml and transferred in a reservoir. Silanized slides were introduced into the reservoir, incubated in the solution for 5 min and then extracted at a constant speed. Before use, combed DNA slides were dried at 60°C overnight.

The protocol used for FISH was essentially the same for metaphase chromosomes, stretched chromosomes and chromatin fibers. For each slide, 250 ng of each satellite probe was used. The cytological preparations were denatured at 75°C in 70% formamide, 2× SSC and immediately incubated with the probe for 16–18 h at 37°C in high-stringency conditions with 50% formamide. Post-hybridization washes were performed in 50% formamide, 2× SSC at 42°C. Biotinylated and DIG-labeled probes were detected with FITC and rhodamine, respectively, using 5 successive layers of antibodies as follows: (i) anti-DIG rhodamine (sheep) 1:200 (Roche); (ii) anti-sheep rhodamine 1:100 (CHEMICON); (iii) ExtrAvidin-FITC 1:100 (Sigma-Aldrich); (iv) biotinylated anti-avidin 1:200 (Sigma-Aldrich); (v) ExtrAvidin-FITC 1:100. All antibody incubations were for 30 min at 37°C. All antibody washes were for 3 × 5 min using 4× SSC, 0.1% Tween 20. Chromosomes were counterstained with DAPI. All the slides were mounted in Dako mounting medium (DAKO). Digital grey-scale images for Cy3, Alexa488, Cy5, rhodamine, FITC, and DAPI 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 published karyotypes [Yang et al., 2003, 2004; Musilova et al., 2007]. At least 20 metaphase spreads were analyzed for each species.

Results and Discussion

Cloning and Characterization of a New Horse Satellite Sequence

We identified a new satellite DNA sequence from the horse genome database (EquCab2.0, http://www.ncbi.nlm.nih.gov/projects/mapview/map_search.cgi?taxid=9796) using available analysis tools (BLAST, BLAT, Tandem Repeat Finder, RepeatMasker, MultAlin). We aligned 120 repeat units of this new satellite and deduced a 137-bp-long consensus sequence [GenBank: JX026961] (fig. 1a). Then, we compared the consensus sequence with that of the major horse satellite DNA families. Sequence analysis demonstrated that the EC137 consensus se-

DOI: 10.1159/000368138

quence has neither significant similarity with known *Equus* satellites, nor with Equidae repetitive sequences deposited in RepBase [Wijers et al., 1993; Sakagami et al., 1994; Broad et al., 1995a, b; Piras et al., 2010; Alkan et al., 2011]. Moreover, comparative in silico analysis demonstrated that the new satellite is not conserved in 9 mammalian species (dog, cat, guinea pig, rat, mouse, rhesus, orangutan, chimpanzee, and human); therefore, this is a new, equid-specific DNA element.

BLAST search on the published horse genome sequence (EquCab2.0) mapped the EC137 at the centromeric region of horse chromosomes 1, 2, 15, 20, 25, 28, and on the long arm of the X. Using primers designed on the EC137 consensus sequence, adapted with *Eco*RI and *Sal*I restriction sites for cloning purposes (see Materials and Methods), we amplified horse genomic DNA by PCR. *Eco*RI/*Sal*I-digested PCR products were cloned in the cloning site of a plasmid vector that we previously constructed [Nergadze et al., 2009]. One of the clones obtained was sequenced and analyzed. It contained a 562-bp insert with 4 copies of the EC137 repeat unit. This plasmid clone was modified to obtain a new plasmid containing 64 copies of the EC137 repeat unit (see Materials and Methods) to be used in the hybridization experiments described hereafter.

In order to study the presence and relative abundance of this satellite DNA family in equid species, we performed Southern blot experiments on total genomic DNA from horse, domestic donkey, Grevy's zebra, and Burchelli's zebra. The results of this experiment are shown in figure 1b. The EC137 sequence was found in all the species analyzed, but its relative abundance was markedly different. Namely, in the horse this repeat was much more abundant than in the other species; the domestic donkey and the Grevy's zebra having the lowest amount. The wide variability of satellite DNA quantity among related species is a well known feature characterizing all highly reiterated sequences [Smith, 1976; Dover, 1982; Plohl et al., 2008; Pertile et al., 2009; Shi et al., 2010]; this may be particularly true for equid species which have an exceptionally high evolutionary rate [Trifonov et al., 2008]. Wichman et al. [1991] examined 4 classes of tandemly repeated elements in 6 equid species and proposed that satellite DNA may be a driving force in chromosomal evolution. This observation also suggested that species with conserved karyotypes have a lower number of rapidly evolving DNA families than those with a karyotype that has undergone rapid evolution [Bradley and Wichman, 1994]. The exceptional heterogeneity that we observed among equid species in the amount and distribution of the EC137 satellite (see following section) and of the main satellites [Piras et al., 2010] is in agreement with the observations of Wichman et al. [1991].

Chromosomal Distribution of EC137 in the 4 Equus Species

Also the chromosomal distribution of the EC137 satellite was very different in the 4 species analyzed here and did not coincide with that of the 2 major equid satellite DNA families (37cen and 2PI) that we previously described [Piras et al., 2010]. Using the EC137 satellite DNA as probe for FISH, we localized it on metaphase chromosomes from horse (fig. 2a), donkey (fig. 2b), Grevy's zebra (fig. 2c) and Burchelli's zebra (fig. 2d).

In the horse (fig. 2a), the centromeric regions of 13 chromosome pairs were labeled (ECA1, 7, 8, 10, 14, 15, 20, 23, 26–30). The FISH signal observed on ECA15 was relatively faint, and its localization was subcentromeric. We previously demonstrated that ECA11 is devoid of 37cen and 2PI satellite DNA families [Piras et al., 2010]; this chromosome was also negative for EC137 hybridization.

We observed a discrepancy in the localization of the EC137 satellite revealed by FISH and that inferred by sequence alignment from the horse genome database. Indeed, FISH analysis could not reveal any signal on ECA-2cen and ECA25cen nor on the long arm of ECAX, which were EC137-positive in BLAST search, whilst a number of centromeres not detected by BLAST search were FISH-positive (ECA7, 8, 10, 14, 23, 26, 27, 29, 30). This apparent contradiction was expected due to 2 factors: (1) FISH allows the identification of repetitive sequences which are located within gaps of the horse genome database sequence; (2) short stretches of reiterated sequences, under the resolution limit of FISH, can be detected only by in silico analysis.

In the domestic donkey, only chromosomes 1 and 2 were FISH-labeled (fig. 2b). However, while the fluorescence signal on EAS2 was clearly centromeric, the one observed on EAS1 was located in an interstitial position in the short arm. In a previous paper [Raimondi et al., 2011], aimed at describing heterochromatin polymorphism, we investigated in detail the distribution of the 2 major equine satellite DNA families (37cen and 2PI) on EAS1. On the short arm of EAS1, 3 hybridization sites were found, none of these coinciding with the cytological position of the EC137 satellite that we observed in the present paper. Our previous results [Piras et al., 2010] demonstrated that, in the donkey, 21 centromeres were

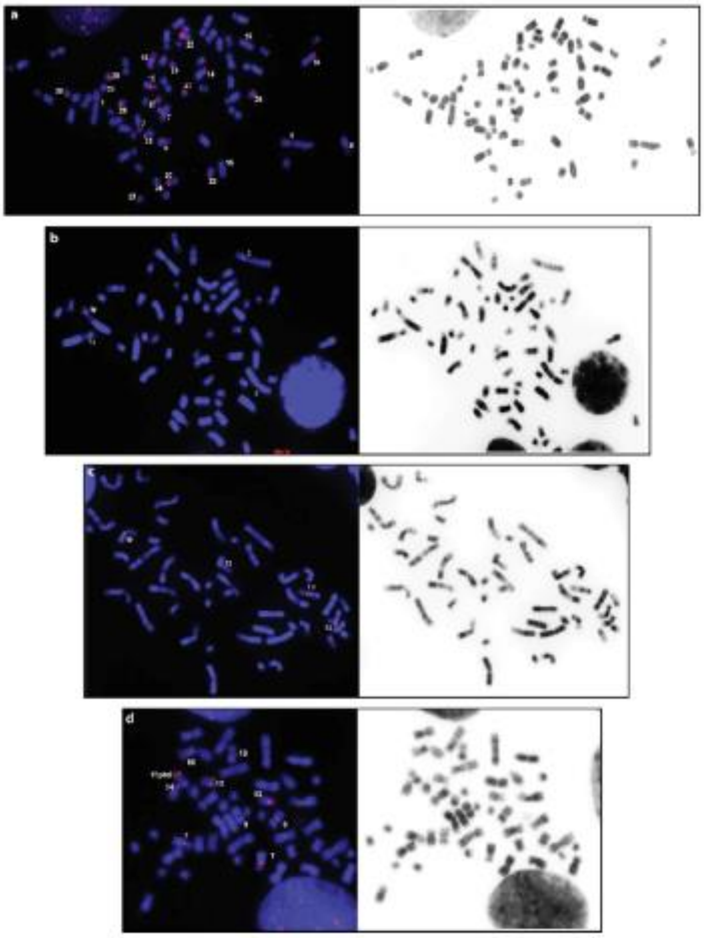

Fig. 2. Localization of the EC137 satellite by FISH on metaphase chromosomes from horse (**a**), domestic donkey (**b**), Grevy's zebra (**c**), and Burchelli's zebra (**d**). For each species, at least 20 metaphase spreads were analyzed. FISH signals are red labeled, while the chromosomes have been counterstained with DAPI (blue) (left panels). The FISH-positive chromosomes were identified by computer-generated reverse DAPI banding (right panels).

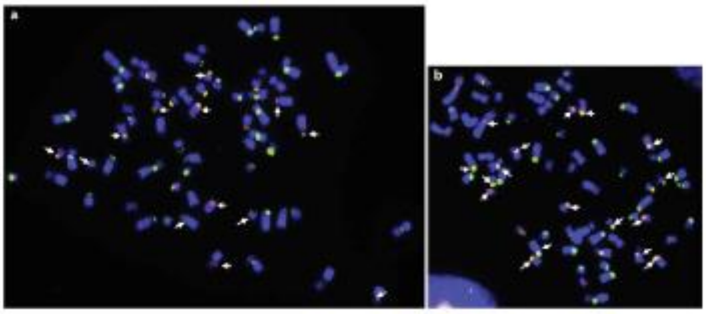

Fig. 3. Two-color FISH with the EC137 (red) and 37cen (green) satellites (**a**) and EC137 (red) and 2PI (green) satellites (**b**) on horse metaphase chromosomes counterstained with DAPI (blue).

negative for both 37cen and 2PI hybridization; none of these was labeled when probed with the EC137 satellite.

In the Grevy's zebra, the FISH signals were located in the centromeric region of 2 chromosomes only, namely EGR10 and EGR12 (fig. 2c). As we previously demonstrated [Piras et al., 2010], 17 Grevy's zebra centromeres are 37cen- and 2PI-negative; among these, EGR10cen was the only EC137-positive one.

In the Burchelli's zebra, in spite of the evolutionary proximity with Grevy's zebra, 5 centromeres were labeled: EBU7cen, EBU9cen, EBU10cen, EBU13cen, and EBU14cen (fig. 2d). In all the analyzed metaphase spreads (total number 25), only one of the homologous chromosomes 14 was FISH-positive (fig. 2d), suggesting a polymorphic variation in the number of EC137 repeats. In addition, the telomeric end of only one of the EBU11 homologs was labeled (fig. 2d) – an indication of likely polymorphism. In the Burchelli's zebra, 12 centromeres were demonstrated to be 37cen- and 2PI-negative [Piras et al., 2010]. Among these, only the centromere of EBU13 was EC137-positive. The polymorphic nature of the FISH signals observed on EBU14cen and on the EBU11 short arm terminus is not surprising due to the well-documented intra- and interspecific variability in the amount and distribution of satellite DNA sequences [Plohl et al., 2008].

Notwithstanding, our previous data [Piras et al., 2010] demonstrated that the 37cen sequence is largely represented only in the horse, being less abundant in the donkey and underrepresented in the zebras that, on the contrary, have the 2PI sequence as the most-represented satellite. Moreover, in the donkey and in the zebras, a number of centromeres are not labeled, neither with the 37cen, nor with the 2PI satellite. Here, we demonstrated that among donkey and zebra centromeres lacking both the major satellite DNA sequences, only 2 (EGR10 and EBU13) were EC137-positive. The present results confirm that besides ECA11cen, whose satellite-less nature was unequivocally demonstrated by sequence data [Wade et al., 2009], a number of centromeres of domestic donkey, Grevy's zebra and Burchelli's zebra chromosomes are apparently devoid of centromeric satellite DNA when analyzed by FISH. We cannot rule out that other centromeric satellite DNA sequences, whose search is hampered by the fact that the genomes of the ass and of the zebras have not been sequenced to date, are present in these species. However, our previous data [Piras et al., 2010], combined with the results reported here, suggest that some donkey and zebra centromeres are bona fide new examples of satellite-free centromeres, thus making the equid species a unique model system to study centromere function and evolution.

Distribution of EC137 Relative to the 2 Major Horse Satellites on Metaphase Chromosomes

To investigate the relations in the distribution of the various classes of equid satellite DNA sequences, we performed 2-color FISH experiments on horse metaphase chromosomes (fig. 3). The results of the co-hybridization of the 37cen and EC137 satellites are shown in figure 3a. Besides being located on a different number of centro-

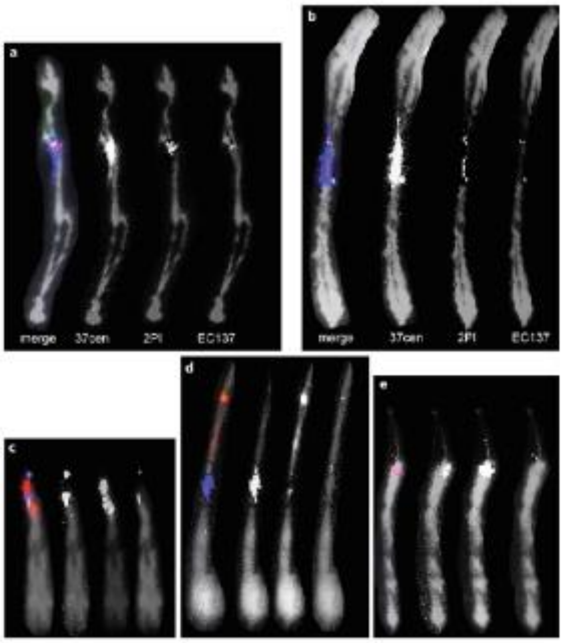

Fig. 4. Three-color FISH on horse mechanically stretched chromosomes with the 37cen, 2PI and EC137 satellites. **a–e** Five organizational patterns of the satellite sequences were found, each represented by 1 sample chromosome here. In each panel, the chromosome on the left represents the merged image, where the chromosome is grey colored, the 37cen probe is blue, the 2PI satellite is red, and the EC137 sequence is green. The separated color channel images of the same chromosome are shown in the adjacent pictures.

meres (the 37cen sequence is present on all the centromeres except those of ECA2 and ECA11, while the EC137 satellite is located only on 26 out of 64 centromeres), the 2 satellite DNA sequences seem to occupy different positions with respect to the primary constriction. Indeed, the 37cen signal always coincides with the primary constriction and, when it is very large, spreads in the pericentromere. On the contrary, the EC137 signal is mostly pericentromeric, with no or limited overlap with 37cen (arrows in fig. 3a). The scenario changes when the location of the EC137 satellite is compared with that of the 2PI satellite (fig. 3b). The majority of the chromosomes sharing the 2PI and the EC137 sequences are yellow labeled (arrows in fig. 3b), the yellow signal demonstrating the overlap of the green (2PI) and red (EC137) fluorescence signals arising from the single probes. The present data, combined with our previous results [Piras et al., 2010], demonstrate that 12 chromosomes carry all 3 satellite DNA sequences (at the FISH resolution level), namely: ECA7, 8, 10, 14, 15, 20, 23, 26–30. Moreover, the 2-color FISH results suggest that, in the horse, the 37cen satellite may be a functional centromeric satellite, while the 2PI and the EC137 sequences may represent accessory pericentromeric elements.

High-Resolution Analysis of Horse Satellite DNA Organization

In an attempt to better define the physical relations between the different satellite DNA families on horse centromeres, 3-color FISH experiments were performed on mechanically stretched chromosomes (fig. 4). A total number of 89 stretched chromosomes were analyzed, on 37 of which all the 3 hybridization signals were present. Five different patterns of physical organization of the satellite sequences were observed (fig. 4a–e). In each panel, the merged image is shown on the left. In the other images, the separate color channels are reported, corresponding to the 37cen probe (blue in the left image), to the 2PI probe (red in the left image) and to the EC137 probe (green in the left image), respectively.

In the majority of the stretched chromosomes analyzed (17 out of 37; 46%), the 37cen sequence covered the whole primary constriction while the 2PI and the EC137 satellites were less abundant and colocalized in the distal portion of the 37cen-positive region, the EC137 satellite being the least represented (fig. 4a).

Seven out of 37 chromosomes (19%) displayed the pattern shown in figure 4b. Again, the 37cen satellite was largely overrepresented, the 2PI sequence was spread along the 37cen-positive region and the EC137 satellite was underrepresented and localized in different positions within the 37cen- and 2PI-positive region.

In 7 out of 37 instances (19%), we observed the organization shown in figure 4c: the 37cen and the 2PI sequences were both very abundant, while the EC137 sequence was extremely scanty and interspersed within the other satellites.

The chromosome shown in figure 4d exemplifies the fourth pattern observed in 4 out of 37 cases (11%): all these chromosomes were metacentric or submetacentric, and the 2PI sequence spread in an uncoiled pericentromeric region which was 37cen- and EC137-negative.

The last pattern (fig. 4e) was observed in 2 out of 37 chromosomes (5%). Both these chromosomes were acrocentric and, peculiarly, centromeric chromatin formed uncoiled extensions protruding out of the main chromosome body. These protruding fibers were 2PI-positive and 37cen- and EC137-negative.

These observations confirm that the 37cen sequence may play a role in centromere function, while 2PI and EC137 may represent accessory elements. An important point is that 1 horse centromere (ECA2cen) is apparently free of the 37cen, the 2PI sequence being the only satellite that we could detect by FISH. This suggests that, at least in this case, 2PI could be able to drive kinetochore assembly. Still, it must be taken into account that a very small amount of the 37cen satellite, under the resolution level of FISH analysis, may be present. Moreover, we want to stress that in the horse (ECA11cen), and presumably in the domestic donkey and in the zebras, satellite-free centromeres exist [Wade et al., 2009; Piras et al., 2009, 2010]; therefore, it cannot be ruled out that key single copy sequences could play a role in centromere function also in some satellite-positive centromeres.

We then performed 2-color FISH on DNA fibers prepared by DNA combing (fig. 5). DNA combing allows obtaining DNA fibers with a controlled and uniform degree of elongation; thereafter, quantitative estimates can be performed. In order to know the average degree of DNA fiber extension, we set up a molecular ruler. To this purpose, we performed parallel FISH experiments on combed DNA using a horse BAC clone of known length as FISH probe. In this way we could relate the length of the hybridization signals, measured in centimeters on digital images, with the length in base pairs of the hybridized target sequence. We estimated that 1 cm, measured on digital images taken with a constant, arbitrarily fixed, magnification, roughly corresponded to 17 kb (white bars in fig. 5). It can be noted that in the merged images shown in figure 5a–c some hybridization signals appear as yellow fluorescence or as overlapping green and red fluorescence, indicating high interspersion of satellite sequences on a small scale. Overlapping signals (yellow) probably represent DNA regions in which the satellite DNA arrays are so closely intermingled to be undistinguishable. Within such regions, the resolution of the method is not sufficient to establish the order of DNA sequences. The 37cen satellite (fig. 5a, c) covers very long continuous regions, extending for hundreds of kilobases (occupying more than one microscopic field). However, when the 2PI (fig. 5a) or EC137 (fig. 5c) satellites are present, these form small stretches (2–8 kb) that are strictly interspersed within the 37cen clusters. The 2PI or EC137 stretches occur every 6–80 kb in the large 37cen blocks. Figure 5b shows the results of 2-color FISH on combed DNA hybridized with 2PI and EC137. These 2 satellite DNA sequences are both organized in small stretches (2–8 kb) which are strictly intermingled. The 2PI satellite appears to be more abundant than EC137. The overall organization of the different classes of horse satellite DNA appears to be a mosaic where the 3 DNA families display an interspersed association of sequence blocks widely variable in size. This organization resembles that described by Zinić et al. [2000] for the 2 centromeric satellite DNA families found in *Tribolium madens* (Insecta, Coleoptera). The authors propose that such an organizational pattern of DNA sequences in heterochromatin might be common in genomes characterized by a high rate of interchromosomal exchange. This hypothesis may well explain what we described here for the horse centromeric satellites: indeed, it is well documented that the genome of the species belonging to the genus *Equus* is exceptionally plastic if compared to that of other mammals [Ryder et al., 1978; Yang et al., 2003, 2004; Trifonov et al., 2008; Piras et al., 2010].

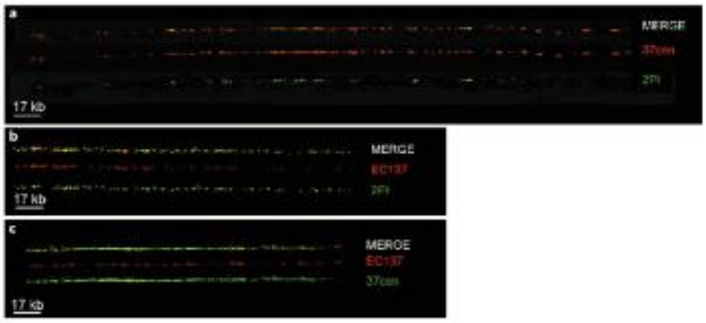

Fig. 5. Two-color FISH on horse combed DNA fibers. **a** 37cen (red) and 2PI (green) satellites. **b** EC137 (red) and 2PI (green) satellites. **c** EC137 (red) and 37cen (green) satellites.

Conclusions

Satellite DNA is organized as stretches of tandem repeats, located in the constitutive heterochromatin. Long arrays of satellite DNA are present at the primary constriction of most eukaryotic chromosomes. However, centromere DNA array length is highly variable, both among homologous and heterologous centromeres. Moreover, centromeric DNA sequences are rapidly evolving among species and within chromosomes of the same species [Plohl et al., 2008]. Surprisingly, satellite DNA is not necessary for centromere function. Actually, satellite DNA-free centromeres have been found in human pathology (neocentromeres) [Voullaire et al., 1993; Marshall et al., 2008] and in extant species (evolutionarily new centromeres). To date, such satellite DNA-free centromeres have been reported in horse, orangutan and chicken [Wade et al., 2009; Shang et al., 2010; Locke et al., 2011]. Equid species are exceptional in this regard. We previously demonstrated [Piras et al., 2010] that apparently satellite-free evolutionarily new centromeres are a common finding in the domestic donkey and in zebras. Moreover, satellite DNA can be present in a non-centromeric position as the fossil relic of an ancient, now repositioned, centromere. In the same study, we gave evidence that other families of centromeric satellite DNA, to date not described, are present at the centromere of some equid chromosomes. In the present paper, we describe a new family of equid pericentromeric satellite DNA, named EC137, the chromosomal distribution of which is defined in detail and compared with that of the 2 major classes of equid satellite DNA.

Our data suggest that the new centromeric horse satellite DNA sequence is an accessory DNA element, presumably contributing to the organization of pericentromeric chromatin. In the domestic donkey, the Grevy's zebra and the Burchelli's zebra, this satellite is much less abundant than in the horse, and only in 2 cases (EGR10cen and EBU13cen), it is present in centromeres apparently free of the 37cen and 2PI satellites.

In all the equid species analyzed here, the amount of the centromeric EC137 satellite is highly variable among chromosomes. Moreover, closely related species, such as the 2 zebras, show a strikingly different pattern of EC137 chromosomal distribution. This observation supports the hypothesis that highly variable satellite DNA may contribute to karyotype evolution in rapidly evolving species such as equids [Wichman et al., 1991; Bradley and Wichman, 1994].

High-resolution FISH analysis on combed DNA fibers demonstrated that satellite DNA clusters at horse centromeres show a peculiar architectural organization, where small arrays of 2PI and EC137 satellites are closely intermingled and immerged within very large stretches of the 37cen sequence. This organization should be the consequence of recombination events among pericentromeric regions containing different types of satel-

DOI: 10.1159/000368138

lites. This observation allows us to hypothesize that, at horse chromosome centromeres, satellite sequence interchanges are a frequent occurrence. This hypothesis is in agreement with the highly plastic nature of equid genomes.

Acknowledgements

This work was supported by grants from the Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR), PRIN and from the Consiglio Nazionale delle Ricerche (CNR), Progetto Bandiera 'Epigenomica'.

References

Alkan C, Cardone MF, Catacchio CR, Antonacci F, O'Brien SJ, et al: Genome-wide characterization of centromeric satellites from multiple mammalian genomes. Genome Res 21:137–145 (2011).

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

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

Bradley RD, Wichman HA: Rapidly evolving repetitive DNAs in a conservative genome: a test of factors that affect chromosomal evolution. Chromosome Res 2:354–360 (1994).

Broad TE, Lde AJ, Forrest JW, Lewis PE, Phua SH, Pugh PA: Families of tandemly repeated DNA elements from horse: cloning, nucleotide sequence, and organization. Genome 38:1285–1289 (1995a).

Broad TE, Forrest JW, Lewis PE, Pearce PD, Phua SH, et al: Cloning of a DNA repeat element from horse: DNA sequence and chromosomal localization. Genome 38:1132–1138 (1995b).

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

Haaf T, Ward DC: High resolution ordering of YAC contigs using extended chromatin and chromosomes. Hum Mol Genet 3:629–633 (1994).

Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, et al: Comparative and demographic analysis of orang-utan genomes. Nature 469:529–533 (2011).

Marshall OJ, Chueh CA, Wong LH, Choo KH: Neocentromeres: new insights into centromere structure, disease development, and karyotype evolution. Am J Hum Genet 82:261–282 (2008).

Michalet X, Ekong R, Fougerousse F, Rousseaux S, Schurra C, et al: Dynamic molecular combing: stretching the whole human genome for high-resolution studies. Science 277:1518–1523 (1997).

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 15:807–813 (2007).

Nergadze SG, Farnung BO, Wischnewski H, Khoriauli L, Vitelli V, et al: CpG-island promoters drive transcription of human telomeres. RNA 15:2186–2194 (2009).

Pertile MD, Graham AN, Choo KHA, Kalitsis P: Rapid evolution of mouse Y centromere repeat DNA belies recent sequence stability. Genome Res 19:2202–2213 (2009).

Piras FM, Nergadze SG, Poletto V, Cerutti F, Ryder OA, et al: Phylogeny of horse chromosome 5q in the genus *Equus* and centromere repositioning. Cytogenet Genome Res 126:165–172 (2009).

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

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

Raimondi E, Piras FM, Nergadze SG, Di Meo GP, Ruiz-Herrera A, et al: Polymorphic organization of constitutive heterochromatin in *Equus asinus* (2n = 62) chromosome 1. Hereditas 148:110–113 (2011).

Ryder OA, Epel NC, Benirschke K: Chromosome banding studies of the Equidae. Cytogenet Cell Genet 20:323–350 (1978).

Sakagami M, Hirota K, Awata T, Yasue H: Molecular cloning of an equine satellite-type DNA sequence and its chromosomal localization. Cytogenet Cell Genet 66:27–30 (1994).

Shang WH, Hori T, Toyoda A, Kato J, Popendorf K, et al: Chickens possess centromeres with both extended tandem repeats and short non-tandem-repetitive sequences. Genome Res 20:1219–1228 (2010).

Shi J, Wolf SE, Burke JM, Presting GG, Ross-Ibarra J, Dawe RK: Widespread gene conversion in centromere cores. PLoS Biol 8:e1000327 (2010).

Smith GP: Evolution of repeated DNA sequences by unequal crossover. Science 191:528–535 (1976).

Trifonov VA, Stanyon R, Nesterenko AI, Fu B, Perelman PL: Multidirectional cross-species painting illuminates the history of karyotypic evolution in Perissodactyla. Chromosome Res 16:89–107 (2008).

Voullaire LE, Slater HR, Petrovic V, Choo KH: 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 (1993).

Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, et al: Genome sequence, comparative analysis and population genetics of the domestic horse (*Equus caballus*). Science 326:865–867 (2009).

Wichman HA, Payne CT, Ryder OA, Hamilton MJ, Maltbie M, Baker RJ: Genomic distribution of heterochromatic sequences in equids: implications to rapid chromosomal evolution. J Hered 82:369–377 (1991).

Wijers ER, Zijlstra C, Lenstra JA: Rapid evolution of horse satellite DNA. Genomics 18:113–117 (1993).

Yang F, Fu B, O'Brien PC, Robinson TJ, Ryder OA, Ferguson-Smith MA: Karyotypic relationships of horses and zebras: results of cross-species chromosome painting. Cytogenet Genome Res 102:235–243 (2003).

Yang F, Fu B, O'Brien PC, Nie W, Ryder OA, Ferguson-Smith MA: 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 (2004).

Žinić SD, Ugarković D, Cornudella L, Plohl M: A novel interspersed type of organization of satellite DNAs in *Tribolium madens* heterochromatin. Chromosome Res 8:201–212 (2000).

Chromosoma (2015) 124:277–287
DOI 10.1007/s00412-014-0493-6

RESEARCH ARTICLE

Centromere sliding on a mammalian chromosome

Stefania Purgato · Elisa Belloni · Francesca M. Piras · Monica Zoli · Claudia Badiale · Federico Cerutti · Alice Mazzagatti · Giovanni Perini · Giuliano Della Valle · Solomon G. Nergadze · Kevin F. Sullivan · Elena Raimondi · Mariano Rocchi · Elena Giulotto

Received: 11 September 2014 / Revised: 31 October 2014 / Accepted: 3 November 2014 / Published online: 21 November 2014

Abstract The centromere directs the segregation of chromosomes during mitosis and meiosis. It is a distinct genetic locus whose identity is established through epigenetic mechanisms that depend on the deposition of centromere-specific centromere protein A (CENP-A) nucleosomes. This important chromatin domain has so far escaped comprehensive molecular analysis due to its typical association with highly repetitive satellite DNA. In previous work, we discovered that the centromere of horse chromosome 11 is completely devoid of satellite DNA; this peculiar feature makes it a unique model to dissect the molecular architecture of mammalian centromeres. Here, we exploited this native satellite-free centromere to determine the precise localization of its functional domains in five individuals: We hybridized DNA purified from chromatin immunoprecipitated with an anti CENP-A antibody to a high resolution array (ChIP-on-chip) of the region containing the primary constriction of horse chromosome 11. Strikingly, each individual exhibited a different arrangement of CENP-A binding domains. We then analysed the organization of each domain using a single nucleotide polymorphism (SNP)-based approach and single molecule analysis on chromatin fibres. Examination of the ten instances of chromosome 11 in the five individuals revealed seven distinct 'positional alleles', each one extending for about 80–160 kb, were found across a region of about 500 kb. Our results demonstrate that CENP-A binding domains are autonomous relative to the underlying DNA sequence and are characterized by positional instability causing the sliding of centromere position. We propose that this dynamic behaviour may be common in mammalian centromeres and may determine the establishment of epigenetic alleles.

Introduction

Centromeres are genetic loci whose identity depends not on the sequence of DNA on which they are formed but on a specific nucleosome configuration containing the centromere-specific histone H3, centromere protein A (CENP-A) (Sullivan 2001; Black and Cleveland 2011). Centromere-associated DNA varies widely in different species and even within a karyotype, but the core protein composition, based on the presence of CENP-A nucleosomes, is a universal feature of eukaryotic chromosomes (Malik and Henikoff 2009). Both CENP-A and its deposition machinery, comprising a distinct pathway for chromatin assembly, are highly conserved during evolution (Maddox et al. 2012; Kato et al. 2013). Precisely how this chromatin architecture is related to its underlying DNA is still poorly understood. Typically, mammalian centromeres are associated with highly repetitive tandem satellite arrays which have limited the detailed molecular dissection of this critical chromatin domain (Karpen and Allshire 1997). Taking advantage of the presence of two alpha satellite subfamilies at the centromere of human chromosome 17,

Electronic supplementary material The online version of this article (doi:10.1007/s00412-014-0493-6) contains supplementary material, which is available to authorized users.

S. Purgato · M. Zoli · G. Perini · G. Della Valle
Dipartimento di Farmacia e Biotecnologie (FABIT), Università di Bologna, Bologna, Italy

E. Belloni · F. M. Piras · C. Badiale · F. Cerutti · A. Mazzagatti · S. G. Nergadze · E. Raimondi (✉) · E. Giulotto (✉)
Dipartimento di Biologia e Biotecnologie "Lazzaro Spallanzani", Università di Pavia, Pavia, Italy
e-mail: elena.raimondi@unipv.it
e-mail: elena.giulotto@unipv.it

K. F. Sullivan
Centre for Chromosome Biology, School of Natural Sciences, National University of Ireland, Galway, Ireland

M. Rocchi (✉)
Dipartimento di Biologia, Università di Bari, Bari, Italy
e-mail: mariano.rocchi@uniba.it

Maloney and colleagues (Maloney et al. 2012) showed that the centromeric function can be linked to different repeated sequence variants generating 'functional epialleles'.

Separation of centromere identity from DNA sequence was first inferred from the analysis of human neocentromeres, in which centromeres form on single-copy sequences in rearranged chromosomes (Barry et al. 1999). Human neocentromeres have been identified in clinical cytogenetic laboratories; most of them arose to stabilize otherwise acentric fragments while a less common type was found in intact chromosomes where the native centromere has been inactivated giving rise to neodicentrics (Marshall et al. 2008). Given the lack of satellite repeats, some human neocentromeres have been deeply analysed by chromatin immunoprecipitation approaches (ChIP-on-chip or ChIP-seq) (Chueh et al. 2005; Alonso et al. 2010; Hasson et al. 2011, 2013); the main conclusions of these studies were that CENP-A binding is largely independent of DNA sequence and that extended herochromatin domains are not required for centromere function. Neocentromere formation on rearranged or engineered chromosomes has also been observed in other species, including *Saccharomyces pombe* (Steiner and Clarke 1994), *Drosophila melanogaster* (Williams et al. 1998), *Candida albicans* (Ketel et al. 2009), maize (Fu et al. 2013) and chicken (Shang et al. 2013).

The formation of novel centromeres can also occur during evolution through the repositioning of the centromere to a new site without chromosomal rearrangement; these evolutionary new centromeres (ENCs) significantly impact karyotype evolution, but their mechanisms of formation are unknown (Kalitsis and Choo 2012; Rocchi et al. 2012). Originally described in primates (Montefalcone et al. 1999), ENCs are particularly prevalent in the genus *Equus* (horses, asses and zebras) (Carbone et al. 2006). Although the majority of ENCs so far described contains satellite DNA arrays, it was proposed that the initial seeding of a new centromere during evolution occurs within an anonymous genomic region and that the acquisition of tandem repeats is a late phenomenon (Amor and Choo 2002; Piras et al. 2010); recent data on rice centromeres suggest that satellite repeats may evolve to stabilize centromeric nucleosomes (Zhang et al. 2013). 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). A multistep model for the birth, evolution and complete maturation of ENCs was proposed: The first step would consist in the shift of the centromeric function to a new position lacking satellite DNA, while the satellite DNA from the old centromere remains in the ancestral position; a subsequent step would be the loss of the leftover satellite DNA; finally, at a later stage, satellite repeats would colonize the new centromere giving rise to completely 'mature' centromeres (Amor and Choo 2002; Piras et al. 2010). During this process, dicentric chromosomes may be transiently generated but, according to the model, epigenetic marks rather than specific DNA sequences may determine the switch of the centromeric function from the old to the new position. Alternatively, the old centromere may be physically lost through chromosome rearrangement, similarly to what has been observed in clinical neocentromeres. A clear example of evolutionarily young neocentromere is the one on horse chromosome 11 which is completely devoid of satellite DNA (Wade et al. 2009). A ChIP-on-chip analysis of this centromere in one individual revealed the presence of two CENP-A binding domains. In order to shed light on the organization of the centromeric function in horse chromosome 11, in the present work, we exploited this satellite-less centromere to examine the detailed functional organization of this native mammalian centromere by analysing five new individuals. We demonstrated that the centromeric function is not fixed and identified at least seven functional alleles scattered in a region of about 500 kb; this surprisingly high positional variation gives rise to multiallelic epigenetic polymorphism. At a molecular level, these results reveal a mobility of CENP-A nucleosome arrays, a property that could be related to the evolutionary mobility of centromeres.

Materials and methods

Horse cells

Primary fibroblast cell lines were obtained from the skin of five different slaughtered animals and designated for convenience HSF-B, HSF-C, HSF-D, HSF-E and HSF-G. We do not know to which breed these animals belong. We tested their relatedness by standard DNA typing using the following microsatellite loci: AHT4, AHT5, ASB2, ASB17, ASB23, CA425, HTG4, HTG6, HTG7, HTG10, HMS2, HMS3, HMS6, HMS7, VHL20, HMS1. These include nine loci recommended by the 'Equine Genetics and Thoroughbred Parentage Testing Standardization Committee' of the International Society for Animal Genetics (ISAG) and eight additional loci commonly used for horse parentage testing and identification (Equine Gentypes Panel 1.1, Thermo Scientific). We then tested likelihood of relation using the Familias 3.1.3 software (http://familias.no).

The cells were cultured in high glucose DMEM (EuroClone) medium supplemented with 15 % foetal bovine serum, 2 mM L-glutamine, 1 % penicillin/streptomycin and 2 % non-essential amino acids at 37 °C with 5 % CO_2. The cell lines were from three male (HSF-B, HSF-C and HSF-G) and two female (HSF-D and HSF-E) animals. Cytogenetic analysis demonstrated that all cell lines had a diploid modal chromosome number (64) and a normal karyotype (Supplementary Fig. 1).

ChIP and ChIP-on-chip analysis

To identify the sequences bound by CENP-A, native chromatin immunoprecipitation analysis was performed, as previously described (Wade et al. 2009). Briefly, native chromatin was prepared from horse fibroblasts by nuclease digestion of cell nuclei; immunoprecipitation was then performed using a polyclonal antibody against the centromeric protein CENP-A (Trazzi et al. 2009). We have previously demonstrated that this antibody is able to recognize horse centromeres (Wade et al. 2009). Both input and immunoprecipitated DNA fragments were purified and amplified using the whole genome amplification (WGA) kit (Sigma-Aldrich, St. Louis, USA). ChIPed DNA was analysed by real-time PCR before and after WGA amplification.

The input and the immunoprecipitated DNAs were co-hybridized to a NimbleGen custom tiling array containing a 3.2 Mb region between nucleotides ECA11:25,566,599-28,305,611 with an average resolution of 100 bp. The array data were deposited in NCBI's Gene Expression Omnibus, and they are accessible through GEO Series accession number GSE57986 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57986). DNA binding peaks were identified by using the statistical model and methodology described at (http://chipanalysis.genomecenter.ucdavis.edu/cgi-bin/tamalpais.cgi) (Bieda et al. 2006) using stringent parameters for peak identification (98th percentile threshold and $p<0.0001$).

Real-time PCR analysis

Real-time PCR was performed using the Go Taq qPCR Master Mix (Promega) on a DNA Engine Opticon 2 System (Bio-Rad). Data were analysed using the Opticon Monitor 3 software.

For each individual, two independent real-time PCR experiments were performed on immunoprecipitated and input DNA using the primer pairs spanning the region of interest listed in Supplementary Table 1. The single-copy gene PRKCi (gene ID: 100063737, forward primer: TGGAGCAAAAGCAGGTGGTA, reverse primer: ATCGTCATCTGGAGTGAGCTG) was used as control. Real-time PCR was performed using the following temperature program: initial denaturation at 95 °C for 2 min; 50 cycles with denaturation at 95 °C for 15 s, annealing at 61 °C for 30 s and elongation at 72 °C for 30 s. Fluorescence detection was performed for 15 s at 80 °C. Final extension at 72 °C for 5 min. For melting curve analysis, a temperature gradient (60–94 °C, 1 °C/s) was applied. Each reaction was carried out in triplicate. For each primer pair, relative standard dilutions of input DNA (1:1, 1:10, 1:100) were included in the experiments. Real-time PCR results were considered reliable only when the r^2 value of the calibration curve was comprised between 0.95 and 1. To evaluate the relative fold enrichment, the $\Delta\Delta$Ct formula was applied where Ct is the cycle threshold.

SNP analysis

SNPs used for the analysis were identified using the website (http://www.broadinstitute.org/mammals/horse/snp).

Firstly, the SNPs were tested on genomic DNA by PCR and sequencing. Genomic DNA was extracted from primary fibroblasts using QIAGEN Blood and Cell culture DNA Midi kit according to manufacturer's instructions. DNA was amplified using the High Fidelity Herculase II Fusion DNA Polymerase (Stratagene, Agilent Technologies), and PCR products were sequenced. SNPs that were heterozygous in genomic DNA (Supplementary Table 2) were analysed both on input and on immunoprecipitated DNA from ChIP experiments.

BAC clones

The DNA segment spanning the centromere of horse chromosome 11 (chr11:27,400,000–28,150,000) was derived from the EquCab2.0 horse genome sequence assembly. The sequence was used as query against NCBI *Equus caballus* Clone End Sequence database. Bacterial artificial chromosome (BAC) end sequences from the horse CHORI-241 BAC library were searched (Leeb et al. 2006). The seven selected clones are reported in Supplementary Fig. 2. Their cytogenetic position was validated by fluorescent in situ hybridization (FISH) on horse metaphase chromosomes (Supplementary Fig. 3).

Immuno-FISH on extended chromatin fibres

Extended chromatin fibres were prepared using published methods (Lam et al. 2006; Maloney et al. 2012) with slight modifications; in particular, an electrical device, equipped with a pulley, was built specifically to raise slides from the lysis buffer perpendicularly and at a constant speed. Immunofluorescence, carried out using a CREST serum (kindly provided by Claudia Alpini, Fondazione I.R.C.C.S. Policlinico San Matteo, Pavia), was followed by FISH with the appropriate BAC clones. Fibres were prepared from at least two independent experiments; combined immunostaining and FISH were performed using different schemes to avoid potential hybridization or detection bias with fluorescent secondary antibodies. DNA fibres were counterstained with 5 mg/mL DAPI and mounted with DAKO mounting medium (DAKO).

Animal rights statement

The horse skin samples were taken from animals not specifically sacrificed for this study; the animals were being processed as part of the normal work of the abattoirs.

Results

Variable position of CENP-A binding domains in different individuals

We established fibroblast cell lines from five horses (HSF-B, HSF-C, HSF-D, HSF-E and HSF-G). Using 17 microsatellite loci (Thermo Scientific Equine Genotypes Panel 1.1), we determined their likelihood of relation with the Familias 3.1.3 software, demonstrating that they were unrelated (see Materials and Methods). The unexpected observation of two CENP-A binding domains in the horse previously analysed (Wade et al. 2009) prompted us to extend the analysis to these five new individuals. Chromatin was immunoprecipitated with an antibody against CENP-A. DNA was then purified and hybridized to a 3.2 Mb tiling array (accession number: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57986) spanning the centromeric region of horse chromosome 11 that we previously defined (Wade et al. 2009). The absence of satellite repeats at this locus (Wade et al. 2009) allowed us to position CENP-A-associated DNA (Fig. 1a). Strikingly, each individual exhibited a distinct arrangement of CENP-A binding domains. These were located across a region of approximately 500 kb, with some individuals (HSF-B, HSF-C and HSF-G) exhibiting two clearly defined peaks while others (HSF-D and HSF-E) showed one.

At least seven functional epialleles were identified in the five horses and are sketched in panel b of Fig. 1; identification was obtained by combining the results of ChIP-on-chip (panel a), qPCR (Fig. 2), SNP analysis (Fig. 3) and fibre immuno-FISH (Fig. 4). Each epiallele occupies about 80–160 kb. These results demonstrate that the centromeric domain of horse chromosome 11 is characterized by great positional variation giving rise to 'epigenetic polymorphism'. No functionally homozygous individuals were observed; therefore, in spite of our limited sample size, we can infer that this epigenetic locus is highly polymorphic.

To define the position and number of CENP-A binding domains with a different approach, we designed a set of 27 primer pairs (Supplementary Table 1) spanning the 500 kb region. Real-time PCR experiments were then carried out on the DNA purified from CENP-A immunoprecipitated chromatin from the five individuals. The q-PCR data confirmed those obtained by the ChIP-on-chip (Fig. 2): Two regions of CENP-A binding were identified in individuals HSF-B, HSF-C and HSF-G, while a single region could be observed in HSF-D and HSF-E.

Analysis of domain organization by single nucleotide polymorphism and immuno-FISH on chromatin fibres

The presence of two domains of CENP-A binding in some individuals could reflect a multidomain centromere structure, shared by both chromosomes 11; alternatively, one of the domains seen in HSF-B, HSF-C and HSF-G could be located on one of the two homologous chromosomes 11 and the second one on the other homolog. To unravel which one of the two possibilities was correct, we sought heterozygous nucleotide positions, SNPs, located within the centromeric domains using the SNP database (see Materials and Methods). Informative SNPs were then identified within the CENP-A binding domains of individuals HSF-D, HSF-G and HSF-E. For HSF-C and HSF-B, the SNPs available in the database were not informative; therefore, in these two horses, informative loci were identified by sequencing PCR products from genomic DNA. These heterozygous positions (Supplementary Table 2 and Fig. 1 black and red dots) would allow us to resolve the two homologs in DNA purified from CENP-A chromatin immunoprecipitations: If the two CENP-A domains were present on both homologs, the immunoprecipitated chromatin would contain similar amounts of the two alleles; on the contrary, if each homolog contained a single CENP-A domain, only one of the two alleles would be enriched in the immunoprecipitated chromatin. The results of all experiments relative to the five horses are summarized in the Supplementary Table 2. In Fig. 3, representative Sanger sequence traces from horses HSF-C, HSF-D, HSF-E and HSF-G are shown.

In Fig. 3 (top panels), Sanger sequence traces from input and CENP-A immunoprecipitated DNA, relative to three SNPs in HSF-G and two SNPs in HSF-C are shown. The position of these SNPs is marked with blue carats in Fig. 1 and are listed, using blue colour, in Supplementary Table 2. At all these SNP positions, both nucleotides were present in input DNA while in the immunoprecipitated DNA, enrichment of only one nucleotide was clearly detected. These results strongly suggest that, in HSF-C and HSF-G, each homolog contains a single CENP-A binding domain. Similarly, in HSF-B, the analysis of the heterozygous microsatellite locus strongly suggests that each one of the two CENP-A domains is located on one homolog (Supplementary Table 2).

In HSF-D and HSF-E, in which a single broad peak of CENP-A binding was observed by ChIP-on-chip (Fig. 1) and q-PCR (Fig. 2), different results were obtained when SNPs at the edges (black dots in Fig. 1) or at the centre (red dots in Fig. 1) of the peak were analysed (Fig. 3). At the edges, in DNA purified from CENP-A immunoprecipitations, a single nucleotide was enriched in the sequence profiles, similarly to what we observed within the HSF-C and HSF-G peaks; on the contrary, at the centre of the broad peak, both SNP nucleotides were bound by CENP-A. The interpretation of this result is that CENP-A binds to different regions in the two homologs, as in horses HSF-C and HSF-G. However, in HSF-D and HSF-E, the CENP-A binding domains are partially overlapping in the horse genome sequence and correspond to the left and the right part of the broad ChIp-on-chip peak,

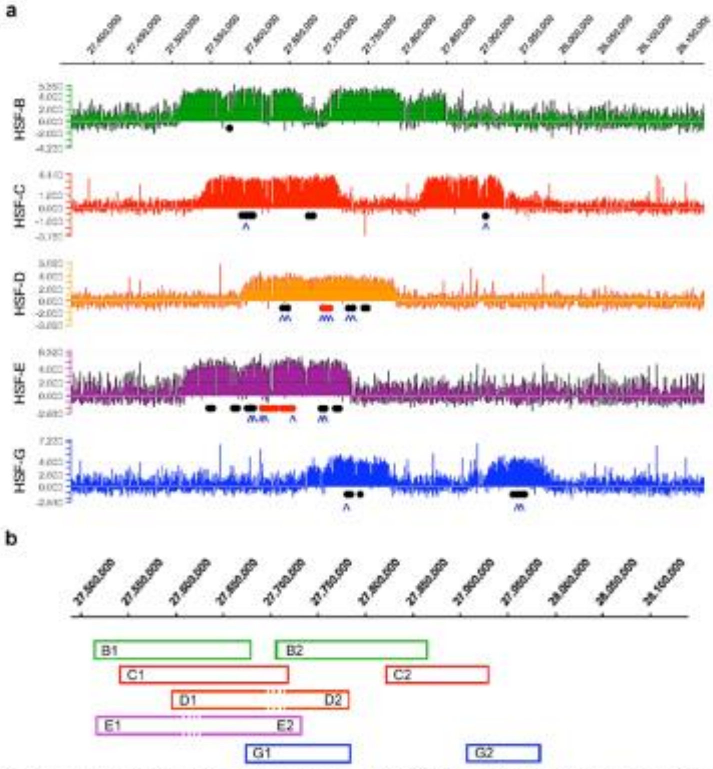

Fig. 1 Variable position of the centromere of horse chromosome 11. **a** DNA obtained by chromatin immunoprecipitation. Using an anti-CENP-A antibody, from five different horse fibroblast cultures was hybridized to a tiling array covering the centromere region. Results are presented as the log2 ratio 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), *red dots* (both SNP alleles are present in immunoprecipitated DNA) and *blue carats* (SNPs shown in Fig. 3). **b** Peak positions are represented as *boxes*. Epiallele identification was obtained by combining ChIP-on-chip, SNP (Fig. 3) and fibre FISH (Fig. 4 and Supplementary Table 2) results. Sequence coordinates refer to the horse EquCab2.0 (2007) sequence assembly, as reported by the UCSC genome browser (http://genome.ucsc.edu). Alleles are designated by the letter of the horse they derive from, followed by '1' or '2' to distinguish the two variants. In HSF-D and HSF-E, where a single broad peak was identified by ChIP-on-chip while two distinct centromeric domains were identified by fibre-FISH (Fig. 4) and SNP analysis (Fig. 3 and Supplementary Table 2), *dotted lines* represent the region of overlap of the two binding domains in the reference sequence. Therefore, at least seven different centromeric domains can be identified: Ba/Ea, Bb, Ca, Cb, Da/Eb, Db/Ga, Gb

respectively; the overlapping region roughly corresponds to the centre of the broad peak. Therefore, also for HSF-D and HSF-E, the results are consistent with the presence of one CENP-A binding domain on each homolog.

The results of SNP analysis were confirmed by an independent approach that is single molecule analysis of centromeric domains by immuno-FISH on chromatin fibres. BACs covering the centromeric domain (Supplementary Fig. 2), as determined by ChIP-on-chip, were used as FISH probes, and a CREST serum was used to detect the functional centromeric domain. In Supplementary Fig. 2a, the BAC clones are listed with their genomic coordinates and their

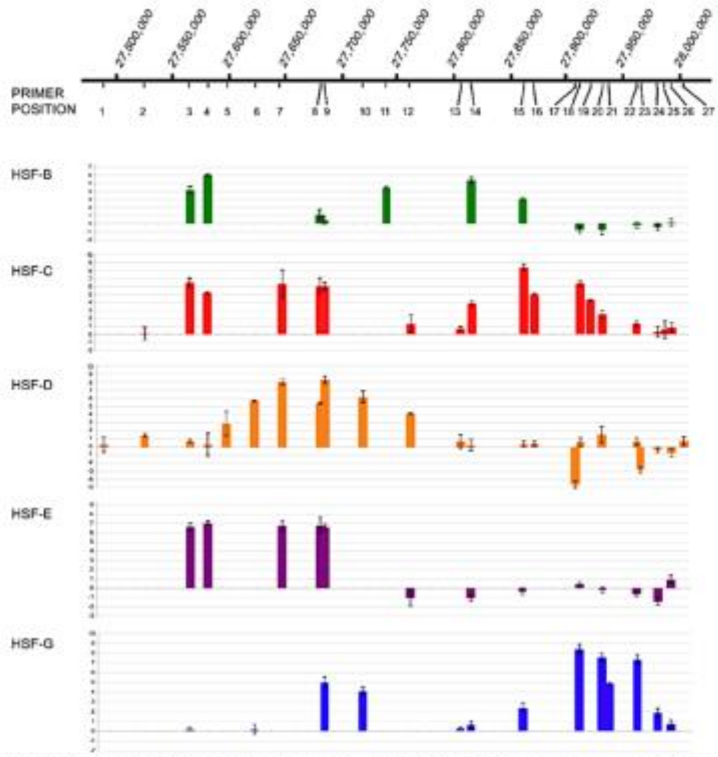

Fig. 2 Real-time PCR analysis of the ChIP-on-chip samples. For each cell line (HSF-B, HSF-C, HSF-D, HSF-E and HSF-G), results are presented as the logarithm of the difference between the cycle threshold obtained with the CENP-A immunoprecipitated sample and the cycle threshold obtained with input sample, normalized for the control region (chr11:28,227,839-28,227,938). The x-axis shows the genomic position of each primer pair along chromosome 11

position on the genome map is sketched in panel b of the same figure. Concerning the CREST serum used, we showed that the signals obtained on DNA fibres is perfectly overlapping with the signal obtained by a monoclonal anti-CENP-A antibody (Supplementary Fig. 4), the CREST serum signal being particularly intense and therefore more suitable for the immuno-FISH experiments in combination with BAC clones. Samples from HSF-B, HSF-D, HSF-E and HSF-G were analysed. We observed two different organization patterns of FISH and immuno-staining fluorescent signals which are exemplified in Fig. 4. The first type of arrangement is reported in Fig. 4a and was observed in samples from horses displaying two clearly separated ChIP-on-chip peaks (HSF-B and HSF-G). Two distinct epialleles could be distinguished, one of which (epiallele 1 in Fig. 4a) had the immuno-staining flanking the FISH signal while in the other one (epiallele 2 in Fig. 4a), the immuno-staining and FISH signals were superimposed. The second type of arrangement, observed in horses HSF-D and HSF-E, is reported in Fig. 4b. These two horses displayed a single broad ChIP-on-chip peak, and SNP data indicated that the broad peak was the result of the partial overlap of two distinct peaks. Immuno-FISH confirmed this interpretation: Indeed, as shown in Fig. 4b, two functional alleles could be observed also in these horses. In one epiallele

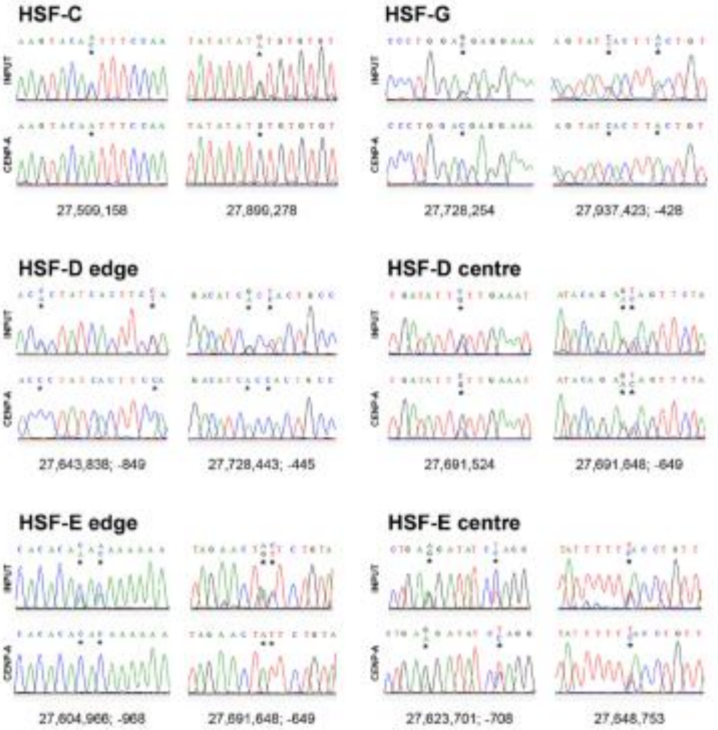

Fig. 3 SNP analysis of centromeric domains. Sanger sequence traces from input (*above*) and CENP-A immunoprecipitated (*below*) samples from HSF-C, HSF-G, HSF-D and HSF-E. SNP coordinates are beneath traces. Stars indicate SNPs. For HSF-C, HSF-G, HSF-D-edge and HSF-E-edge, both nucleotides are present in input DNA while the immunoprecipitated DNA is enriched for one of the two nucleotides. For HSF-D centre and HSF-E centre, the two nucleotides are present in both input and CENP-A immunoprecipitated samples

(epiallele 1 in Fig. 4b), the immuno-staining partially covered the FISH signal and extended in the flanking region, while in the other epiallele (epiallele 2 in Fig. 4b), the immuno-staining covered the FISH signal. The immuno-labelled regions of epiallele 1 and epiallele 2 were partially overlapping.

Sequence analysis of the DNA region containing the CENP-A binding domains

To test whether any peculiar DNA sequence composition may account for the presence of centromeric domains, we carried out a detailed analysis of the region under study and of 64 control regions (two interstitial regions from each horse chromosome were chosen at random) of the same size, using the RepeatMasker software (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker); subtelomeric and heterochromatic regions were intentionally excluded from the analysis. The results are reported in Supplementary Table 3 and summarized in Fig. 5. For ECA11, the analysis was performed on the entire centromeric region and on each individual CENP-A binding domain identified by ChIP-on-chip. In the control

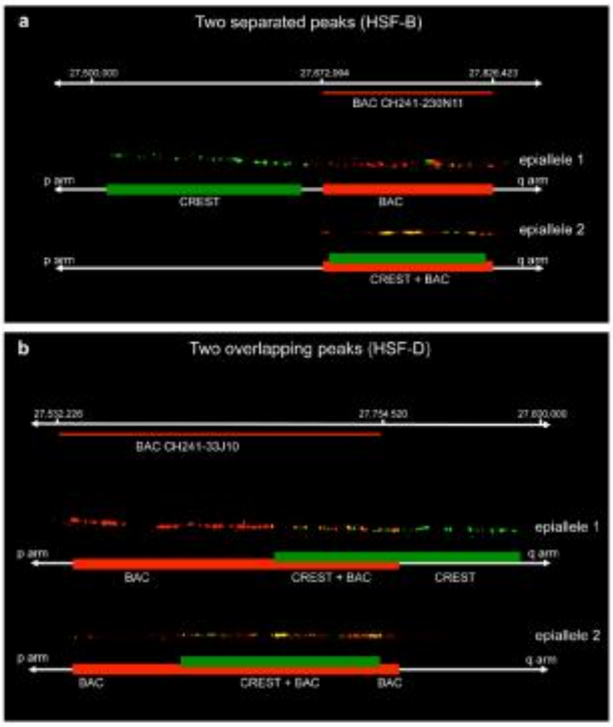

Fig. 4 Single molecule analysis of centromeric epialleles on chromatin fibres by immuno-FISH. **a** Organization pattern of functional allelels in horses displaying two separated ChIP-on-chip peaks (HSF-B and HSF-G). The example shown refers to horse HSF-B, and the BAC used was CH241-230 N11. **b** Pattern of functional alleles organization in horses displaying two overlapping ChIP-on-chip peaks (HSF-D and HSF-E). The example shown refers to horse HSF-D, and the BAC used was CH241-33 J10. At the *top* of each panel, the coordinates of the regions occupied by the centromeric domains are reported, and BAC coverage is represented by a *red line*. CREST immunostaining is green labelled while the BAC FISH signals are *red* labelled. Under each fibre image, a schematic representation is depicted with *green rectangles* corresponding to centromeric domains and *red rectangles* indicating BAC hybridization. Two (HSF-B and HSF-G) or three (HSF-D and HSF-E) independent experiments were performed for each horse, and at least 10 chromatin fibres were analysed. The ratio of epialleles 1 and 2 observed in the individual horses was close to 50 %: HSF-B 5/11 vs 6/11; HSF-G 4/10 vs 6/10; HSF-D 14/26 vs 12/26; HSF-E 6/16 vs 10/16

regions, the guanine-cytosine (GC) content ranged between 34.74 and 48.52 % with a mean value of 40. 25 %. Consistently, in the entire centromeric region, the GC content was 39.12 %; little variation around this value was observed among single CENP-A binding domains (data on peaks in Supplementary Table 3). It is important to note that the GC content of the entire region does not correspond to the mean of the single peaks due to peak overlapping. Student's *t* test indicated that the average GC content of the control regions was not significantly different from the ECA11 centromeric region ($p=0.75$, Fig. 5). A similar comparison was carried out for the following classes of repetitive elements: SINEs, LINEs, LTRs, DNA transposable elements, small RNAs and low-complexity repeats; *p* values comprised between 0.32 and 0.89 indicated that the repeated element composition of the ECA11 centromeric domain was comparable to that of the control regions.

Discussion

The results presented here reveal a remarkable plasticity of the satellite-less centromere of horse chromosome 11. In this analysis of ten horse chromosomes 11, at least seven distinct CENP-A binding domains, each one extending for about 80–160 kb, were found across a region of about 500 kb. These results demonstrate that, in a native mammalian centromere, the positioning of CENP-A binding domains is unrelated to the sequence of the DNA the centromere is associated with and that centromere position can be flexible across a relatively

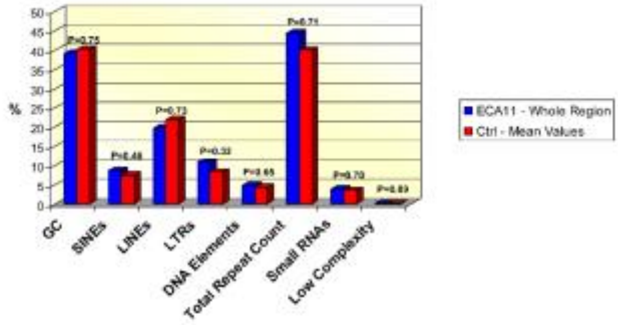

Fig. 5 Sequence analysis. *Blue bars* represent the percentage of each class of sequence in the entire ECA11 centromeric region. *Red bars* correspond to mean values from 64 control regions of the percentage of each sequence class (data are reported in Supplementary Table 3). Satistical analysis was performed using the Student's *t* test, and the *p* values are reported for each comparison, indicating that the differences between the centromeric and the control regions are not statistically significant

wide single-copy genomic region. Indeed, the sequence features (GC and repetitive elements content) of the ECA11 centromeric region are comparable to those of random intrachromosomal genomic regions. The analysis of the GC content of this genomic region was performed taking into consideration the isochore theory (Bernardi 1993). According to this theory, stretches of more than 300 kb, uniform for GC content, characterize the genomes of 'worm-blooded' vertebrates. With this analysis, we intended to test whether the centromeric region of horse chromosome 11 was inserted in an AT reach isochore, as previously suggested for other mammalian neocentromeres (Marshall et al. 2008).

Although the size and organization of mammalian and fission yeast centromeres are remarkably different, it was recently shown that, also in the small centromere of *S. pombe*, the positioning of CENP-A/Cnp1 nucleosomes varies relative to the underlying DNA sequence among genetically homogeneous cell lines (Yao et al. 2013).

When neocentromeres were experimentally induced in chiken DT40 cells, most of them were formed at multiple positions close to the original centromere; interestingly, detectable levels of CENP-A were found in a 2 Mb region surrounding the original centromere (Shang et al. 2013). The proposed hypothesis was that epigenetic marks favouring 'centromerization' were present around the original centromere, and this may be the reason why neocentromeres were preferentially seeded in that region. In spite of the positional variation of neocentromeres induced by chromosome engineering, in the chicken system, centromere spreading seems to be prevented in wild-type cells (Shang et al. 2013). Here, we demonstrated that the wild-type centromere of horse chromosome 11, unlike chicken wild-type centromeres, moves considerably within a 500 kb region. It is important to underline that we analysed an evolutionary neocentromere that was established about one million years ago, after the divergence of horses from the other species of the genus *Equus* (asses and zebras) (Piras et al. 2010). The centromeric domains detectable nowadays are the result of a positional sliding that occurred during the evolution of the horse lineage; we are therefore taking a 'snapshot' of an ongoing evolutionary process whose initial shots are unavailable.

It is possible that removal of the centromere of horse chromosome 11 from a typical heterochromatic environment has revealed or exacerbated an underlying dynamic behaviour of CENP-A chromatin, as proposed for experimentally induced neocentromeres in *Drosophila* (Maggert and Karpen 2001). Some human neocentromeres have been shown to be very poor in heterochromatin, and this feature has been correlated with defects of sister chromatid cohesion (Alonso et al. 2010). This observation is in agreement with the hypothesis that evolutionary neocentromeres tend to be 'stabilized' through the recruitment of satellite DNA. Indeed, it has been proposed that the mosaicism observed for some clinical neocentromeres may be due to their intrinsic mitotic instability (Marshall et al. 2008). On the contrary, the neocentromere of horse chromosome 11 must be sufficiently stable to be present in all individuals of the species. Heterochromatin has been shown to limit spreading of protein domains in *S. pombe* (Partridge et al. 2000) and to specifically exclude CENP-A incorporation in *Drosophila* (Heun et al. 2006). In addition, although the role of the centromeric protein CENP-B is not well understood, it has been suggested that this protein might contribute to the organization of centromeric heterochromatin both in fission yeast (Nakagawa et al. 2002) and in humans (Okada et al. 2007). Since we did not find any evidence for the presence of CENP-B boxes (the consensus sequence binding CENP-B) in the ECA11 centromeric region (data not shown), it is tempting to speculate that the absence or low level of binding to chromatin of this protein may contribute to the sliding of CENP-A domains described here. We propose that fluctuations in CENP-A nucleosome positioning may give rise to a diffusion-like behaviour, a form of un-anchored chromatin spreading, that could account for 'centromere sliding'. Such dynamic behaviour might be one reason for the great variability of centromere-associated DNA sequences.

It is worth noticing that cytogenetic approaches on metaphase chromosomes never revealed positional variation of the primary constriction on horse chromosome 11, indicating that the polymorphism described here involves a defined genomic region whose size is under the resolution limit of cytogenetic analysis; indeed, this region occupies about 500 kb. In any case, the phenomenon described here is distinct from larger scale centromere repositioning observed during karyotype evolution (Carbone et al. 2006; Rocchi et al. 2012).

It is possible that the centromere studied here is particularly dynamic because it is evolutionarily young and lacks satellite tandem repeats (Wade et al. 2009; Piras et al. 2010). As mentioned above, some positional variation, affecting centromeric domains on alphoid DNA, was observed on the mature human chromosome 17 (Maloney et al. 2012). In that case, two adjacent alpha satellite arrays were shown to possess centromere activity. In our system, the lack of satellite DNA at the centromere of horse chromosome 11 is a stable feature in all individuals of the horse species and was maintained for many generations during evolution; therefore, the mechanisms of satellite DNA recruitment and the precise role of repetitive sequences in centromere function and stabilization remain to be established. Satellite DNA recruitment appears to be a late step in centromere repositioning events, with repetitive DNA arrays proposed to play a role in stabilizing centromere position. We suggest that the colonization of a CENP-A domain by satellite DNA may progressively reduce the positional flexibility of the centromere through a satellite-mediated stabilization mechanism.

We do not know the probability of centromere movement per cell per generation nor how far from their original position CENP-A binding domains can move. We have evidence that the position of these domains is endowed with a certain degree of stability as we did not detect any positional variation in our fibroblast cell lines at different culture passages (data not shown). Another open question is the evolutionary timescale of centromere movement; the great variability of CENP-A domain position in our ten chromosome sample suggests that this phenomenon is quite frequent, at least in horse chromosome 11.

We previously described, in non-horse species of the genus *Equus*, a number of centromeres at different maturation stages, some of which seem to be devoid of extended clusters of tandemly repeated DNA (Piras et al. 2010). These satellite-less equid centromeres represent a new and powerful model system offering a clear advantage with respect to engineered or clinical neocentromeres: They are natural, stably present in all individuals of a given species and can therefore be used as an ideal tool to study the maturation and fixation of evolutionary new centromeres. In addition, the non-repetitive nature of a number of equid centromeres and the availability of the complete sequence of the horse genome provide the chance to analyse, at the molecular level, the architecture, plasticity and evolution of natural centromeres.

Acknowledgments We thank Margherita Bonuglia (Laboratorio di Genetica Forense Veterinaria, UNIRELAB, Settimo Milanese, Italy) for performing the microsatellite analysis, Claudia Alpini (Fondazione I.R.C.C.S. Policlinico San Matteo, Pavia, Italy) for the CREST serum, Tosso Leeb and Maren Scharfe (Helmholtz Centre for Infection Research, Braunschweig, Germany) for the BAC clones, Sergio Comincini and Luca Ferretti (Dipartimento di Biologia e Biotecnologie, Università di Pavia, Italy) for helpful suggestions, reagents and instruments. We are grateful to Cesare Rognoni, DVM, (Equicenter Monteleone, Inverno e Monteleone, Pavia, Italy) for his precious collaboration.

References

Alonso A, Hasson D, Cheung F, Warburton PE (2010) A paucity of heterochromatin at functional human neocentromeres. Epigenetics Chromatin 3:6

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

Barry AE, Howman EV, Cancilla MR, Saffery R, Choo KH (1999) Sequence analysis of an 80 kb human neocentromere. Hum Mol Genet 8:217–227

Bernardi G (1993) The vertebrate genome: isochores and evolution. Mol Biol Evol 10:186–204

Bieda M, Xu X, Singer MA, Green R, Farnham PJ (2006) Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. Genome Res 16:595–605

Black BE, Cleveland DW (2011) Epigenetic centromere propagation and the nature of CENP-A nucleosomes. Cell 144:471–479

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

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

Fu S, Lv Z, Gao Z, Wu H, Pang J, Zhang B, Dong Q, Guo X, Wang XJ, Birchler JA, Han F (2013) De novo centromere formation on a chromosome fragment in maize. Proc Natl Acad Sci U S A 110:6033–6036

Hasson D, Alonso A, Cheung F, Tepperberg JH, Papenhausen PR, Engelen JJ, Warburton PE (2011) Formation of novel CENP-A domains on tandem repetitive DNA and across chromosome breakpoints on human chromosome 8q21 neocentromeres. Chromosoma 120:621–632

Hasson D, Panchenko T, Salimian KJ, Salman MU, Sekulic N, Alonso A, Warburton PE, Black BE (2013) The octamer is the major form of CENP-A nucleosomes at human centromeres. Nat Struct Mol Biol 20:687–695

Santagostino *et al. BMC Genetics* (2015) 16:126
DOI 10.1186/s12863-015-0281-1

BMC Genetics

RESEARCH ARTICLE Open Access

Genome-wide evolutionary and functional analysis of the Equine Repetitive Element 1: an insertion in the myostatin promoter affects gene expression

Marco Santagostino[1†], Lela Khoriauli[1†], Riccardo Gamba[1†], Margherita Bonuglia[2], Ori Klipstein[1], Francesca M. Piras[1], Francesco Vella[1], Alessandra Russo[2], Claudia Badiale[1], Alice Mazzagatti[1], Elena Raimondi[1], Solomon G. Nergadze[1*] and Elena Giulotto[1*]

Abstract

Background: In mammals, an important source of genomic variation is insertion polymorphism of retrotransposons. These may acquire a functional role when inserted inside genes or in their proximity. The aim of this work was to carry out a genome wide analysis of ERE1 retrotransposons in the horse and to analyze insertion polymorphism in relation to evolution and function. The effect of an ERE1 insertion in the promoter of the myostatin gene, which is involved in muscle development, was also investigated.

Results: In the horse population, the fraction of ERE1 polymorphic loci is related to the degree of similarity to their consensus sequence. Through the analysis of ERE1 conservation in seven equid species, we established that the level of identity to their consensus is indicative of evolutionary age of insertion. The position of ERE1s relative to genes suggests that some elements have acquired a functional role. Reporter gene assays showed that the ERE1 insertion within the horse myostatin promoter affects gene expression. The frequency of this variant promoter correlates with sport aptitude and racing performance.

Conclusions: Sequence conservation and insertion polymorphism of ERE1 elements are related to the time of their appearance in the horse lineage, therefore, ERE1s are a useful tool for evolutionary and population studies. Our results suggest that the ERE1 insertion at the myostatin locus has been unwittingly selected by breeders to obtain horses with specific racing abilities. Although a complex combination of environmental and genetic factors contributes to athletic performance, breeding schemes may take into account ERE1 insertion polymorphism at the myostatin promoter.

Keywords: Horse genome, SINEs, Equids, Myostatin gene expression

Background

A large fraction of the genome of mammals is occupied by interspersed repeats that were generated during evolution by the propagation of transposable elements [1–3]. Short INterspersed Elements (SINEs) are non-autonomous retrotransposons that make use of a transposition process in which an RNA intermediate is reverse transcribed and the resulting cDNA is inserted into a new genomic location [4, 5]. Sequence analysis of SINE elements suggested that most of them derive from ancestral tRNAs, but there are examples of 5S- or 7SL-like sequences [6]. These elements are characterized by two internal RNA-polymerase III promoters that make them transcriptionally independent, but their retrotranscription and integration processes are catalyzed by enzymes encoded by autonomous Long INterspesed Elements (LINEs) [4, 5]. The primate Alu family is an example of SINE; Alu repeats are the most abundant transposable elements in the human genome accounting

* Correspondence: solomon.nergadze@unipv.it; elena.giulotto@unipv.it
†Equal contributors
[1]Dipartimento di Biologia e Biotecnologie "Lazzaro Spallanzani", Università di Pavia, Via Ferrata 1, 27100 Pavia, Italy
Full list of author information is available at the end of the article

BioMed Central

for more than one million copies [7–9]. The majority of human Alu elements are present in all individuals because they were inserted in the genome before the radiation of extant humans; however, some Alu elements, that were integrated recently in the human lineage, are characterized by insertion polymorphism [9–12]. In humans, an inverse correlation between the evolutionary age of Alu subfamilies and the percentage of polymorphic elements was demonstrated: 20–25 % of the elements belonging to the youngest subfamily (AluY) are polymorphic [13].

Because of their abundance and mechanism of origin, transposable elements were considered "junk DNA", albeit, in a number of examples it was shown that they can acquire a functional role, a process termed "exaptation" [14–17]; in particular, the insertion of transposable elements inside genes or in their proximity may alter gene structure or expression through gene interruption, introduction of promoter sequences or splice sites [18–20]. In some rare cases, transposons are implicated in genetic disease or cancer [21–23].

In the present paper, taking advantage of the published horse genome sequence [24], we carried out a genome wide analysis of the perissodactyl-specific SINE family of Equine Repetitive Elements (ERE) focusing our attention on insertion polymorphism in relation to sequence conservation. ERE retrotransposons derive from tRNA^{Ser} and occupy about 4 % of the horse genome [25, 26]; to date, four main ERE subfamilies were identified: ERE1-4 [27, 28]. To our knowledge, before the present study, no data were available on the involvement of horse transposable elements in the modulation of gene expression. The description of a polymorphic ERE1 insertion in the promoter of the myostatin gene [29] prompted us to investigate the possible functional role of this insertion.

Myostatin or growth/differentiation factor 8, a member of the transforming growth factor-β family, is a repressor of muscle growth that regulates myoblast proliferation and differentiation. It has been shown previously that mutations in the myostatin gene can cause muscle hypertrophy in a range of mammals such as mice [30], cattle [31, 32] and sheep [33]. In 2004, Schuelke and collaborators reported the case of an extraordinarily muscular child whose mother appeared muscular, although not to the extent observed in her son, and was a professional athlete [34]. The authors discovered that the boy carried a single base substitution in both copies of the myostatin gene generating a premature termination codon while the mother was heterozygous for the mutation. Particularly relevant in this context is also the "bully" phenotype in whippet racing dogs, which depends on a frameshift mutation causing the production of a truncated protein. Individuals homozygous for the mutation show a double-muscle-phenotype, called "bully", while heterozygotes display an intermediate phenotype. While heterozygous animals have significantly greater racing ability than wild-type and mutated homozygous dogs, the excessive muscle mass of homozygotes for the mutation is detrimental for performance [35].

In the horse, the myostatin gene, which comprises three exons and two introns, is located on chromosome 18; several sequence variants were identified in this gene and in its flanking regions [29, 36–41]; among these variants the SNP g.66493737C > T, which is contained within the first intron, was associated with regulation of gene expression in Thoroughbred race horses and proposed as the best predictor of optimum racing distance [29, 38, 42]. The same variant was also associated with high values of body weight/withers height ratio, which, in the horse, is considered a good indicator of skeletal muscle mass [43]. Four additional SNPs, located in the regions adjacent to the myostatin gene, have been identified on chromosome 18 and were associated to performance [43–45]. Finally, as mentioned above, the insertion of an ERE1 element within the promoter region of the myostatin gene was described in some Thoroughbreds [37]. Recently, the presence of this insertion has been associated with a different muscle fiber composition [40, 46]. In the present paper we tested whether this insertion affects gene expression, contributes to breed differentiation and is relevant for sport aptitude and racing performance.

Results and discussion

Insertion polymorphism of ERE loci in the reference genome

A large body of evidence suggests that the horse genome is in a state of rapid evolution [24, 47–50]. Therefore, we may expect that several transposon insertions may have occurred in the horse lineage in relatively recent evolutionary times.

A preliminary *in silico* analysis of the four ERE subfamilies (ERE1 to ERE4) was carried out. To this purpose, the consensus of each ERE subfamily [27, 28] was used as query for a BLAT search (BLAST-Like Alignment Tool) in the reference sequence of the horse [51, 52], which derives from the assembly of the genomic sequence of the Thoroughbred horse named Twilight [24]. From each ERE subfamily, the 200 loci with the highest identity to their consensus were analyzed in search of empty alleles (i.e., alleles in which the ERE element is not present, ERE–) that may be present in the reference genome, thus identifying heterozygous loci in the genome of Twilight. ERE– alleles were found for 3.5 % of the ERE1, 0.5 % of the ERE2 and none of the ERE3 and ERE4 loci. Since the frequency of insertion polymorphism of transposable elements is related to the age of their insertion in the host genome [11], these results strongly suggest that ERE1s are the elements that were inserted most recently in the horse genome. It must be underlined that, since the reference sequence derives from the genome of a single horse, the frequencies of polymorphic loci reported above are largely underestimated being based on the analysis of two alleles per locus.

We then focused on the youngest subfamily, the ERE1, and carried out an extensive genome wide search of these elements in the reference genome sequence (Broad/equCab2). A list of 45,713 ERE1 loci was obtained using the consensus sequence deposited at the RepBase database as query [53] for a BLAST search (Additional file 1: Table S1A). The sequences were then filtered to include only elements with sizes similar to the ERE1 consensus (225 bp ± 10 bp) and with minimum identity of 84 % to the consensus. This operation left 34,131 loci (Additional file 1: Table S1B). The ERE1 sequences located inside other repetitive elements were also excluded from the analysis to avoid false positive results; this operation left 27,396 loci (Additional file 1: Table S1C). In order to obtain a comprehensive view of polymorphic ERE1 loci in Twilight, we analyzed the horse trace database, which includes unassembled traces [54] (center_project number G836). The sequence of each one of the 27,396 ERE1 loci was used as query for a BLAST search. The results of this analysis showed that Twilight is heterozygous at 377 ERE1 loci, possessing an ERE1+ and an ERE1− allele. A complete list of these polymorphic loci is reported in Additional file 2: Table S2. It is important to point out that an undefined number of ERE1 insertions, that are present in the horse population, is not detectable in the reference genome because Twilight may carry two ERE1− empty alleles at such loci. A clear example of this situation is the insertion in the myostatin gene promoter described below.

Since the fixation of insertion elements in the genome of a phylogenetic lineage requires many generations, the presence of empty alleles suggests that the insertion event occurred in relatively recent evolutionary times. In addition, mutations tend to accumulate in the inserted element and therefore a high degree of sequence conservation is considered indicative of a young evolutionary age of insertion, as previously shown for primate and rodent interstitial telomeric sequences [55, 56] and for human transposable elements [57]. In light of these considerations, we can hypothesize that ERE1 elements with higher identities to the consensus may have greater probabilities of being polymorphic compared to less conserved elements. To test this hypothesis, we evaluated the frequency of polymorphic loci in eight classes of ERE1 elements, characterized by different degrees of identity to the consensus (Fig. 1 and 2; data file 1: Table S3). In the class including ERE1 loci with the highest identity to the consensus (98–100 %), the percentage of loci that are polymorphic in Twilight is surprisingly high (4.6 %); this fraction decreases with the decrease of identity to the consensus reaching values as low as 0.1 % (Fig. 1). The correlation between fraction of polymorphic loci and percentage of identity to the ERE1 consensus sequence is highly significant (Pearson's correlation $\rho = 0.93$, $p = 8.5 \times 10^{-4}$). These results suggest that sequence conservation and insertion polymorphism of ERE elements are both related to the time of their appearance in the horse lineage.

Insertion polymorphism in the horse population, evolutionary history and sequence conservation of ERE1 loci

To evaluate the frequency of insertion polymorphism in the horse population, we analyzed 80 ERE1 loci in 30 unrelated domestic horses of different origin (see Materials and Methods). The 80 loci were chosen randomly

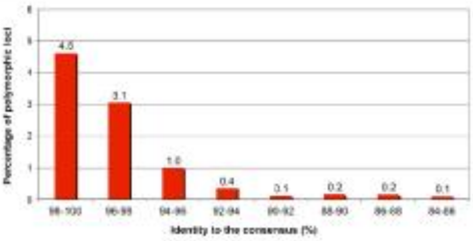

Fig. 1 Percentage of ERE1 polymorphic loci in the horse genome reference sequence. The ERE1 elements were grouped in eight classes according to their identity to the ERE1 consensus sequence published in Repbase. The percentage of polymorphic loci in each class is reported

from four classes (20 loci per class) with different degrees of identity to the ERE1 consensus sequence (≥98, 95, 90 and 85 % identity). For each locus, a primer pair flanking the ERE1 element was designed (Additional file 2: Table S4) and the genomic DNA of the 30 horses was amplified by PCR. The analysis of these loci in the 30 horses is summarized in Fig. 2, where different colours indicate the genotypes of each individual: ERE1+/+, green; ERE1+/−, yellow; ERE1−/−, red. For 71 loci (Fig. 2) only individuals homozygous for the presence of the ERE1 element (ERE1+/+) were found, suggesting that either the insertion is fixed in the population or the frequency of ERE1- alleles is very low. The remaining 9 loci were characterized by insertion polymorphism (Fig. 2). At these 9 loci, the fraction of ERE1- alleles per locus is highly variable ranging from 1.7 (locus 51) to 97 % (locus 11). Although the number of loci analyzed in each class as well as the number of individuals are relatively small, the results are in agreement with the *in silico* results described above: polymorphic loci are more represented in the class with the highest similarity to the ERE1 consensus sequence (6 loci out 20) whereas no polymorphic loci were identified in the class with the lowest identity to the consensus. These results confirm the observation, reported in the previous paragraph, that elements with high similarity to the consensus sequence, have a greater probability of being polymorphic compared to less conserved elements. We previously observed a high frequency of insertion polymorphism in the horse, involving NUMT elements (NUclear sequences of MiTochondrial origin) [49]. Similarly to NUMT sequences, the fraction of ERE1 polymorphic loci described here is particularly high compared to that reported for SINE elements in the human genome [9], thus providing further evidence for the rapid evolution of the horse genome.

We also analyzed the 80 loci in 20 Przewalski's horses, in three individuals from *E. asinus* and in one individual each from *E. burchellii, E.grevyi, E. zebra hartmannae, E. kiang* and *E. hemionus onager*, respectively (Fig. 2); since the results of the three *E. asinus* individuals were identical, only one column is reported in Fig. 2. As shown in Fig. 3, from the evolutionary point of view, ERE1 loci can be classified in three groups: elements which are conserved in all species of the genus *Equus* (53 loci) and thus were inserted in a common ancestor of all extant equids, at least 3.8 Ma ago (Mya); elements which are conserved in all analyzed horses (*E. caballus* and *E. przewalskii*) but absent in the other *Equus* species (25 loci), thus inserted after the separation of the horse lineage, that is about 3.8 Mya [58, 59]; elements which are present in *E. caballus* only (two loci: 11 and 35 in Fig. 2) and therefore were probably inserted after the separation of the two horse species. To this regard, it must be pointed out that, in the middle of the twentieth century, Przewalski's horses were close to extinction and the extant population derives from a very limited number of individuals [60]; therefore, the absence of an ERE1 element in Przewalski's horses may be related either to the date of its insertion or to genetic drift. Nine loci (number 1, 6, 9, 11, 13, 15, 28, 35, 51 in Fig. 2) are polymorphic in one or both horse species and absent in the other species, suggesting that these insertions occurred in a relatively recent evolutionary time, after the separation of the horse lineages, and are not yet fixed.

In conclusion, these results showed that the fraction of ERE1 insertions conserved in all *Equus* species increases with the decrease of their identity to the consensus (Fig. 3): only 3 out of the 20 horse ERE1 elements with 98–100 % identity were present in the other species while 13, 17 and 20 loci out of 20 were conserved in the classes with 95, 90 and 85 % identity, respectively (Fig. 3). On the contrary, the majority of ERE1s that are present in the horse lineage only (16/20) share a high identity to the consensus (98–100 %). The loci that were conserved in all *Equus* species were not polymorphic in the horse (Fig. 2) confirming that they were inserted earlier during evolution, in a common ancestor of the extant *Equus* lineages. Since only three individuals from *E. asinus* and one individual from *E. burchellii, E. grevyi, E. zebra hartmannae, E. kiang* and *E. hemionus onager* were analyzed, we cannot exclude that, at some ERE1 loci, insertion polymorphism may be present in one or more *Equus* species, however, the results confirm that the level of identity to the consensus not only is related to their polymorphism but is also indicative of their evolutionary age. Therefore, ERE1 insertion polymorphism can be used for evolutionary analyses and population studies.

Position of ERE1 loci relative to genes

Since transposable elements, when inserted within or near genes, may influence gene expression, we used an algorithm developed in our laboratory (see Material and Methods) to classify ERE1 elements according to their position relative to genes. The coordinates of the horse genes were obtained using the tool "UCSC Table Browser" [61, 62]. Horse genes are poorly mapped, therefore we included in the analysis the coordinates of putative horse genes listed in a table generated by UCSC, based on homology with human and bovine genes. The results (Fig. 4) showed that 45.4 % of ERE1 elements were located inside introns of validated or putative genes. The fraction of the human genome occupied by introns has been estimated to be between 26 and 38 % [2, 63–67]; since no data are available for the horse, we are unable to conclude whether the fraction of ERE1 elements contained within introns is simply due to random insertion. Given the high number of ERE1 elements within introns, it is possible that some have acquired a functional role by modifying the splicing pattern as documented for other SINEs [68–70]. The remaining ERE1s (54.6 %) were located at variable distances from genes. Our data suggest that there are no hotspots for ERE1 integration sites in the horse genome and that insertion events may have occurred at random. Counter-selection may be responsible for the lack of insertions within exons. Moreover, only 170 ERE1 insertions (0.5 %) were found at less than 1 kb from the 5' end of validated or putative genes suggesting that some of them may affect gene expression.

Sequence organization of the myostatin gene promoter and mechanism of ERE1 insertion

As mentioned above, a polymorphic ERE1 insertion was identified at the myostatin locus [29]. In Fig. 5, the wild type myostatin locus (Fig. 5a), the ERE1+ allele (Fig. 5e), and a model for the transposition mechanism (Fig. 5b–d) are shown. At the wild type myostatin locus, the regulatory elements, located upstream and in close proximity of the putative transcription start site (Fig. 5a), comprise: two TATA boxes (TATA box1 and 2, located 24 and 1 bp upstream the transcription start site, respectively) and one CAAT box (70 bp upstream the transcription start site). In addition, two E-boxes (E1 and E2), which are muscle gene control elements [71, 72], are located 49 and 16 bp upstream the transcription start site, respectively. Given their position relative to the putative transcription start site, the TATA Box 1 and the CAAT box are likely to constitute the core promoter directing transcription of the horse wild type myostatin gene.

Sequence comparison of the wild type and ERE1+ alleles suggested that this insertion may have occurred according to the previously proposed mechanism of SINE elements retrotransposition in the human genome leading to a direct duplication of the target site [16, 73, 74]. According to this model, during the first step of the process (Fig. 5a), the target site was cleaved inside the TATA box 1 (black arrowhead); the 3' end of the ERE1 RNA (light blue) annealed through microhomology to the single-stranded 5'-TTTTT-3' sequence generated after the nick in the TATA box 1 (Fig. 5b). The free 3'OH group created after the cleavage was then used to prime the reverse transcription of the ERE1 RNA and synthesize the first strand of the cDNA (dark blue, Fig. 5b). The second strand of the DNA was then cleaved one bp downstream the E-box E2 (black arrowhead, Fig. 5c), producing a 3' end that was used to prime the synthesis of the second strand of the ERE1 DNA (Fig. 5d). Through a gap filling reaction, the entire ERE1 sequence was integrated into the myostatin promoter with the formation of the Target Site Duplication. Fig. 5e shows the ERE1+ allele of the myostatin promoter obtained as a result of the retrotransposition event. The inserted ERE1 (dark blue) is located 29 bp upstream the transcription start site. The size of the Target Site Duplication (14 bp) falls into the range described for SINE elements in the human genome [16, 73, 74]. The consequence of the ERE1 insertion was a modification of the core promoter with the formation of a variant TATA Box 1 and the displacement of the CAAT box. This rearrangement likely affects the strength of the core promoter.

Reporter gene assay of the two variants of the myostatin gene promoter

To test the hypothesis that the ERE1 insertion alters the expression of the myostatin gene, we performed a reporter gene assay using a plasmid containing the enhanced Green Fluorescent Protein (*eGFP*) gene and the puromycin resistance gene. The two variants of the myostatin promoter (ERE1+ and ERE1-) were cloned from the genomic DNA of a heterozygous Thoroughbred horse and inserted into the plasmid cloning site upstream of the *eGFP* reporter gene. The ERE1- variant plasmid contained a 2042 bp genomic fragment comprising 31 bp from the myostatin UTR; the ERE1+ plasmid contained an insert differing from the previous one only for the ERE1 insertion.

To test whether the ERE1 insertion can affect promoter strength the two plasmids were transfected in human HeLa cells and in a horse fibroblast cell line that we immortalized using the procedure described in Vidale et al. [75]. Since transfection efficiency in horse fibroblasts is extremely low (3–5 %), transient short term transfections could not be performed. Long-term selection with puromycin had to be carried out in order to isolate stably transfected cell populations. The expression of *eGFP* was evaluated by fluorescence microscopy, western blotting and quantitative real-time PCR (Fig. 6). Both in human and in horse cells, the ERE1 insertion caused a reduction of eGFP fluorescence signals to almost undetectable levels (Fig. 6a). The effect of the insertion on promoter strength was also demonstrated by immunoblotting of protein extracts with an anti-eGFP antibody (Fig. 6b): while a strong band could be detected in protein extracts from cells transfected with the plasmid containing the ERE1- promoter, only a very faint band could be observed in extracts from cells transfected with the ERE1+ plasmid. We then carried out a quantitative real-time PCR reaction using eGFP specific primers (Additional file 2: Table S4B) to amplify reverse transcribed mRNA from the transfected cell lines (Fig. 6c): in human cells transfected with the ERE1+ plasmid the expression level of the reporter gene showed a 6.4-fold reduction compared with that observed in cells transfected with the vector carrying the ERE1-

promoter; similarly, a 4.9-fold reduction was observed in horse fibroblasts. These results demonstrate that the ERE1 insertion affects the ability of the myostatin gene promoter to drive transcription of a reporter gene and strongly suggest that the myostatin gene may be underexpressed in horses containing this variant promoter sequence.

ERE1 insertion polymorphism at the myostatin locus: sport aptitude and racing performance

Given the role of myostatin in the regulation of muscle development and considering the relevance of muscular mass in athletic performance, we wondered whether the genotype of horses relative to the ERE1 insertion may influence their sport aptitude and racing abilities.

Using primers flanking the myostatin gene promoter (Additional file 2: Table S4B), we set up a PCR assay to identify the two alleles: the ERE1 containing allele, ERE1 +, produces a 441 bp band, while the allele lacking the insertion, ERE1-, produces a 214 bp band. We then analyzed the frequency of the two alleles, in 5 horse breeds (Quarter Horse, Andalusian, Lipizzaner, Norwegian Fjord and Icelandic Pony) and in Przewalski's horse. As shown in Table 1A, in Quarter horses, although the number of individuals analyzed is limited (20), the frequency of the ERE1+ allele seems particularly high (57 %). In the Andalusian breed, the ERE1+ allele was observed only in 3 heterozygous individuals, while in the other breeds and in Przwelaski's horse the ERE1+ variant was not present. Since the ERE1 insertion was present only in horse populations in which Thoroughbred blood is known to have been introduced (Quarters, Andalusians, Show Jumpers), it is likely that it appeared recently in the horse lineage and probably occurred in a Thoroughbred ancestor, as previously suggested [46].

Although the number of individuals tested for each breed is relatively small (19–23 animals per breed), the striking frequency variation of the two alleles suggests that the two variants may have been under selection during the establishment and improvement of some breeds in relation to specific aptitude and performance traits. In particular, the high frequency of ERE1+ alleles in Quarter horses suggests that this variant may favor the ability of sprinting short distances. To this regard, it is important to point out that the name of this breed came from its excellence in races of a quarter mile or less.

Therefore, to test the hypothesis that the ERE1 insertion at the myostatin locus may affect the aptitude for specific sport abilities, we initially analyzed the frequency of the two allelic variants in 30 horses competing in show-jumping at various levels, in 90 horses registered in the Italian Trotter studbook, bred for harness racing, and in 75 horses registered in the Italian Thoroughbred studbook mainly bred for flat racing (Table 1B). Although Italian Trotters derive from English Thoroughbred stallions crossed with mares of different origins, and Thoroughbreds have been introduced in several bloodlines of Show Jumpers, the allelic frequencies in the three groups were strikingly different (Table 1B): the ERE1+ allele was completely absent in the Trotters and, in the Show Jumpers, only one individual was heterozygous for the variant; on the contrary, among the flat racing horses, the percentage of ERE1+ alleles was 43. These observations suggest that the ERE1+ allele may have been selected in the Thoroughbreds and in the Quarter Horses together with flat racing aptitude.

To test whether the ERE1+ variant may influence racing performance in the Thoroughbreds, we selected a group of 117 elite horses classified in the top three places in at least one high level race in Italy in the period ranging from 2005 to 2011. In this selected group, the ERE1+ allele was significantly more frequent compared to the general Thoroughbred population ($p = 9.31 \times 10^{-6}$, Table 1 B). To test whether the ERE1 insertion influences performance relatively to race distance, the elite horses were grouped according to Best Race Distance, defined as the distance of the highest grade race won. When multiple races of the same grade were won, the distance of the race with the most valuable prize was considered. The results of this analysis are shown in Fig. 7: in short distance races (1000 and 1200 m), the majority of winning horses (18 out of 30) were homozygous for the ERE1+ allele and no homozygous individuals for the ERE1- allele were found; in the long distance races (>2000 m), only heterozygotes and ERE1- homozygotes were observed and, in medium distance races (1400–2000 m), all the three genotypes were represented although the ERE1+ homozygotes were relatively more frequent in the groups winning up to 1600 m races compared to horses winning 1700–2000 m races. When the genotypic frequencies in horses winning short distance (1000–1200 m), medium distance (1400–2000 m) and long distance (>2000 m) races were compared, the differences were highly significant ($p = 1.94 \times 10^{-6}$).

Since the ERE1+ variant is associated with better performance in short distance races, it may have been artificially selected through breeding, consequently, its frequency increased in the Thoroughbred population, although it was not fixed. The empty allele might also have been subjected to artificial selection. Thoroughbreds are also used for long distance races, in which individuals homozygous for the ERE1- alleles have the best performance while heterozygous animals seem to be advantaged in average distance races. It should be pointed out that among the Italian Trotters, a breed derived from English Thoroughbreds, no ERE1+ allele was identified. This is probably due to the fact that Italian Trotters are bred for harness racing at a trot gait in relatively long distance races and this artificial selection led to the loss of the ERE1+ allele. Finally, although Quarter Horses derive from the crossing of Thoroughbreds with horses from other breeds, the frequency of the ERE1+ allele was even higher than in the Thoroughbreds themselves (Table 1); this observation can be related to the fact that these horses have been selected for their sprinting ability in flat races of a quarter mile or less.

As mentioned in the introduction, the g.66493737C > T SNP in the first intron of the myostatin gene was shown to be predictive of athletic performance [29, 37]: C/C horses are suited for short-distance, C/T for middle-distance and T/T for long-distance races. Comparing the ERE1 and the g.66493737C > T genotypes (Fig. 7), we observed that in 112 out of 117 horses the two genotypes were concordant, with the C SNP allele associated with ERE1+ and the T SNP allele associated with the ERE1- promoter. These results show that the two polymorphic loci are tightly linked, as expected by their close proximity in the genome (1605 bp). Although the ERE1 insertion was previously described [37], its influence on myostatin gene expression was not investigated. In the present work, we demonstrate that the ERE1 insertion affects gene expression supporting the hypothesis that this is the genotype that drove selection [46]. In particular, we showed that the ERE1 insertion causes a 5–6 fold decrease in the

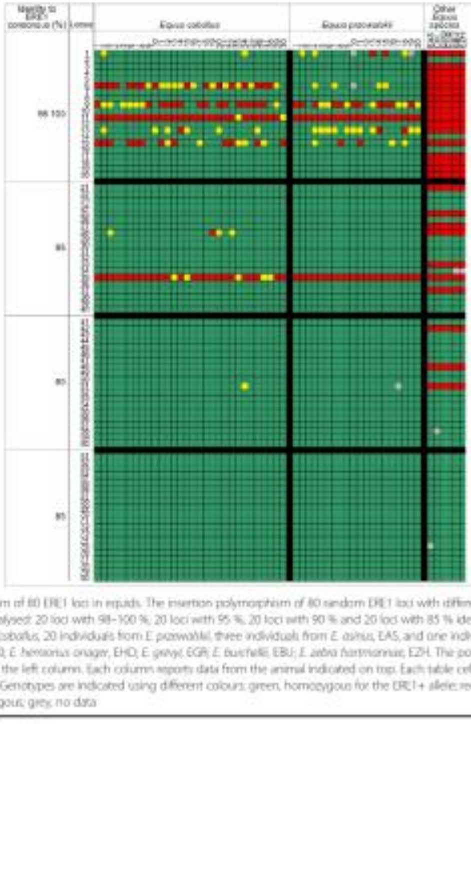

Fig. 2 Insertion polymorphism of 80 ERE1 loci in equids. The insertion polymorphism of 80 random ERE1 loci with different percentage of identity to the ERE1 consensus were analysed: 20 loci with 98–100 %, 20 loci with 95 %, 20 loci with 90 % and 20 loci with 85 % identity. The analysis was carried out in 30 individuals from *E. caballus*, 20 individuals from *E. przewalskii*, three individuals from *E. asinus*, EAS, and one individual from each one of the following species: *E. kiang*, EKI; *E. hemionus onager*, EHO; *E. grevyi*, EGR; *E. burchelli*, EBU; *E. zebra hartmannae*, EZH. The position of each locus in the horse genome is reported in the left column. Each column reports data from the animal indicated on top. Each table cell shows the genotype of an individual at a specific locus. Genotypes are indicated using different colours: green, homozygous for the ERE1+ allele; red, homozygous for the ERE1- allele; yellow, heterozygous; grey, no data

transcription of the reporter gene (Fig. 6), providing the first example of a SINE element influencing gene expression in the horse genome.

Although the g.66493737C > T SNP showed an association with racing performance [29], this sequence variation does not provide an immediate functional explanation of this trait. On the contrary, our experimental data strongly suggest a direct influence of the ERE1 insertion on myostatin expression. Since the g.66493737C > T SNP is located only 1605 bp away from the ERE1 insertion site in the promoter, the ERE1 insertion, rather than the g.66493737C > T SNP (located in the first intron), may functionally influence racing performance, the two polymorphisms being in linkage disequilibrium ($r^2 = 0.73$) as previously observed [29, 46]. In other words, the results presented here on myostatin expression provide a physiological interpretation of the correlation between ERE1 insertion and racing performance; moreover, the previously described correlation among the g.66493737C > T SNP, muscle mass [43] and muscle fiber composition [46] can also be reinterpreted on the basis of the linkage disequilibrium between the two polymorphic loci.

Conclusions

In the work presented here we provide a catalogue of the most abundant SINE retrotransposons, ERE1, in the horse genome. Through the analysis of sequence conservation, insertion polymorphism and presence in other equids, we provide an evolutionary dating of ERE1 elements appearance in the *Equus* lineage. Therefore, similarly to other mammalian SINE elements, ERE1 insertion polymorphism can be used for evolutionary analyses and population studies.

The analysis of ERE1s position relative to genes suggests that some may have acquired a functional role by modifying the splicing pattern, when interrupting an intron, or by altering gene expression, when inserted inside regulatory regions. To this regard, we studied the effect of an ERE1 insertion in the promoter of the myostatin gene showing that it causes a reduction of promoter strength in a reporter gene assay. Therefore, we suggest that this ERE1 insertion may decrease the levels of myostatin thus modifying muscle development.

The ERE1 insertion at the myostatin locus is polymorphic in the horse population and seems to be related to specific racing aptitude, the ERE1+ allele being particularly common in breeds characterized by sprinting ability, such as the Quarter Horse, and absent in other breeds, such as the Italian Trotter, which are used for long distance racing. In a sample of Thoroughbred elite horses, classified in the top three places in at least one high level race in Italy, we observed a statistically significant correlation between the ERE1+ variant and good performance in short distance races; on the other hand, the empty allele was more frequent in Thoroughbreds winning long distance races. We propose that the two variants have been unwittingly selected by breeders in order to obtain horses with specific racing abilities.

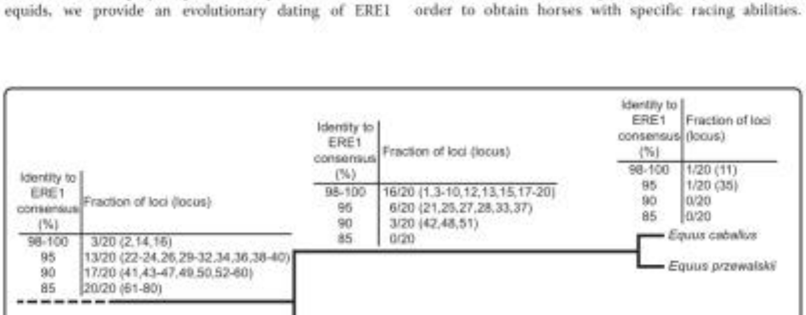

Fig. 3 Phylogenetic tree of equids. The time of insertion of each one of the 80 ERE1s is marked on the phylogenetic tree (adapted from [58, 59]). ERE1 loci are classified according to the percentage of identity to the consensus sequence, the fraction of inserted loci in each class of identity is shown. Each ERE1 is indicated by a unique locus number (see Fig. 2 and Additional data file 1: Table S3A). The lineage "Other *Equus* species" comprises the following non-horse species: *E. asinus*, *E. kiang*, *E. hemionus onager*, *E. burchelli*, *E. grevyi*, *E. zebra hartmannae*

Table 1 ERE1+ and ERE1- genotyping at the myostatin locus

			Number of alleles (%)		Homozygous individuals (%)		Heterozygous individuals (%)
		Number of individuals	ERE1+	ERE1-	ERE1+/+	ERE1-/-	ERE1+/-
A	Quarter Horse	20	23 (57.5)	17 (42.5)	9 (45)	6 (30)	5 (25)
	Andalusian	20	3 (7.5)	37 (92.5)	0	17 (85)	3 (15)
	Lipizzaner	23	0	46 (100)	0	23 (100)	0
	Norwegian Fjord	20	0	40 (100)	0	20 (100)	0
	Icelandic Pony	19	0	38 (100)	0	19 (100)	0
	Przewalski's Horse	20	0	40 (100)	0	20 (100)	0
B	Show Jumpers	30	1 (1.7)	59 (98.3)	0	29 (96.7)	1 (3.3)
	Italian Trotters	90	0	180 (100)	0	90 (100)	0
	Unselected Italian Thoroughbreds	75	65 (43.3)	85 (56.7)	18 (24.0)	28 (37.3)	29 (38.7)
	Elite Italian Thoroughbreds	117	135 (57.7)	99 (42.3)	33 (28.2)	15 (12.8)	69 (59.0)

(A) Analysis of individuals from five breeds of the domestic horse and from Przewalski's horse. (B) Analysis of individuals bred for different sport aptitude.

Our results indicate that, although racing performance is certainly influenced by environmental factors, like training and nutrition, and by several genetic factors, breeding schemes may also take into account the differential effect of these two ERE1 allelic variants.

Methods

Ethics statement

Horse blood and hair samples were collected in the stables where the animals were kept, during veterinary practices carried out for routine clinical analysis, animal care or registration requirements. Since blood samples were not collected for experimental purposes, according to the Italian law (Decreto Legislativo 4/03/2014 n.26), the procedures do not require approval by an ethical committee. Written consent from the owners was not required because the identity of horses and owners cannot be established from the data presented in this work. DNA samples from endangered Equus species were shipped to Italy from the San Diego zoo together with the appropriate international CITES permit. Horse 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.

Preliminary *in silico* analysis of the polymorphism of the four ERE subfamilies

The consensus sequences of the ERE subfamilies ERE1 (accession number: D26566) [53], ERE2 [76], ERE3 [77], ERE4 [78] were downloaded from Repbase [27, 28] and used as queries for a BLAT search against the horse genome reference sequence (September 2007 Broad/equCab2.0 assembly) [51, 52]. For each ERE subfamily the 200 loci with the highest identity to their consensus sequence were identified. Their sequence was used as query for a BLAST search against the horse Trace

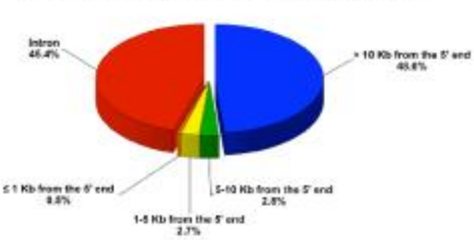

Fig. 4 Distribution of ERE1 elements relative to genes. The percentage of ERE1 loci located in introns of validated or putative genes (red) and in non-genic regions is indicated. ERE1 elements located in non-genic regions are classified according to their distance from the 5' end of the nearest gene (>10 Kb, 5–10 Kb, 1–5 Kb, ≤ 1 Kb)

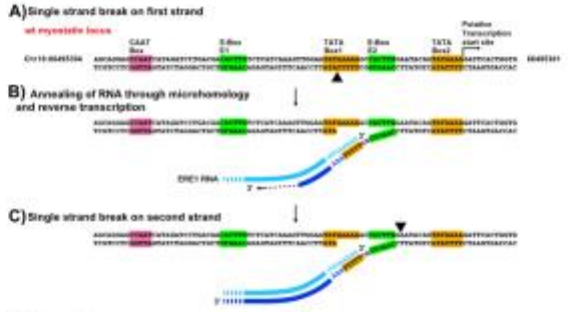
A) Single strand break on first strand
wt myostatin locus
CAAT Box
E-Box E1
TATA Box1
E-Box E2
TATA Box2
Putative Transcription start site
B) Annealing of RNA through microhomology and reverse transcription
ERE1 RNA
3'
C) Single strand break on second strand
3'
D) DNA synthesis
E) Formation of Target Site Duplications
ERE1+ myostatin locus
CAAT Box
E-Box E1
TATA Box1
E-Box E2
ERE1
Mutated TATA Box1
E-Box E2
TATA Box2
Putative Transcription start site
Target Site Duplication
213 bp
Target Site Duplication

Fig. 5 Model for ERE1 integration via retrotransposition at the myostatin promoter. **a** Sequence of the empty wild type myostatin locus, the coordinates of the locus in the horse genome reference sequence (equCab2.0) are indicated on both sides. The promoter elements located upstream the transcription start site are shown: CAAT-box (pink), E-boxes (green), TATA boxes (yellow). The black arrowhead points to the position of the single strand break. **b** Annealing through microhomology (TTTTT/AAAAA) of the ERE1 RNA (light blue) to the single stranded DNA end originated after the nick and reverse transcription. The cDNA produced by retrotranscription is shown in dark blue. **c** Cleavage of the second DNA strand (black arrowhead). **d** Synthesis of the second strand of ERE1 cDNA and gap filling. **e** Sequence of the ERE1+ myostatin locus: the ERE1 element is integrated (dark blue) and the target site is duplicated (boxed nucleotides)

database [54], which is a collection of short sequences (<1 Kb) generated during large-scale sequencing projects. From the Trace database we selected the dataset Equus caballus-WGS, which contains reads that were not included in the final assembly of the horse genome reference sequence. We then used the sequences flanking each ERE insertion as query to search for traces corresponding to the same loci but lacking the ERE insertion (empty alleles).

Search of ERE1 loci characterized by insertion polymorphism in the horse genome reference sequence

Our preliminary search, based on the analysis of 200 loci from each ERE subfamily, showed that ERE1s have the highest proportion of empty alleles. We then focused further analyses on this subfamily.

In order to obtain a comprehensive catalog of ERE1 polymorphic loci in the horse genome reference sequence, we developed a pipeline using the C# programming language (Microsoft Visual Studio 2008) and Microsoft SQL Server 2008 as the database management system. The ERE1 consensus sequence downloaded from RepBase (accession number D26566) [53] was used as query for a BLAST search against the horse genome reference sequence (September 2007 Broad/equCab2.0 assembly) [79]. The BLAST search was performed using "megablast" as optimization algorithm and standard search parameters. Results were downloaded as hit table. Only the loci with identity to the consensus greater than 84 % were considered. To exclude loci that were subject to deletions or insertions, only the hits with length similar to that of the ERE1 consensus sequence (225 ± 10 bp) were considered. Since the coordinates of the hits inside the table were referred to contig sequences, they were converted into genomic coordinates using the conversion table "seq_contig.md" at [80]. ERE1s located inside unplaced regions were discarded. Since our method is based on similarity, ERE1s inserted inside other transposons could give rise to false positive hits because several uninterrupted transposons are scattered through the genome. Therefore, before starting the search for polymorphic loci we identified and discarded ERE1 elements inserted inside other transposons. To this purpose, we downloaded the list of the horse transposable elements

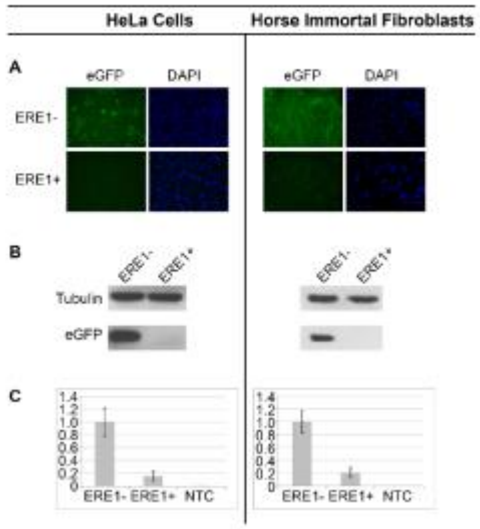

Fig. 6 Reporter gene assay. The reporter gene assay (*eGFP expression*) was carried out in human HeLa cells (left) and in horse immortalized fibroblasts (right). **a** Fluorescence microscopy images of cells transfected with the two constructs containing the *eGFP* gene under the control of the ERE1- (top) and ERE1+ (bottom) promoter. DAPI-staineing is shown on the right of each image. **b** Western blots with anti-tubulin (loading control, top) and anti-eGFP (bottom) antibodies using protein extracts from cells transfected with the plasmid containing the ERE1- (left) and the ERE1+ (right) promoter. **c** Quantification of *eGFP* expression by quantitative RT-PCR. The expression levels of the eGFP transcript are indicated in arbitrary units. eGFP levels in cells transfected with the ERE1- plasmid were used as reference and set to 1.0. NTCs, no-template controls

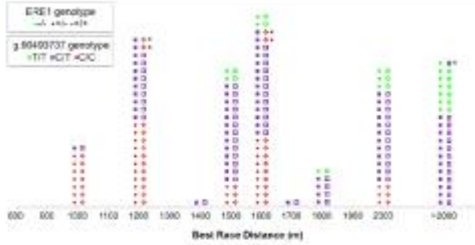

Fig. 7 Genotyping of elite Thoroughbred horses. Elite Thoroughbreds (*n* = 117) were classified according to the distance of the highest grade race won (Best Race Distance), expressed in meters. For each individual, the myostatin promoter (full symbols) and the SNP g.66493737C > T (empty symbols) genotypes are shown. The majority of the ERE1−/−, ERE1+/− and ERE1+/+ individuals have a T/T, C/T and C/C SNP genotype, respectively; the five individulas with a different combination of genotypes at the two loci are marked with an asterisk

from the site UCSC Genome Bioinformatics using the tool "Table Browser" [61, 62]. The list of transposons is found in the data table called "rmsk" (Group "Variation and Repeats", Track "RepeatMasker") that was generated using the software RepeatMasker [81] during the horse genome sequencing project [24]. The coordinates of each ERE1 were compared with those of the boundaries of other transposable elements. If an ERE1 interrupted a repetitive element the locus was discarded.

To identify empty alleles, for each locus we downloaded a 2.2 Kb sequence from UCSC Genome Browser [24, 82, 83] containing the transposon (about 225 bp), 1 Kb from the 5' flanking region and 1 Kb from the 3' flanking region. These sequences were then used as queries for a BLAST search [54] against the horse "Traces – WGS sequence" database. The BLAST search was performed using "megablast" as optimization algorithm and standard search parameters. If the hit contained a 225 ± 10 bp gap and was at least 98 % identical to the sequences flanking the transposon, it was considered an ERE1- locus. Only traces from the reference genome of Twilight were considered identifying them as belonging to "center_project number" G836. The specificity of each trace sequence was manually checked using BLAT [51, 52] and MultAlin [84, 85]. In order to focus on the loci inserted in single copy sequences, the ERE1 loci that were found at multiple positions during the BLAT search, and were probably located inside segmental duplications, were discarded. The complete list of single copy polymorphic ERE1 loci and the accession codes of the traces (Trace id) corresponding to the empty alleles is reported in Additional file 2: Table S2.

In silico localization of ERE1 elements relative to genes

The position of ERE1 elements relative to horse genes was defined using the genomic coordinates of known horse validated and putative genes. Horse validated genes and their coordinates are listed in the data table "refGene" (assembly "Sep. 2007") downloaded from the site UCSC Genome Bioinformatics using the tool "Table Browser" [61, 62]. The "refGene" table contains, among other information, the name of each gene, the coordinates of the transcription start and stop sites, the coordinates of the boundaries of each exon. Since the number of known horse genes is relatively small, we also included in the search the genomic coordinates of putative genes defined by sequence homology with those from human and bovine as listed in the data table "Other RefSeq (xenoRefGene)". The data table (xenoRefGene) was downloaded from using the tool "Table Browser" [61, 62] and was used to define the coordinates of the beginning and end of putative genes in horse that are orthologous to those from human and bovine. This track was prepared by the UCSC genome browser group as described in the information page (https://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=442242277_zw0eu9-Hy93E8wIE62c8BxvE3BJox&c=chr11&g=xenoRefGene): as stated in the information page, this track shows known protein-coding and non-protein-coding genes for organisms other than horse. The RNAs were aligned against the horse genome using blat. This track was produced at UCSC from RNA sequence data generated by scientists worldwide and curated by the NCBI RefSeq project.

Genomic DNA samples

Genomic DNA was extracted from blood or hair samples, or from cultured primary fibroblasts using standard protocols. The 30 *E. caballus* samples shown in Fig. 2 derive from: peripheral blood of 22 show jumping horses which, according to their pedigree chart, do not share common ancestors up to the third generation (they were also used for the analysis of the myostatin gene polymorphism shown in Table 1, see below); fibroblast cell lines established from the skin of 8 slaughtered animals which were shown to be unrelated by microsatellite analysis as described in [86]. The *E. asinus* samples derive from fibroblast cell lines established from the skin of 3 slaughtered animals. The *E. grevyi* sample derives from a fibroblast cell line purchased from Coriell Repositories and *E. burchellii* fibroblasts were a kind gift from Mariano Rocchi (University of Bari, Italy) [50, 87]. *E. zebra hartmannae*, *E. kiang* and *E. hemionus onager* fibroblasts were provided by Oliver Ryder (Genetics Division of San Diego Zoo, San Diego, California, USA) [48]. DNA samples from Quarter Horses, Andalusian, Norwegian Fjord, Icelandic Ponies (Table 1) and *E. przewalskii* (Fig. 2 and Table 1) were provided by Cecilia Penedo (UC Davis, California, USA). Lipizzaner DNA samples (Table 1) were described in [88]. The 30 Show Jumpers in Table 1, which comprise the 22 *E. caballus* individuals of Fig. 2, were animals kept in Italian sport riding stables and competing at the National and International level; they derived from different stud farms in Italy, France, Germany, Holland, Belgium and were chosen by the owners for their show jumping aptitude. Genomic DNA from Italian Trotters and Italian Thoroughbreds was extracted from blood spotted on FTA® filter papers (Whatman Bioscience, Cambridge, UK). All samples came from horses belonging to the Italian Stud Book of MiPAAF (Ministero Delle Politiche Agricole Alimentari e Forestali). The performance information were provided by ANAC (Associazione Nazionale Allevatori Cavalli Purosangue).

PCR and SNP analysis

Eighty ERE1 insertions with different degrees of identity relative to the consensus sequence were randomly selected from the list of 27,396 loci obtained by *in silico* analysis. The coordinates of the 80 loci are reported in Additional file 2: Table S4A together with the sequence of the primers deduced from the sequences flanking the transposon (Additional file 2: Table S4A). Twenty ng of genomic DNA were used as template for PCR experiments performed in a 10 µl-final volume with 8 pmoles of each primer, 0.2 mM dNTPs, 1× Green Buffer (Promega) and 0.4 units of GoTaq DNA polymerase (Promega). After a denaturation step at 95 °C for 2 min, the following amplification cycle was performed 3 times: 95 °C for 50 s, appropriate annealing temperature (Additional file 2: Table S4A) for 45 s, 72 °C for 1 min. The first 3 cycles were followed by 27 cycles: 95 ° C for 30 s, appropriate annealing temperature for 35 s, 72 °C for 1 min. Final extension was carried out at 72 °C for 5 min. PCR products were checked by electrophoresis in 1 % agarose gel.

To analyze the ERE1 insertion polymorphism at the myostatin promoter, we amplified genomic DNAs using primers from the sequences flanking the insertion site (MyostProm-F0 and MyostProm-R, Additional file 2: Table S4B). The expected length of the PCR products from the ERE1+ and the ERE1- alleles were 441 and 214 bp, respectively. The reactions were carried out as described above.

The Analysis of SNP g.66493737C > T was performed using the "Custom TaqMan SNP Assay" (Applied Biosystems) on a 7500 Fast Real Time PCR Instrument.

Preparation of plasmids for reporter gene assay

In order to clone the entire promoter and the transcription start site of the myostatin gene we PCR-amplified the locus chr18:66495283–66497324 (equCab2.0) from the genomic DNA of a horse heterozygous for the ERE1 insertion.

PCR reaction was performed using the primers MyostProm-F and MyostProm-R (Additional file 2: Table S4B), which contain *Hind*III and *Bam*HI restriction sites, respectively. After a denaturation step at 95 °C for 2 min, the following amplification cycle was repeated for 30 times: 94 °C for 40 s, 65 °C for 40 s and 72 °C for 4 min. The final extension was carried out at 72 °C for 10 min. The reaction products corresponding to the ERE1- and the ERE1+ allele (2058 and 2285 bp, respectively) were separated by electrophoresis on 1 % agarose gel and purified using the Wizard SV Gel and PCR Clean-Up System (Promega). The two alleles differed only for the presence of the ERE1 element and the target site duplication (see Fig. 3).

The purified PCR products were digested with *Hind*III and *Bam*HI and then cloned, upstream of the enhanced Green Fluorescent Protein (*eGFP*) cDNA, into an expression vector that was previously constructed in our laboratory [89]. Our vector contains the puromycin and ampicillin resistance genes. All constructs were checked by Sanger sequencing.

Cell culture and transfection

Horse Immortal Fibroblasts [75] and HeLa (human cervical carcinoma) cells were cultured in high-glucose D-MEM supplemented with 10 % fetal calf serum (Euroclone), 2 % non-essential amino acids, 2 mM L-glutamine and 1× penicillin-streptomycin (Sigma). For primary fibroblast cell lines, the culture medium was supplemented with 20 % fetal calf serum. Cells were routinely cultured at 37 °C in 5 % CO_2.

Plasmid DNA for promoter reporter assays was prepared using QIAGEN Plasmid Midi kit. Transfections were carried out using the Lipofectamine 2000 reagent (Invitrogen) according to the manufacturer's protocol.

Twenty-four hours post-transfection, cells were selected adding 300 ng/ml (horse immortal fibroblasts) or 1 µγ/ml (HeLa cells) puromycin to the medium. Cells were cultured with selective medium until the emergence of drug-resistant colonies, that is 3 weeks for horse fibroblasts and 2 weeks for HeLa cells. Pools of about 50 colonies were obtained and grown as stably transfected cell populations.

Western Blot experiments

Protein extracts were prepared from samples three million cells as follows: the cells were washed twice with ice cold 1xPBS, resuspended in lysis buffer (50 mM Tris–HCl pH 6.8, 86 mM β-mercaptoethanol, 2 % SDS) and boiled for 10 min. Proteins were separated by 10 % SDS-PAGE and transferred to nitrocellulose membranes (Amersham Protran Premium 0.45 µm NC) through wet transfer. Membranes were incubated with a rat monoclonal antibody against eGFP (Chromotek, code 3H9), diluted 1:1000, and with a mouse monoclonal antibody against tubulin (NeoMarkers, Ab-4, code MS-719-P1ABX), diluted 1:3000. Secondary antibodies, conjugated to horseradish peroxidase, were a chicken anti-rat IgG-HRP (Santa Cruz Biotechnology, code sc-2956), diluted 1:5000, and an ImmunoPure goat anti-mouse monoclonal (H + L) (Pierce, code 31430), diluted 1:10,000. Detection was performed using Immun-Star WesternC Kit (Bio-Rad) according to the manufacturer's protocol. Pre-incubation of the membranes and dilutions of the antibodies were performed in 1xPBS containing 0,05 % Tween20 and 7.5 % skim milk.

eGFP fluorescence analysis

Cells for eGFP fluorescence analysis were grown on coverslips (24 × 24 mm), washed with cold 1xPBS and fixed in 2 % paraformaldehyde in PBS for 10 min. Fixed cells were then stained with DAPI (4,6-

diamidino-2-phenylindole) and observed with a ZEISS Axioplan fluorescence microscope at 63× magnification. Pictures were captured using a CoolSNAP CCD camera (RS Photometrics) and processed using the software IPLab 3.5.5 (Scanalytics inc).

RNA preparation and quantitative RT-PCR

Total RNA from transfected HeLa and horse fibroblast cells was extracted using TRizol Reagent (Invitrogen) according to the manufacturer's protocol. The extracted RNA was purified using the RNA Clean & Concentrator-25 kit (Zymo Research) and treated three times with RQ1 RNase-free DNase (Promega).

For quantitative RT-PCR experiments we reverse transcribed 2.5 μg of total RNA using oligo-d(T)$_{17}$ primers and Revert Aid Premium First Strand cDNA synthesis kit (Fermentas) according to the manufacturer's protocol.

The cDNA was PCR amplified using GoTaq qPCR Master Mix (Promega) containing the appropriate oligonucleotides (Additional file 2: Table S4B). Oligonucleotides eGFP-F and eGFP-R were used to detect the eGFP transcript. *GAPDH* (glyceraldehyde 3-phosphate dehydrogenase, primer pair GAPDH-F and GAPDH-R) or *PRKCI* (protein kinase C iota, primer pair humcavPRKC-RealT-F and cavPRKC-RealT-R) were used as control genes for quantitative RT-PCRs carried out with the cDNA from HeLa cells or horse immortal fibroblasts, respectively. Each sample was prepared in triplicate. Negative controls (No template controls, NTCs) were included in the experiments. Reactions were carried out using an Opticon 2 System instrument (MJ Research). Cycling parameters comprised an initial denaturation at 95 °C for 2 min followed by 50 cycles at 95 °C for 15 s, 62 °C for 30 s and 72 °C for 30 s coupled to fluorescence detection. Experiments were repeated twice for each transfected cell line. Data were analyzed with the Opticon Monitor 3 software. Levels of expression were calculated using the standard ΔΔCq method [90], the level of expression in cells transfected with the plasmid containing the wild type allele was used as reference.

Statistical analysis

The correlation between the percentage of identity of the ERE1 loci and the natural logarithm of the frequency of polymorphic loci in each class was tested calculating Pearson's product moment correlation coefficient.

The significance of the difference of the allelic frequencies at the myostatin promoter in the populations of Elite and Unselected Thoroughbreds was tested using a Chi-Square test goodness of fit. The allelic frequencies in the 75 Unselected Thoroughbreds were adopted as expected values.

The significance of the correlation between the Best Race Distance and the genotype of the 117 Elite Thoroughbreds for the ERE1 insertion at the myostatin promoter was tested using a Chi-Square test for independence.

All statistical analyses were performed using R [91].

Availability of supporting data

The data sets supporting the results of this article are included within the article and its additional files.

Additional files

Additional file 1: Table S1.A lists the 45,713 loci identified using the ERE1 consensus sequence deposited at the RepBase database as query for a BLAST search against the horse genome reference sequence. **Table S1B** reports the 34,131 ERE1 loci with sizes similar to the ERE1 consensus (225 ± 10 bp) and with minimum identity to the consensus of 84 %. **Table S1.C** lists the 27,396 ERE1 loci that are not located inside other repetitive elements. (XLSX 3293 kb)

Additional file 2: Table S2. lists the ERE1 polymorphic loci identified in the horse reference genome sequence. **Table S3.** reports the frequency of ERE1 polymorphic loci in eight classes of ERE1 elements grouped according to consensus identity. The values reported in this table were used to draw Fig. 1. **Table S4A** lists the genomic position of the 80 ERE1 loci analysed in Fig. 2 and the sequence of the primers used for each locus. Table S4B lists the primers used to clone the myostatin promoter region and those used to perform quantitative RT-PCR experiments for reporter gene assay. (PDF 184 kb)

Abbreviations

BLAST: Basic Local Alignment Search Tool; BLAT: BLAST-Like Alignment Tool; DAPI: 4',6-diamidino-2-phenylindole; EAS: Equus asinus; EBU: Equus burchelli; eGFP: Enhanced Green Fluorescent Protein; EGR: Equus grevyi; EZH: Equus zebra hartmannae; EHO: Equus hemionus onager; EKI: Equus kiang; ERE: Equine Repetitive Element; GAPDH: Glyceraldehyde 3-phosphate dehydrogenase; LINE: Long INterspersed Element; Mya: Million years ago; NTC: No Template Control; NUMT: Nuclear sequences of MiTochondrial origin; PCR: Polymerase Chain Reaction; PRKCI: Protein kinase C iota; RT-PCR: Real Time Polymerase Chain Reaction; SINE: Short INterspersed Element; SNP: Single Nucleotide Polymorphism.

Competing interests

The authors declare that there are no competing interests.

Authors' contributions

MS carried out sequence analysis and alignment, performed some PCR experiments, contributed to data analysis, manuscript drafting and figure preparation. LK carried out gene reporter assays, including immunoassays and quantitative RT-PCR, and participated to figure preparation. RG carried out some PCR experiments, participated to gene reporter assays, performed the statistical analysis and participated to figure preparation. MB and AR carried out PCR and SNP analysis in racing Thoroughbreds and in Italian trotters. OK contributed to bioinformatic analysis. FMP and CR established fibroblast cell lines from horses and donkeys. FV and AM carried out some PCR experiments and sequence analysis. ER contributed to conception and design and to manuscript preparation. SGN carried out and supervised molecular and bioinformatics studies, contributed to the design of the study and to data analysis and interpretation. EG conceived the study, designed and coordinated the experiments, prepared the manuscript. All authors have read and approved the manuscript.

Acknowledgements

We thank Antonella Lisa (Istituto di Genetica Molecolare, CNR, Pavia) for helpful suggestions on the statistical analysis and Oliver Ryder (San Diego Zoo, Institute for Conservation Research, San Diego, CA, USA) for providing cell lines from *E. zebra hartmannae*, *E. kiang* and *E. hemionus onager*. We are grateful to Cecilia M. Penedo for providing DNA samples from horse breeds (Veterinary Genetics Laboratory, University of California, Davis, Old Davis Road, Davis, 95616 California, USA), Cesare Rognoni, DVM, (Equicenter Monteleone, Inverno e Monteleone, Pavia, Italy) and Simone Vignati, DVM, for their precious collaboration.

This work was supported by grants from the Ministero dell'Istruzione dell'Università e della Ricerca (MIUR-PRIN) and from the Consiglio Nazionale delle Ricerche (CNR-Progetto Bandiera Epigenomica).

Author details

[1]Dipartimento di Biologia e Biotecnologie "Lazzaro Spallanzani", Università di Pavia, Via Ferrata 1, 27100 Pavia, Italy. [2]Laboratorio di Genetica Forense Veterinaria, UNIRELAB srl, Via A. Gramsci 70, 20019 Settimo Milanese (MI), Italy.

Received: 4 August 2015 Accepted: 13 October 2015
Published online: 26 October 2015

References

1. Smit AF. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev. 1999;9:657–63.
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
3. Deininger PL, Moran JV, Batzer MA, Kazazian HH. Mobile elements and mammalian genome evolution. Curr Opin Genet Dev. 2003;13:651–8.
4. Kramerov DA, Vassetzky NS. Short retroposons in eukaryotic genomes. Int Rev Cytol. 2005;247:165–221.
5. Jurka J, Kapitonov VV, Kohany O, Jurka MV. Repetitive sequences in complex genomes: structure and evolution. Annu Rev Genomics Hum Genet. 2007;8:241–59.
6. Luchetti A, Mantovani B. Conserved domains and SINE diversity during animal evolution. Genomics. 2013;102:296–300.
7. Schmid CW, Jelinek WR. The Alu family of dispersed repetitive sequences. Science. 1982;216:1065–70.
8. Mighell AJ, Markham AF, Robinson PA. Alu sequences. FEBS Lett. 1997;417:1–5.
9. Batzer MA, Deininger PL. Alu repeats and human genomic diversity. Nat Rev Genet. 2002;3:370–9.
10. Roy-Engel AM, Carroll ML, El-Sawy M, Salem A-H, Garber RK, Nguyen SV, et al. Non-traditional Alu evolution and primate genomic diversity. J Mol Biol. 2002;316:1033–40.
11. Salem A-H, Kilroy GE, Watkins WS, Jorde LB, Batzer MA. Recently integrated Alu elements and human genomic diversity. Mol Biol Evol. 2003;20:1349–61.
12. Wang J, Song L, Gonder MK, Azrak S, Ray DA, Batzer MA, et al. Whole genome computational comparative genomics: A fruitful approach for ascertaining Alu insertion polymorphisms. Gene. 2006;365:11–20.
13. Carroll ML, Roy-Engel AM, Nguyen SV, Salem AH, Vogel E, Vincent B, et al. Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. J Mol Biol. 2001;311:17–40.
14. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, et al. Ultraconserved elements in the human genome. Science. 2004;304:1321–5.
15. Santangelo AM, de Souza FSJ, Franchini LF, Bumaschny VF, Low MJ, Rubinstein M. Ancient exaptation of a CORE-SINE retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene. PLoS Genet. 2007;3:1813–26.
16. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. Nat Rev Genet. 2009;10:691–703.
17. Okada N, Sasaki T, Shimogori T, Nishihara H. Emergence of mammals by emergency: exaptation. Genes Cells. 2010;15:801–12.
18. Speek M. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol Cell Biol. 2001;21:1973–85.
19. Wheelan SJ, Aizawa Y, Han JS, Boeke JD. Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. Genome Res. 2005;15:1073–8.
20. Matlik K, Redik K, Speek M. L1 antisense promoter drives tissue-specific transcription of human genes. J Biomed Biotechnol. 2006;2006:71753.
21. Druker R, Whitelaw E. Retrotransposon-derived elements in the mammalian genome: a potential source of disease. J Inherit Metab Dis. 2004;27:319–30.
22. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, et al. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature. 2006;441:87–90.
23. Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ, et al. Landscape of somatic retrotransposition in human cancers. Science. 2012;337:967–71.
24. 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.
25. Sakagami M, Ohshima K, Mukoyama H, Yasue H, Okada N. A novel tRNA species as an origin of short interspersed repetitive elements (SINEs). Equine SINEs may have originated from tRNA(Ser). J Mol Biol. 1994;239:731–5.
26. Gallagher PC, Lear TL, Coogle LD, Bailey E. Two SINE families associated with equine microsatellite loci. Mamm Genome. 1999;10:140–4.
27. RepBase. http://www.girinst.org/repbase/.
28. 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.
29. Hill EW, McGivney BA, Gu J, Whiston R, MacHugh DE. 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.
30. Szabó G, Dallmann G, Müller G, Patthy L, Soller M, Varga L. A deletion in the myostatin gene causes the compact (Cmpt) hypermuscular mutation in mice. Mamm Genome. 1998;9:671–2.
31. Grobet L, Martin LJ, Poncelet D, Pirottin D, Brouwers B, Riquet J, et al. A deletion in the bovine myostatin gene causes the double-muscled phenotype in cattle. Nat Genet. 1997;17:71–4.
32. McPherron AC, Lee SJ. Double muscling in cattle due to mutations in the myostatin gene. Proc Natl Acad Sci U S A. 1997;94:12457–61.
33. Clop A, Marcq F, Takeda H, Pirottin D, Tordoir X, Bibé B, et al. A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. Nat Genet. 2006;38:813–8.
34. Schuelke M, Wagner KR, Stolz LE, Hübner C, Riebel T, Kömen W, et al. Myostatin mutation associated with gross muscle hypertrophy in a child. N Engl J Med. 2004;350:2682–8.
35. Mosher DS, Quignon P, Bustamante CD, Sutter NB, Mellersh CS, Parker HG, et al. A mutation in the myostatin gene increases muscle mass and enhances racing performance in heterozygote dogs. PLoS Genet. 2007;3, e79.
36. Dall'Olio S, Fontanesi L, Nanni Costa L, Tassinari M, Minieri L, Falaschini A. Analysis of horse myostatin gene and identification of single nucleotide polymorphisms in breeds of different morphological types. J Biomed Biotechnol. 2010;2010:542945.
37. Hill EW, Gu J, Eivers SS, Fonseca RG, McGivney BA, Govindarajan P, et al. A Sequence Polymorphism in MSTN Predicts Sprinting Ability and Racing Stamina in Thoroughbred Horses. PLoS ONE. 2010;5, e8645.
38. Tozaki T, Miyake T, Kakoi H, Gawahara H, Sugita S, Hasegawa T, et al. A genome-wide association study for racing performances in Thoroughbreds clarifies a candidate region near the MSTN gene. Anim Genet. 2010;41 Suppl 2:28–35.
39. Baron EE, Lopes MS, Mendonça D, da Câmara MA. SNP identification and polymorphism analysis in exon 2 of the horse myostatin gene. Anim Genet. 2012;43:229–32.
40. Petersen JL, Mickelson JR, Rendahl AK, Valberg SJ, Andersson LS, Axelsson J, et al. Genome-wide analysis reveals selection for important traits in domestic horse breeds. PLoS Genet. 2013;9, e1003211.
41. Li R, Liu D-H, Cao C-N, Wang S-Q, Dang R-H, Lan X-Y, et al. Single nucleotide polymorphisms of myostatin gene in Chinese domestic horses. Gene. 2014;538:150–4.
42. McGivney BA, Browne JA, Fonseca RG, Katz LM, MacHugh DE, Whiston R, et al. MSTN genotypes in Thoroughbred horses influence skeletal muscle gene expression and racetrack performance. Anim Genet. 2012;43:810–2.
43. Tozaki T, Sato F, Hill EW, Miyake T, Endo Y, Kakoi H, et al. Sequence variants at the myostatin gene locus influence the body composition of Thoroughbred horses. J Vet Med Sci. 2011;73:1617–24.
44. Binns MM, Boehler DA, Lambert DH. Identification of the myostatin locus (MSTN) as having a major effect on optimum racing distance in the Thoroughbred horse in the USA. Anim Genet. 2010;41 Suppl 2:154–8.
45. Tozaki T, Hill EW, Hirota K, Kakoi H, Gawahara H, Miyake T, et al. A cohort study of racing performance in Japanese Thoroughbred racehorses using genome information on ECA18. Anim Genet. 2012;43:42–52.

46. Petersen JL, Valberg SJ, Mickelson JR, McCue ME. Haplotype diversity in the equine myostatin gene with focus on variants associated with race distance propensity and muscle fiber type proportions. Anim Genet. 2014;45:827–35.
47. Trifonov VA, Stanyon R, Nesterenko AI, Fu B, Perelman PL, O'Brien PCM, et al. Multidirectional cross-species painting illuminates the history of karyotypic evolution in Perissodactyla. Chromosome Res. 2008;16:89–107.
48. Piras FM, Nergadze SG, Poletto V, Cerutti F, Ryder OA, Leeb T, et al. Phylogeny of horse chromosome 5q in the genus Equus and centromere repositioning. Cytogenet Genome Res. 2009;126:165–72.
49. Nergadze SG, Lupotto M, Pellanda P, Santagostino M, Vitelli V, Giulotto E. Mitochondrial DNA insertions in the nuclear horse genome. Anim Genet. 2010;41 Suppl 2:176–85.
50. 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.
51. BLAT. http://genome.ucsc.edu/cgi-bin/hgBlat .
52. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12:656–64.
53. Jurka J. ERE1. http://www.girinst.org/protected/repbase_extract.php?access=ERE1 .
54. BLAST Trace database. https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&BLAST_SPEC=TraceArchive&PAGE_TYPE=BlastSearch&PROG_DEFAULTS=on .
55. Nergadze SG, Rocchi M, Azzalin CM, Mondello C, Giulotto E. Insertion of telomeric repeats at intrachromosomal break sites during primate evolution. Genome Res. 2004;14:1704–10.
56. Nergadze SG, Santagostino MA, Salzano A, Mondello C, Giulotto E. Contribution of telomerase RNA retrotranscription to DNA double-strand break repair during mammalian genome evolution. Genome Biol. 2007;8:R260.
57. Giordano J, Ge Y, Gelfand Y, Abrusán G, Benson G, Warburton PE. Evolutionary history of mammalian transposons determined by genome-wide defragmentation. PLoS Comput Biol. 2007;3, e137.
58. Steiner CC, Ryder OA. Molecular phylogeny and evolution of the Perissodactyla. Zool J Linn Soc. 2011;163:1289–303.
59. Trifonov VA, Musilova P, Kulemzina AI. Chromosome evolution in Perissodactyla. Cytogenet Genome Res. 2012;137:208–17.
60. Wakefield S, Knowles J, Zimmermann W, van Dierendonck M. Chapter 7: status and action plan for the Przewalski's horse (equus ferus przewalskii). In: Moehlman PD, editor. Equids: zebras, asses and horses: status survey and conservation action plan. Gland: IUCN; 2002. p. 82–92.
61. Table Browser. https://genome.ucsc.edu/cgi-bin/hgTables .
62. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004;32:D493–6.
63. Sakharkar MK, Chow VTK, Kangueane P. Distributions of exons and introns in the human genome. In Silico Biol. 2004;4:387–93.
64. Fedorova L, Fedorov A. Puzzles of the human genome: Why Do We need Our introns? Curr Genomics. 2005;6:589–95.
65. Gregory TR. Synergy between sequence and size in large-scale genomics. Nat Rev Genet. 2005;6:699–708.
66. Patrushev LI, Minkevich IG. The problem of the eukaryotic genome size. Biochem Mosc. 2008;73:1519–52.
67. Shepard S, McCreary M, Fedorov A. The peculiarities of large intron splicing in animals. PLoS ONE. 2009;4, e7853.
68. Krull M, Brosius J, Schmitz J. Alu-SINE exonization: en route to protein-coding function. Mol Biol Evol. 2005;22:1702–11.
69. Keren H, Lev-Maor G, Ast G. Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet. 2010;11:345–55.
70. Ponicsan SL, Kugel JF, Goodrich JA. Genomic gems: SINE RNAs regulate mRNA production. Curr Opin Genet Dev. 2010;20:149–55.
71. Apone S, Hauschka SD. Muscle gene E-box control elements. Evidence for quantitatively different transcriptional activities and the binding of distinct regulatory factors. J Biol Chem. 1995;270:21420–7.
72. Spiller MP, Kambadur R, Jeanplong F, Thomas M, Martyn JK, Bass JJ, et al. The myostatin gene is a downstream target gene of basic helix-loop-helix transcription factor MyoD. Mol Cell Biol. 2002;22:7066–82.
73. Jurka J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci U S A. 1997;94:1872–7.
74. Szak S, Pickeral O, Makalowski W, Boguski M, Landsman D, Boeke J. Molecular archeology of L1 insertions in the human genome. Genome Biol. 2002;3:research0052.
75. Vidale P, Magnani E, Nergadze SG, Santagostino M, Cristofari G, Smirnova A, et al. The catalytic and the RNA subunits of human telomerase are required to immortalize equid primary fibroblasts. Chromosoma. 2012;121:475–88.
76. Smit AF. ERE2. http://www.girinst.org/protected/repbase_extract.php?access=ERE2 .
77. Jurka J. ERE3. http://www.girinst.org/protected/repbase_extract.php?access=ERE3.
78. Wade CM. ERE4. http://www.girinst.org/protected/repbase_extract.php?access=ERE4.
79. Equus caballus (horse) Nucleotide BLAST. http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&PROG_DEF=blastn&BLAST_PROG_DEF=megaBlast&BLAST_SPEC=OGP__9796__11760 .
80. seq_contig.md. ftp://ftp.ncbi.nih.gov/genomes/Equus_caballus/mapview/seq_contig.md.gz.
81. Smit AFA, Hubley R, Green P. RepeatMasker. http://www.repeatmasker.org/.
82. UCSC Genome Browser ftp. ftp://hgdownload.cse.ucsc.edu/goldenPath/equCab2/chromosomes/.
83. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, et al. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 2014;42:D764–70.
84. MultAlin. http://multalin.toulouse.inra.fr/multalin/.
85. Corpet F. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res. 1988;16:10881–90.
86. 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.
87. Carbone L, Nergadze SG, Magnani E, Misceo D, Francesca Cardone M, Roberto R, et al. Evolutionary movement of centromeres in horse, donkey, and zebra. Genomics. 2006;87:777–82.
88. Anglana M, Bertoni L, Giulotto E. Cloning of a polymorphic sequence from the nontranscribed spacer of horse rDNA. Mamm Genome Off J Int Mamm Genome Soc. 1996;7:539–41.
89. Nergadze SG, Farnung BO, Wischnewski H, Khoriauli L, Vitelli V, Chawla R, et al. CpG-island promoters drive transcription of human telomeres. RNA. 2009;15:2186–94.
90. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods. 2001;25:402–8.
91. R Development Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2008. http://www.R-project.org.

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.

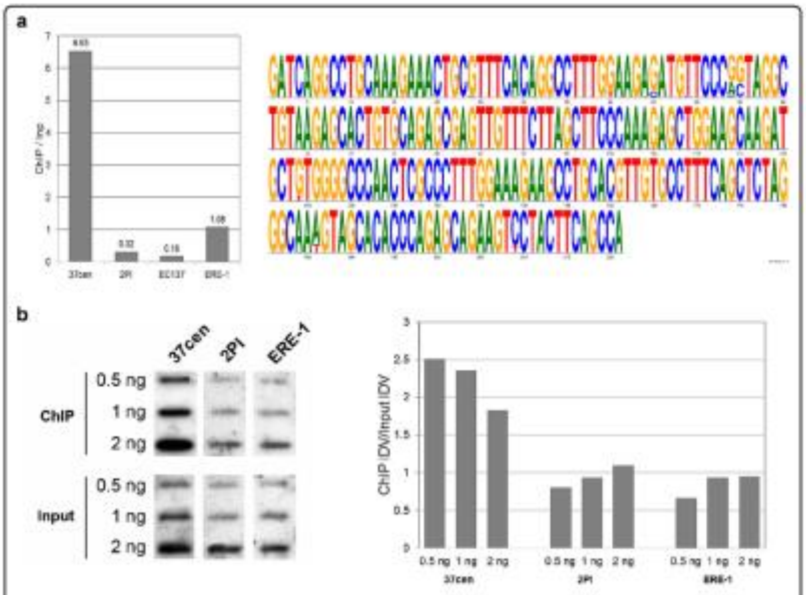

Fig. 1 Identification and analysis of the CENP-A bound horse satellite. **a**: In the left panel, the enrichment of the 37cen, 2P1 and EC37 satellites was measured as ratio between normalized read counts in immunoprecipitated and in input DNA. The same calculation was performed for the ERE1 retrotransposon 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

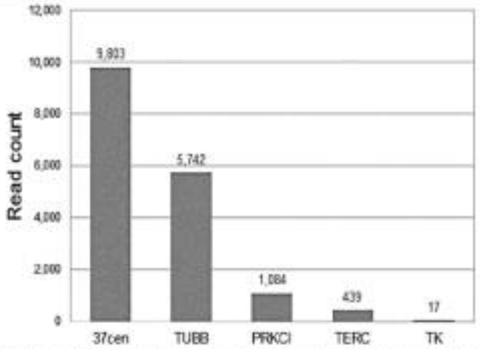

Fig. 2 Transcription of the 37cen satellite by RNA-seq. The graph reports the number of reads corresponding to the consensus sequences of 37cen, tubulin beta (TUBB), protein kinase C iota (PRKCI), telomerase RNA component (TERC), thymidine kinase (TK)

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 intraspecific 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

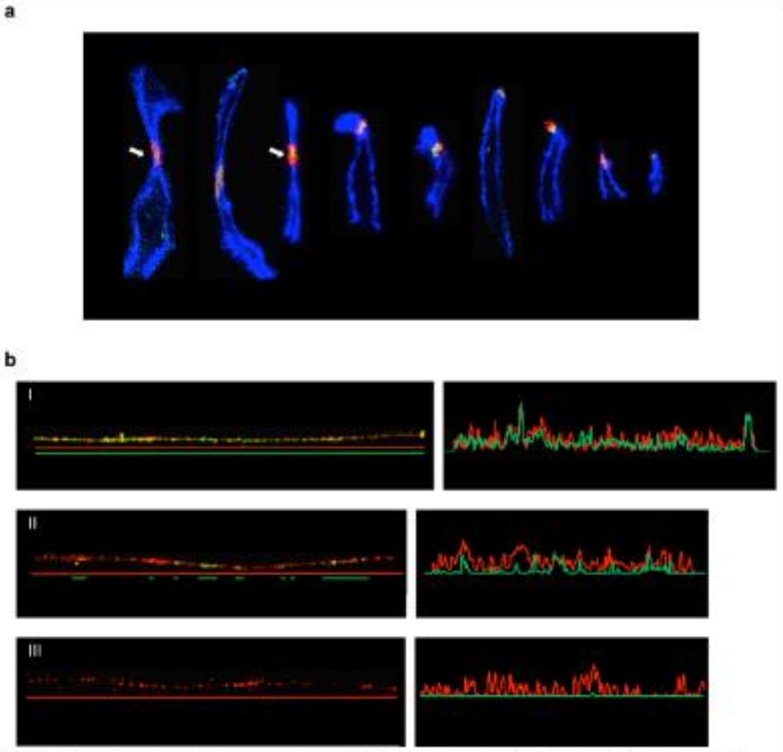

Fig. 3 a: Immuno-FISH on mechanically stretched chromosomes. 37cen is red labelled while CENP-A, detected by a CENP-A enriched CREST serum is green labelled. A total number of 99 stretched chromosomes was analysed. A sample of representative images is reported in the figure. **b**: Immuno-FISH on extended 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 IRCCS 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, Meštrović 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, Koblížková 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.