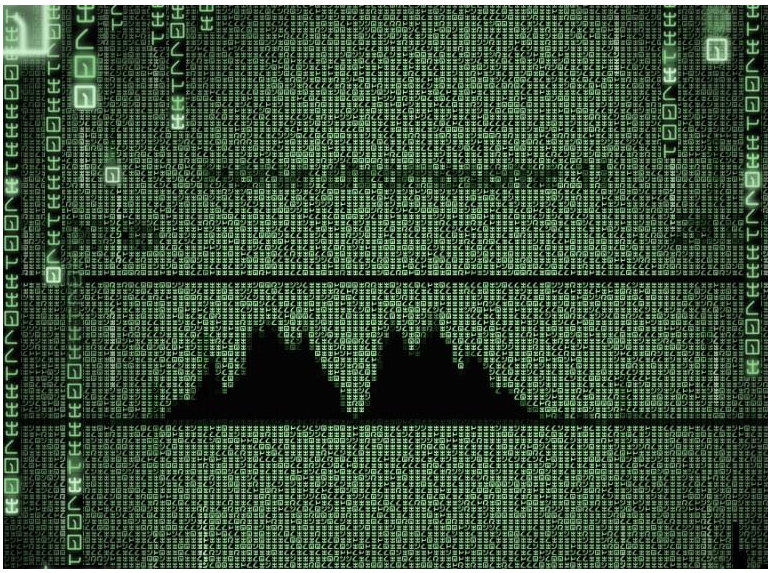
UNIVERSITA' DEGLI STUDI DI PAVIA

Dipartimento di Biologia e Biotecnologie "L. Spallanzani"

"Centromeric protein A in the genus *Equus* model system: functional and evolutionary aspects"



Marco Alfonso Rosario Corbo

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

UNIVERSITA' DEGLI STUDI DI PAVIA

Dipartimento di Biologia e Biotecnologie "L. Spallanzani"

"Centromeric protein A in the genus *Equus* model system: functional and evolutionary aspects"

Marco Alfonso Rosario Corbo

Supervised by Prof. Solomon Nergadze

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

ABSTRACT

The centromere is the chromosomal structure required for chromosome segregation during cell division. In different species, centromeric proteins show high levels of conservation while centromeric DNA sequences are highly variable. This discordance is related to the fact that the centromeric function is not defined by the DNA sequence but by epigenetic factors. The histone-H3 variant, CENP-A, is the main centromeric determinant.

The molecular characterization of mammalian centromeres has been a challenge due to the repetitive nature of its DNA sequence (satellite DNA). Previous studies from our laboratory demonstrated that the ECA 11 centromere of *Equus caballus* (Wade CM et al. 2009) is devoid of satellite sequences, having acquired its function in recent evolutionary times. Recently we discovered that, in *E. asinus*, 16 out of 31 centromeres are satellite-less (Nergadze SG et al. 2018). During my PhD thesis work, using an NGS approach (ChIP -seq with anti-CENP-A antibodies), we identified a total of 65 satellites-less neocentromeres in other species of the genus *Equus*: *E. zebra hartmannae, E. grevyi, E. burchelli, E. kiang* and *E. hemionus onager*.

We then carried out a comparative analysis of the localization and sequence of satellite-less centromeres in the different *Equus* species revealing that "centromerization" hot spots may have been used for the formation of neocentromeres during evolution.

We also mapped extra-centromeric CENP-A binding sites in horse, donkey, mouse and human cells. The results of this analysis strongly suggest that CENP-A may play a role in gene expression regulation.

The analysis of several epigenetic markers at the horse and donkey satellite-less centromeres was previously carried out in our laboratory (Riccardo Gamba PhD thesis 2017) using NGS approaches. Two additional markers (H4k20me1 and H3k4me3) were recently studied during my thesis work. Taken together, the results of these experiments allowed us to define a peculiar epigenetic landscape of these loci.

Finally, in the context of the FAANG (Functional Annotation of Animal Genomes) international collaboration, we evaluated the transcriptional status of the satellite-less centromere of horse chromosome 11 in different tissues of two horses. The preliminary results suggest that these regions are silent for polyA-RNA transcription and a few microRNAs are transcribed at a low level.

Taken together, this analysis provides new insights into the nature, evolution and structure of the mammalian centromeric domain at molecular level and on new possible epigenetic regulatory function of CENP-A, confirming the remarkable plasticity of the genus *Equus* genomes.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Professor Solomon Nergadze; he helped me during my whole journey, addressing me on the right way of thinking and making science. Thanks to his personality, he always helped me on taking very seriously my work, without forgetting that a smile or a joke is the key to face every day problems. Grazie Solomon.
I would like to thank Professor Elena Giulotto for the amazing chance of working in her laboratory and on her projects. She has been a mentor and a guide during my studies, helping me growing as scientist and as person. Working with her has been one of the most important experiences I ever had in science, but also in life. I won't be the person I am without her. Grazie Elena.
Of course, my stay in the laboratory wouldn't be the same without my colleagues. Francesca, besides being one of the most important members of the team, is really an amazing person. She carried out the majority of the experiments from which I started and began my project. She has been there every time I was in trouble.
Grazie Francesca.
Marco, Lela and Eleonora deserve also a mention since they have been of incredible help in the laboratory with my work and especially on the great memories we created together. Grazie Marco, Lela e Eleonora.
I also would like to thank all the past members of the laboratory, Francesco, Riccardo, Claudia, Antonio, Simone and all the present and past students. Grazie ragazzi.
My biggest thanks go to my parents and my brothers who helped me on each moment of my experience, during the bad and the good days; they have been patiently listening to me, giving me the best suggestions to solve any possible issue I had. Grazie Mamma, Papà, Gero e Toni.
A special mention goes to my cousin Erika, to my friends and PhD colleagues who contributed on making me the person I am now. Grazie Erika, Antonello, Giuseppe, Mario, Francesco, Daniele, Roberto, Marco, Elisa, Marco e tutti gli altri.
Finally, I would like to thank my will, my dreams and my passions on giving me the strength to go on and overcome the obstacles of my journey when I wanted to quit.

ABBREVIATIONS

BAC: Bacterial Artificial Chromosome
BS-seq: Bisulfite sequencing
CCAN: Constitutive Centromere Associated Network
CEN: Centromere
CENP: Centromeric Protein
ChIP: Chromatin Immuno-Precipitation
CR: Centromere Repositioning
CREST: Calcinosis, Raynaud's syndrome, Esophageal dismotility, Sclerodactyly and Telangiectasia
DMEM: Dulbecco's Modified Eagle Medium
EAS: Donkey (*Equus asinus*) chromosome
EBU: Burchell's zebra (*Equus burchelli*) chromosome
ECA: Horse (*Equus caballus*) chromosome
EGR: Grevy's zebra (*Equus grevyi*) chromosome
EHO: Onager (*Equus hemionus onager*) chromosome
EKI: Kiang (*Equus kiang*) chromosome
ENC: Evolutionary New Centromere
ERE-1: Equine repetitive element 1
EZH: Hartmann's zebra (*Equus zebra hartmannae*) chromosome
FISH: Fluorescence *In Situ* hybridization
H3K36me2: Histone H3 dimethylated at Lysine 36
H3K4me2: Histone H3 dimethylated at Lysine 4
H3K4me3: Histone H3 trimethylated at Lysine 4
H3K9me3: Histone H3 trimethylated at Lysine 9
H4K20me1: Histone H4 monomethylated at Lysine 20
HJURP: Holliday Junction Recognition Protein
LINE: Long INterspersed Element
NGS: Next Generation Sequencing
PCR: Polymerase Chain Reaction
qPCR: quantitative Polymerase Chain Reaction
RNA pol II: RNA polymerase II
RNA-seq: RNA sequencing
SNP: Single Nucleotide Polymorphism

TABLE OF CONTENTS

INTRODUCTION

1 The centromere

The centromere is one of the most important cell structures in all eukaryotes. It is necessary for the faithful chromosome segregation during both mitosis and meiosis.
It is the site for the kinetochore assembly, from which microtubules propagate during metaphase, allowing sister chromatids to correctly separate into the two daughter cells. Centromeres occur within a well characterized chromatin domain and surrounded by pericentromeric regions which are made of defined epigenetic markers.
Centromere malfunction can lead to genome instability, missegregation of sister chromatids and chromosome breakage. All these negative outcomes lead to aneuploidy and sometimes to cancer.

Centromeres could be grouped in three categories among eukaryotes [Figure I1]. They are divided in regional centromeres, point centromeres and holocentric centromeres:

i) Regional centromeres, typical of higher eukaryotes, span large regions (0.1-0.4 MB) and form a well-defined structure during metaphase, defined as primary constrictions if observed microscopically through cytogenetics. These centromeres are typically characterized by long stretches of repeated DNA sequences (satellite DNA) and retrotransposable elements. Satellite DNA is made by large arrays of tandemly reiterated DNA assembled with a head-to-tail structure [Figure I2]. Centromeres characterized by satellite DNA are more competent, as compared to satellite-less centromeres, in recruiting the centromeric proteins. Furthermore, gene deserts associated with satellite DNA regions may form an advantageous environment for centromere formation. Thus, one role of satellite DNA may be to stabilize the centromeric core.

ii) Point centromeres are typical of *S. cerevisiae,* span only few hundred nucleotides and the kinetochores bind only a single microtubule. DNA sequence specificity is a key factor in the formation and establishment of point centromeres in budding yeast unlike higher eukaryotes.

iii) Holocentric centromeres, typical of nematodes, insects and some plants, usually span the entire chromosome and the centromeric function is spread across the whole chromosome.

CENP-A is necessary for recruiting all the proteins required for the kinetochore assembly (Black BE and Cleveland DW 2011). CENP-A formation machinery is highly conserved during evolution (Maddox PS et al. 2012; Kato H et al. 2013). Proteins which form the centromere are highly conserved among different species

but the sequence underneath this structure greatly varies due to DNA recombination which leads to a rapidly evolving sequence (Bensasson D 2011).
Highly divergent sequences and highly conserved centromeric proteins are commonly described as the "centromere paradox" (Henikoff S et al. 2001).
This paradox is solved by the fact that now is known that the centromere is epigenetically defined (Black BE and Cleveland DW 2011), since CENP-A is the main centromeric determinant.
The main difficulty for the centromere study is the presence of highly repeated DNA elements called satellite sequences (Plohl M et al. 2014). A detailed dissection of the epigenetic factors associated to this locus has been hindered by the typically repetitive nature of centromeric DNA (Amor DJ and Choo KH 2002).

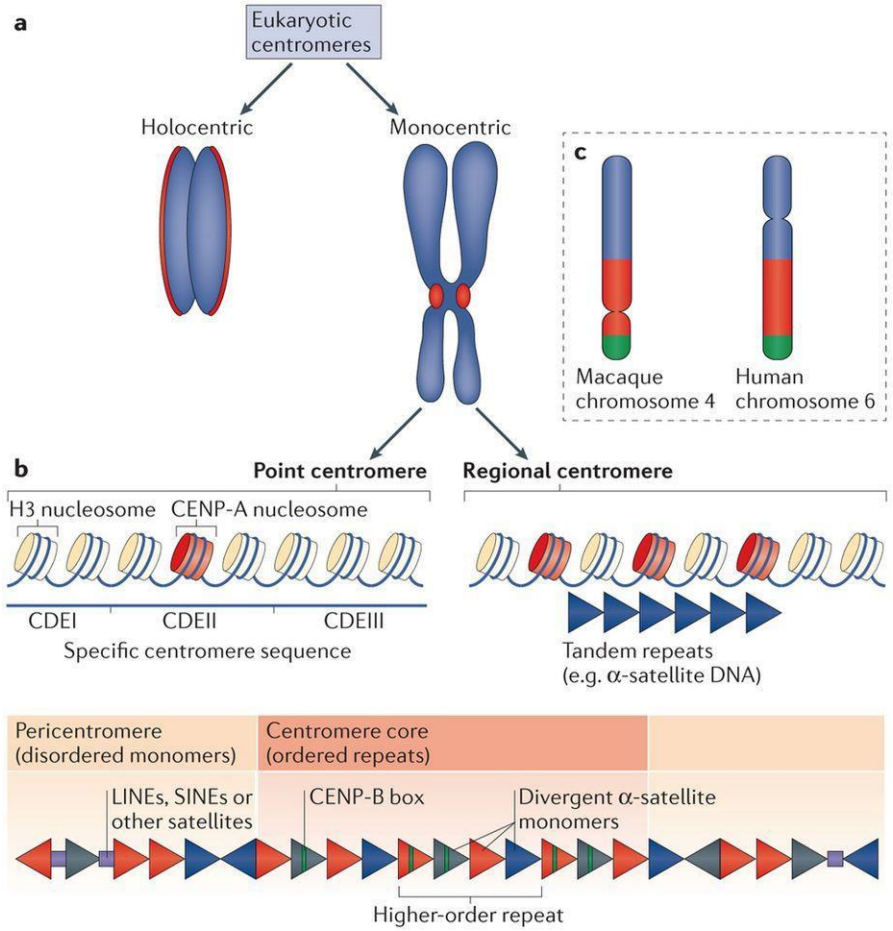


Figure I1: Centromere specification. *A) different types of centromeres; B) α-satellite monomers are the DNA component of primate centromeres; C) macaque and human orthologous chromosomes (McKinley KL and Cheeseman IM 2016).*

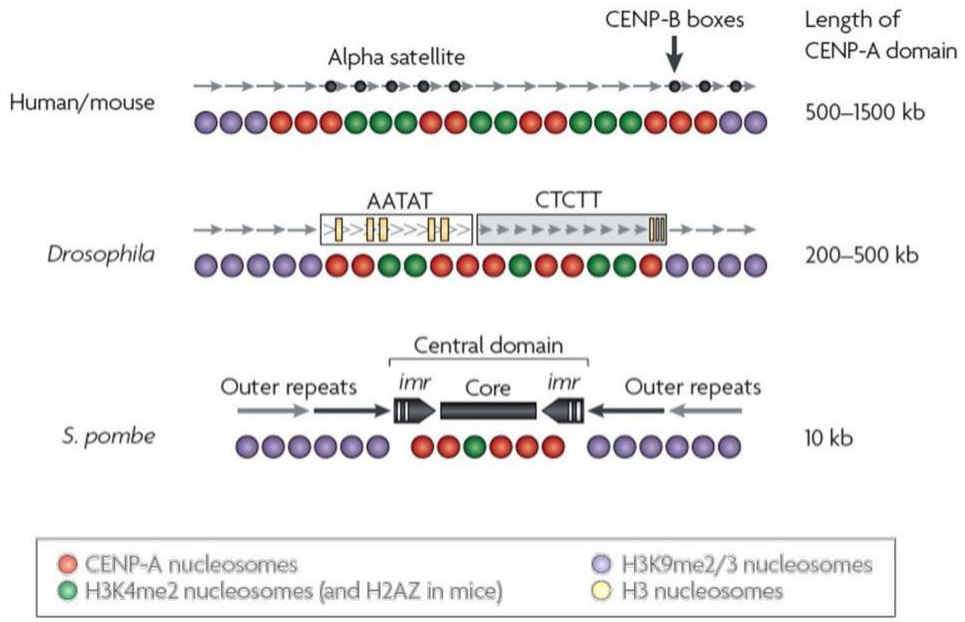


Figure I2 Centromere structure and organization. *Schematic representation of centromeric DNA and nucleosomes in humans and mice, fruit flies and fission yeast (readapted from Allshire RC and Karpen GH 2008).*

2 The kinetochore

Despite the different localization and sequence composition of centromeres, its unique function is shared across species, and it is the genome locus of the kinetochore machinery assembly.

The kinetochore is the protein structure needed for the correct segregation of sister chromatids during cell division.

It is made by two regions, the inner kinetochore which is tightly associated with the centromeric DNA, and the outer kinetochore which interacts with the microtubules [Figures I3 and I4]. Kinetochores lead chromosomal movements during cell division. Cohesion of sister chromatids is mainly compelled by cohesins, and each of the sister chromatids has its own kinetochore which is linked to the opposite pole of the mitotic spindle through microtubules. During anaphase, the opposite movement of the sister chromatids towards their facing mitotic spindle, ensures the correct division of the chromosomes between daughter cells.

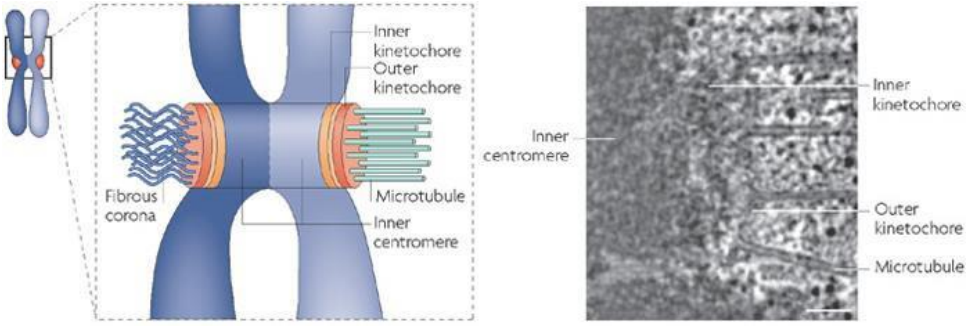


Figure I3: Vertebrate kinetochore ultrastructure. *A) representation of paired sister chromatids the chromatid attached to microtubules (right) and unattached (left). Kinetochore layers and inner centromere are shown. B) Electron micrograph of a human kinetochore (readapted from McEwen BF et al. 2007).*

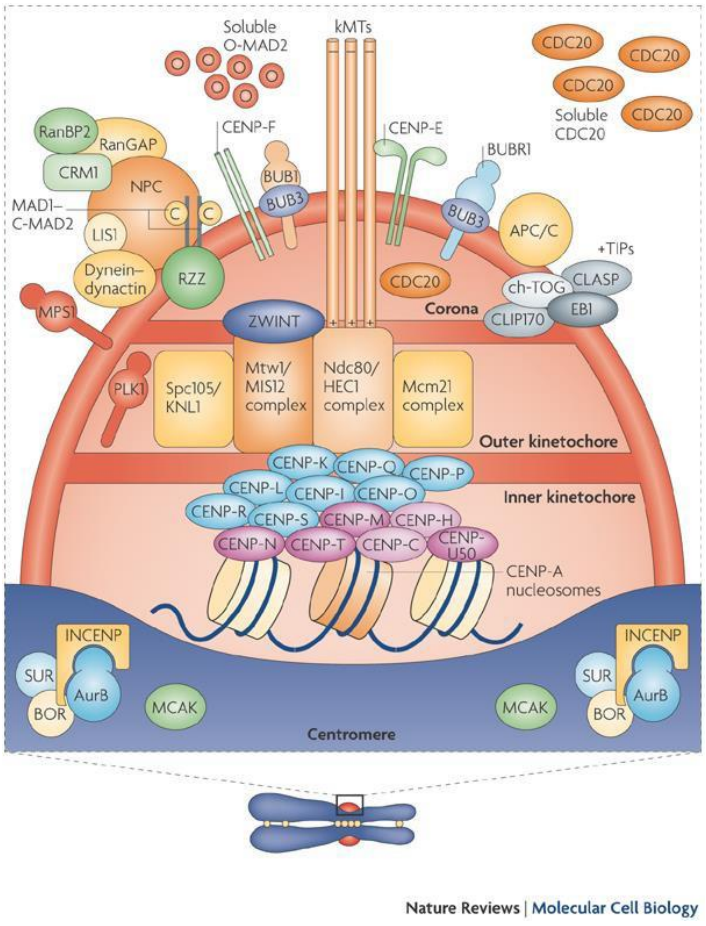


Figure I4: The centromere–kinetochore region. *Schematic representation of the proteins assembling into the centromeric region to form the kinetochore (Musacchio A and Salmon ED 2007).*

3.1 CENP-A protein

CENP-A is the one of the main component of the centromere structure. It is a histone H3-variant which replaces H3 in the nucleosomes at the centromere core.
It is a constitutive component of the centromere and one of the leader protein in the segregation pathway. CENP-A over-expression in mutant cell lines causes its misincorporation in several other genomic regions, although not forming active centromeric sites (Van Hooser AA et al. 2001). Previous studies reported that, CENP-A is localized on active centromeres and not in inactive sites (Earnshaw WC and Migeon BR 1985; Choo KH 1997), so its localization and position is not sequence-related (Vafa O and Sullivan KF 1997; Warburton PE et al. 1997).
CENP-A participates in the formation of an octameric nucleosomal structure [Figure I5], formed by one CENP-A/H4 heterotetramer and two H2A/H2B heterodimers, which bind to 120-150 bp DNA sequence (Allshire RC and Karpen GH 2008). Studies showed that CENP-A nucleosomes have more plasticity than nucleosomes containing the normal form of H3 protein meaning that this chromatin region is more prone to unwrapping and conformational changes in its 3D structure (Verdaasdonk JS and Bloom K 2011).
Crucial, for understanding the propagation of centromere identity, is the study of CENP-A loading onto chromatin [Figures I6 and I7]. CENP-A is positioned at the centromeric locus by two mechanisms: a *de novo* pathway occurring in artificial chromosomes and a maintenance pathway occurring in new replicated cells where half of H3 protein is preloaded on the new synthetized DNA and then replaced by CENP-A (Buscaino A et al. 2010).
The protein insertion and the nucleosome formation can have different kind of deposition models such as the looping model, the solenoid model, and the sinusoidal patch model (Verdaasdonk JS and Bloom K 2011).
CENP-A-containing nucleosomes are usually displayed in the centromeric outer surface, thus having the spindle pole on the same side, facilitating the microtubule attachment to the kinetochore. The looping model proposes that CENP-A chromatin is looped out from the bulk chromatin towards the spindle pole. In the solenoid model, CENP-A and H3 nucleosomes occupy different sides of the arrangement of CEN chromatin, in which CENP-A is sorted towards the kinetochore and H3 faces the inner centromere. According to the sinusoidal patch model, alternating CENP-A and H3 domains fold in a sinusoidal fashion into various layers, stacked on the top of each other. Further studies will be required to demonstrate which model for the physical organization of eukaryotic centromeric chromatin is correct.
CENP-A is just one the initial protein required for the kinetochore assembly. It recruits other centromeric proteins (CENP proteins, which will form the necessary scaffold for complete kinetochore formation) in that specific locus.

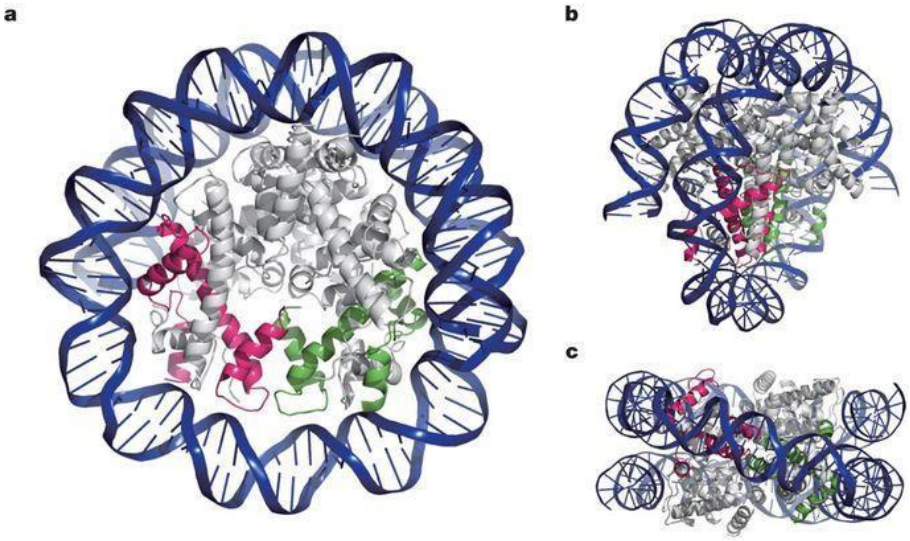


Figure I5: Crystal structure of the human CENP-A nucleosome. Three views of CENP-A nucleosome. *a) frontal view of the DNA supercoil. b, c) side views of the DNA supercoil. CENP-A molecules are shown in magenta and green. The central 121 bp DNA region is shown in dark blue (Tachiwana H et al. 2011).*

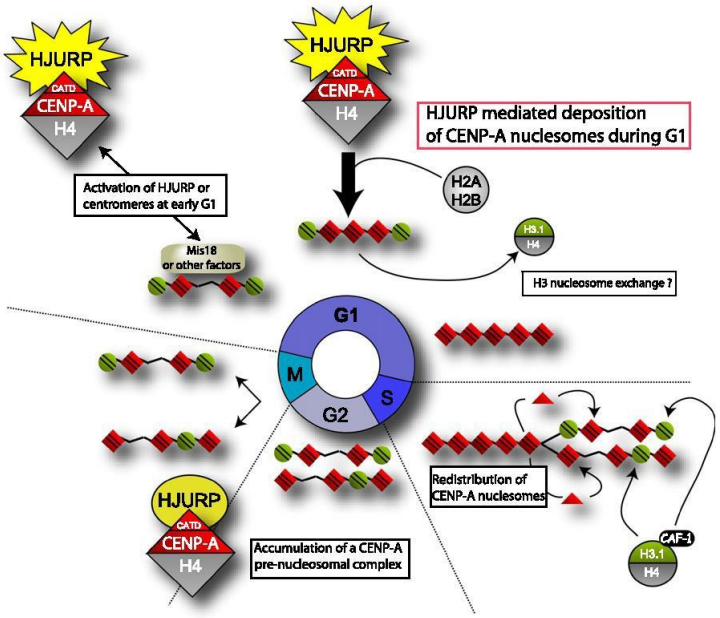


Figure I6 Model of CENP- A nucleosome deposition at centromeres. *CENP-A nucleosomes are distributed to sister chromosomes during DNA replication. Cells have half of the total number of CENP-A nucleosomes till mitosis. CENP-A nucleosomes are deposited during G1 phase through HJURP mediation (Foltz DR et al. 2009).*

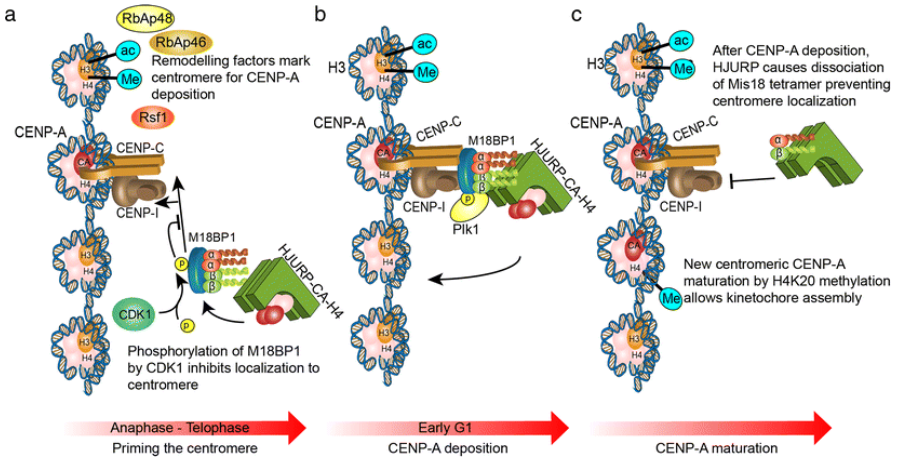


Figure I7 Maintenance of CENP-A chromatin. *a) The priming step involves the action of different regulatory proteins such as RbAp46/48, RSF1, FACT, and RNAPII which create a feasible chromatin environment for new CENP-A recruitment and deposition through different histone modifications such as H3K4 methylation or H3K9 acetylation. b) The phosphorylated Mis18 complex (by PLK1) allows the HJURP complex to load new CENP-A. c) Mis18 complex is destroyed, inhibiting further CENP-A deposition. CENP-A nucleosomes are methylated at the lysine 40 of the H4 protein (Nagpal H and Fukagawa T 2016).*

3.2 CENP-A in non centromeric regions

CENP-A is present not only in the centromeric core, but also at genome-wide level (Bodor DL et al. 2014). Centromeric loci comprising CENP-A molecules are determined by the density of the histone H3 variant distributed on the DNA sequence unit (Bodor DL et al. 2014). Centromeric domains are highly enriched in CENP-A, although normal H3 histone variants are still present on the centromeric domain (Blower MD et al. 2002; Sullivan BA and Karpen GH 2004; Sullivan LL et al. 2011). CENP-A nucleosomes tend to localize on transcription factor sites in human cancer genome, being possibly involved in gene expression and regulation (Athwal RK et al. 2015) and they tend to nucleate around heterochromatin spots (Gonzalez M et al. 2014)

CENP-A containing nucleosomes localized throughout the genome were demonstrated to be relevant for the epigenetic state of chromatin, influencing gene transcription and regulation

4 CENP-B

CENP-B is another important protein which localizes within centrochromatin (Cooke CA et al. 1990; Pluta AF et al. 1992) despite the presence or absence of CENP-A. Composed of 599 amino acids, it functions as a dimer (Earnshaw WC and Rothfield N 1985; Sullivan KF and Glass CA 1991) due to its C-terminal domain. Its two domains share high sequence homology among species (Sullivan

KF and Glass CA 1991; Pluta AF et al. 1992; Muro Y et al. 1992; Yoda K et al. 1996). CENP-B is the only centromeric protein having a precise DNA binding specificity and recognizing a 17 bp long sequence (Fujita R et al. 2015). Only 9 out of the 17 bp of the CENP-B box are essential for binding specificity (Masumoto H et al. 1993). CENP-B boxes with those nine essential nucleotides were found within the satellite DNA of centromeres of different species (Kipling D et al. 1995; Yoda K et al. 1996). In human centromeres CENP-B seems to stabilize CENP-A and CENP-C at this locus, increasing the centromere strength and fidelity of chromosome segregation (Fachinetti D et al. 2015). However, centromeres lacking satellite DNA like active human neocentromeres lack the CENP-B box (Choo KH 2000; Amor DJ and Choo KH 2002). Furthermore, it seems not to be essential also in CENP-B knock-out mice which live normally (Hudson DF et al. 1998).

5 CENP-C

CENP-C is necessary for the establishment of functional centromeres and is only detectable within active centromeres (Sullivan BA and Schwartz S 1995; Fukagawa and Brown WR 1997); its absence causes mitotic delay, chromosome missegragation, apoptosis (Fukagawa and Brown WR 1997). The atomic mass of CENP-C is 107 KDa and, like CENP-A and CENP-B, it is highly conserved during evolution (Henikoff S et al. 2001). CENP-C is expressed through the whole cell cycle (Knehr M et al. 1996). Its level increases during different cell cycles reaching its maximum in G1 before being partially degraded. CENP-C is a fundamental protein of the Constitutive Centromere Associated Network (CCAN). It is a molecular bridge between CENP-A nucleosomes and the NDC80 complex. CENP-C also appears to bind DNA directly. However, it lacks sequence specificity (Sugimoto K et al. 1994; Yang CH et al. 1996).

6 Histone modification and centromere transcription

CENP-A is not the only key factor for centromere identification, function and maintenance. Additional molecular markers contribute to the establishment of the centromere. These markers are known to be associated with centromeric DNA transcription, centromeric and pericentromeric chromatin and histone modifications (Summarized in Figure I8 and discussed below).

Early studies identified centrochromatin as being associated with heterochromatin (Lima-De-Faria 1949), and later on studies have found out that pericentromeric regions are particularly rich in histone H3 trimethylated at Lysine 9 (H3K9me3), which is a marker of constitutive heterochromatin (Peters AH et al. 2003; Rice JC et al. 2003). Some studies in *S. pombe* and Drosophila (Partridge JF et al. 2000; Heun P et al. 2006;) suggested that, one of the possible roles of the heterochromatic environment at pericentromeric loci seems to be creation of a physical boundary to limit the propagation of CENP-A domains.

The existence of centromere transcription was first reported in mouse satellite DNA (Cohen AK et al 1973)

At the centromeric core, CENP-A nucleosomes are interspersed with nucleosomes containing histone H3 modified with transcriptionally permissive markers, such as dimethylated Lysine 4 (H3K4me2) and dimethylated Lysine 36 (H3K36me2) in human and *D. melanogaster*. This strengthens the fact that satellite DNA at centromeres is transcribed. H3K4me2, found at centromeres, is characteristic of "poised" chromatin, "ready to be transcribed". On the contrary, its trimethylated form (H3K4me3), associated with actively transcribed DNA, was not found at the centromeric core other transcription-associated modifications, such as acetylation of histone 3 and 4 have not been detected at centromeres (Bailey AO et al. 2015).

Interestingly, in chicken centromeres which do not contain satellite DNA, the centromeric core lacks H3K4me2 and H3K36me2, suggesting that their presence is dispensable for centromeric function (Hori T et al. 2014).

In CENP-A containing nucleosomes, CENP-A is known to associate to the H4K20me1 marker (histone H4 monomethylated at Lysine 20), which is essential for kinetochore assembly (Bailey AO et al. 2015).

The different histone modifications that have been identified at the centromeric and pericentromeric regions are believed to assemble into a complex 3D structure, with specific histone marks making contact with the kinetochore machinery (Stellfox ME et al. 2013). Despite the accumulating evidence on several model systems, a comprehensive understanding of this aspect is far from clear.

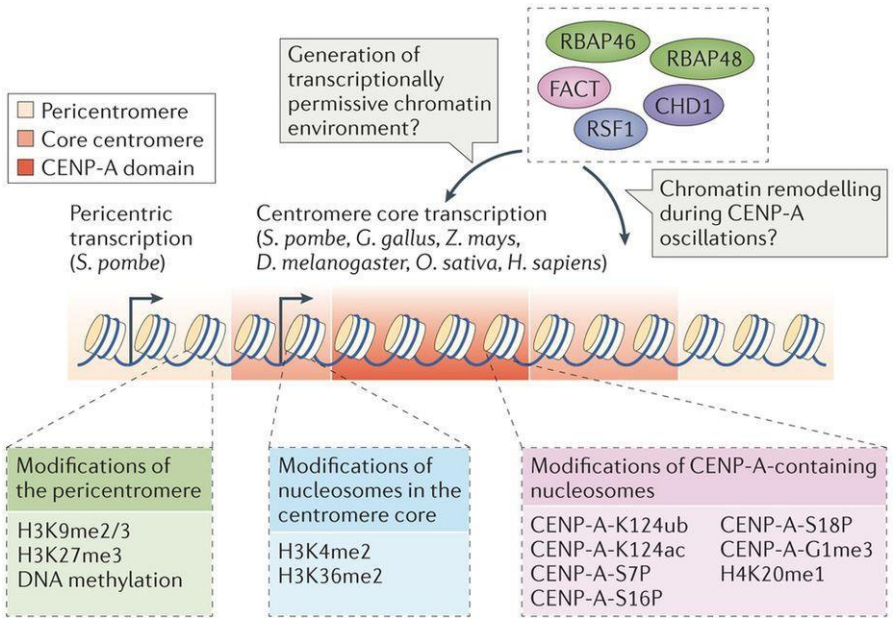


Figure I8: Schematic representation of the histone modifications associated to the centromeric region (McKinley KL and Cheeseman IM 2016).

7 Satellite DNA

Mammalian centromeres are typically characterized by the presence of highly repetitive sequences, called satellite DNA, that make up the centromeric and pericentromeric heterochromatin (Gartenberg M 2009; Torras-Llort M et al. 2009). These sequences undergo rapid evolution, as the repeat monomers are extremely variable in sequence composition and length even in closely related species. However, they seem to have similar modes of evolution, maybe due to a possible role in the stabilization of the centromeric function (Melters DP et al. 2013). Often, sequence features in satellite repeats are different between the centromeric core region and the surrounding pericentromere.

Concerning the homogeneity of the satellite repeat across the karyotype, different species may have different patterns. For example, in humans, a defined number of monomers are organized into chromosome specific satellite DNA families (Schueler MG and Sullivan BA 2006). In other cases, satellite arrays may be nearly identical at centromeres of all the chromosomes (Macas J et al. 2010).

In primates, the centromeric repeat has been termed α-satellite DNA and it was identified in all primates studied so far (Alexandrov I et al. 2001; Willard HF 1991). Originally, the α-satellite array was characterized as divergent 170 bp monomers organized in a tandem head-to-tail fashion (Maio JJ 1971) [Figure I9]. This type of satellite is termed monomeric and has been identified in the centromeric and pericentromeric region of 21 human chromosomes (Alexandrov IA et al. 1993; Rudd MK and Willard HF 2004). In the centromere, α- satellite monomers arrange into homogeneous higher-order repeats (HORs), each spanning 3-5 megabases. Functional centromeres form on a portion of the HOR region (Fukagawa T and Earnshaw WC 2014). The majority of data describing human α-satellite arrays implicates unequal crossover between sister chromatids as the primary force driving change and evolution of these sequences (Warburton PE et al 1996). Segmental duplication is also an important factor causing amplification of satellite DNA arrays (Horvath JE et al., 2005; Ma J and Jackson SA 2006).

Since the biological role of satellite DNA remains elusive, several groups have tried to infer its function by identifying common features of the centromeric repeat in different species. As CENP-A is essential for kinetochore nucleation, it has been hypothesized that centromeric repeat monomers may tend to be about the size of DNA embedded in one nucleosome (Willard HF 1991; Shelby RD et al. 1997) which is another feature proposed as being common among satellite sequences of different species was low GC content (Henikoff S et al. 2001). However, these hypotheses were dismissed by an extended analysis of tandem repeats from hundreds of species (Melters DP et al. 2013), which showed an extreme variability both in length and in sequence composition of satellite repeats. In some species, monomers much longer than the size of DNA embedded in a nucleosome or with high GC content were identified.

Therefore, the only common feature characterizing satellite DNA appears to be their repetitive nature.

Several functions have been assigned to centromeric satellites specific roles. For example, they may act as “binding sequence donors”. In a number of species,

satellite repeats contain the CENP-B box motif, which allows binding of the protein CENP-B directly to DNA, thus improving chromosomal stability (Kipling D et al. 1995; Yoda K et al. 1996). 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 (Plohl M et al. 2008, 2014). Pericentromeric repetitive DNA might inhibit spreading of the centromere over neighboring genic regions (Sullivan BA 2002). It has also been proposed that the satellite DNA may improve the cohesion and the separation of sister chromatids.

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 T and Earnshaw WC 2014).

Whatever their function, it is important to remember that centromeric DNA sequences are dispensable, and that satellite-free functional chromosomes exist (Plohl M et al. 2012) (see section 8 Neocentromeres).

Centromeric satellite DNA is transcribed, and this transcription appears to be very important for centromere maintenance (Steiner FA and Henikoff S 2015). Indeed, transcription of repeat elements is part of the regulatory mechanisms of centromeres, and defects in transcriptional competence lead to chromosome missegragation (Hsieh CL et al. 2011; Ohkuni K and Kitagawa K 2011; Chan FL et al. 2012). Conversely, hypermorphic expression of centromeric RNAs impairs CENP-A loading (Carone DM et al. 2013). Transcripts of different length that are homologous to centromeric and pericentromeric repetitive sequences have been identified in several organisms such as yeast (Ohkuni K and Kitagawa K 2011; Choi ES et al. 2012), mouse (Ferri F et al. 2009), wallaby (Carone DM et al. 2009) and humans (Saffery R et al. 2003; Wong LH et al. 2007).

Although satellite transcription was proposed to promote the formation of a heterochromatic environment (Verdel A et al. 2004; Maida Y et al. 2014) or the deposition of CENP-A nucleosomes (Quénet AR and Dalal IM 2014) the precise function of these transcripts remains unclear.

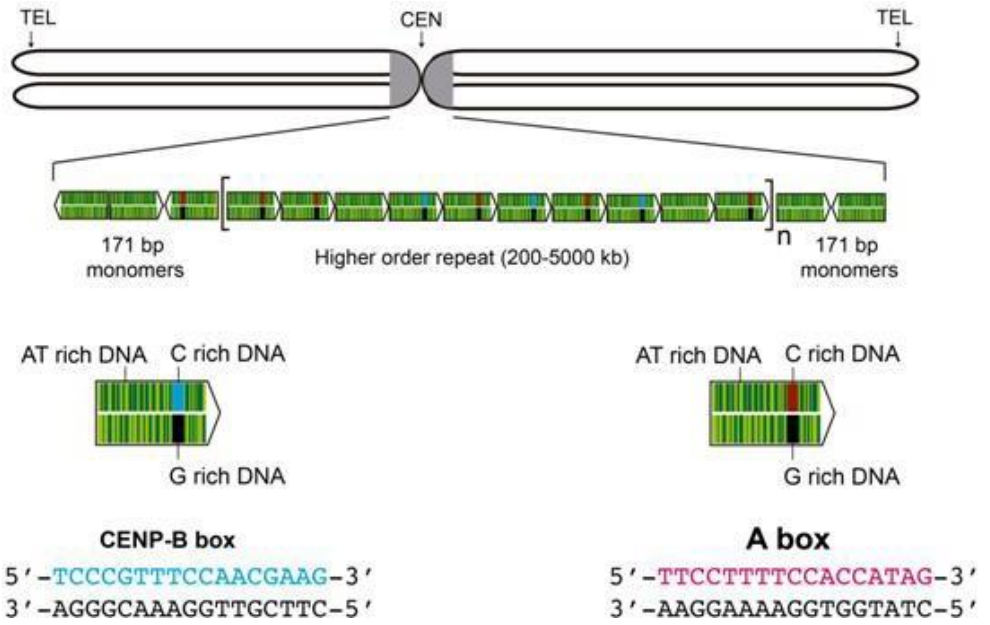


Figure I9. Schematic representation of the centromeric alpha-satellite DNA showing the position of the A and B boxes at the centromere (readapted from Garavís M et al. 2015).

8 Neocentromeres

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

The term "centromerization" was coined by Choo to define the process of centromere formation in a chromosomal region. Centromerization normally concerns the propagation of an existing centromere during replication. Rarely, this phenomenon occurs in regions which are normally non-centromeric. The ectopic centromere that appears occasionally in otherwise non-centromeric chromosomal regions is called "neocentromere" (Amor DJ and Choo KH 2002; Choo KH 2000; Kalitsis J and Choo KH, 2012). Two different types of neocentromeres have been identified: clinical neocentromeres and evolutionary new centromeres. While clinical neocentromeres are sporadic cases that are not fixed in the population, evolutionary new centromeres are fixed in the species and represent an aspect of karyotype evolution.

Such neocentromeres are different from the "classical" plant neocentromeres first described by Rhoades and Vilkomerson (Rhoades MM and Vilkomerson H 1942). Plant neocentromeres are accessory centromeres coexisting with the functional

normal centromere, their activity is confined to meiosis and they do not form a typical kinetochore (Rhoades MM and Vilkomerson H 1942; Amor DJ and Choo KH 2002; Dawe RK and Hiatt EN 2004).

8.1 Clinical neocentromeres

Since the discovery of the first neocentromere (Voullaire LE et al. 1993), more than 90 cases of human neocentromeres have been described (Kalitsis J and Choo KH, 2012; Marshall OJ et al. 2008). Generally, neocentromerization is a rare rescue mechanism to avoid the loss of an acentric chromosomal fragment originating from a chromosomal rearrangement. Beyond neocentromere formation, these chromosomal rearrangements result in karyotype instability and are usually detrimental to the individual, explaining why human neocentromeres are occasional and not fixed in the population (Amor DJ and Choo KH 2002; Marshall OJ et al. 2008). However, as previously mentioned, at least in one case, a clinical neocentromere was transmitted through three generations (Tyler-Smith C et al. 1999), showing the possible inheritance of these ectopic centromeric domains.

Human clinical neocentromeres are functional centromeres which are completely devoid of satellite DNA (Amor DJ and Choo KH 2002; Marshall OJ et al. 2008; Kalitsis J and Choo KH 2012). They typically arise in gene-poor euchromatic regions, although heterochromatic markers have been detected, suggesting that neocentromeres carry certain features of heterochromatin (Amor DJ and Choo KH 2002; Kalitsis J and Choo KH, 2012). Despite the absence of sequence preference for neocentromere seeding, centromerization has not been reported at random sites along chromosomes. It has been hypothesized that genomic "hotspots" for centromerization exist in certain region of the genome. These genomic locations may favour neocentromerization because of specific epigenetic hallmarks or the persistence of recombinogenic duplicons. It has been proposed that regions of the genome with a high content of duplications are predisposed to rearrangements, which then lead to neocentromere formation through epigenetic changes in the chromatin after DNA repair (Marshall OJ et al. 2008).

Despite their full functionality as centromeric domains, some differences have been identified when comparing neocentromeres to satellite containing centromeres. Irvine and colleagues (Irvine DV et al., 2004) demonstrated a decreased level of overall CENP-A binding at neocentromeres compared to typical centromeres. More recently, a neocentromere was shown to bind ~25% less CENP-A compared to the satellite containing centromeres within the same cell line (Bodor DL et al. 2014). According to several reports, the average size of the neocentromere is between 40 and 500 kb (Lo AW et al. 2001; Alonso A et al. 2010; Hasson D et al. 2013; Shang WH et al. 2013), making them smaller than typical satellite-containing centromeres (04.-4.2 Mb according to (Sullivan LL et al. 2011).

As with satellite containing centromeres, neocentromeres can be transcribed. A human neocentromere was found to **be laying on a DNA region which is rich in transcriptionally active LINE retrotransposons** (Chueh AC et al. 2009).

8.2 Satellite-free centromeres

Neocentromeres are structures formed on unique regions containing no repetitive elements (differently from the active old centromeres) that can form spontaneously after sequence recombination or centromere damage.
These new structures usually form following these sequence rearrangement events but they rarely can acquire autonomously centromeric function (Amor DJ et al. 2004; Liehr T et al. 2010). They then can be transmitted to daughter cells maintaining the active function (Kalitsis P and Choo KH 2012).
These structures seem to be formed on region with no relevant characteristics, especially in gene poor regions. Neocentromere formation does not randomly happen. For example some human neocentromeres form at sites that, in the past, had an active centromere function (Ventura M et al. 2003; Ventura M et al. 2004). Sometimes they form in regions that seem to have a high tendency for recombination. Chromosomal ends seem to be preferential genomic region for the neocentromere formation.
This can be driven by some epigenetic marks helping the formation of an active centromere structure. Studies (Castillo AG et al. 2013) demonstrated that an over expression of the centromeric protein preferentially aggregates at the chromosome arms.

9 The *Equus* model and discovery of the first natural satellite-less centromere

The karyotypes of the extant *Equus* species are characterized by the presence of a variable number of meta- and submetacentric chromosomes derived from fusions between ancestral acrocentric elements (Trifonov VA et al. 2008). Indeed, equids are considered a representative example of recently diverged organisms. The eight living species of the genus *Equus* comprise two horses (*E. caballus* and *E. przewalskii)*, two Asiatic asses (*E. kiang and E. hemionus*), one African ass (*E. asinus*) and three zebras (*E. grevyi*, *E. burchelli and E. zebra*) (Steiner CC et al. 2012). All these species recently divergence and their number of chromosomes greatly vary, from 32 in *E. zebra* to 66 in E. *przewalskii*. Our laboratory has focused on the cytogenetics and biology of the genus *Equus*, to investigate the biology of the centromere. We investigated the position of the centromere, with respect to flanking markers, in the horse, in the donkey, and in the Burchell's zebra. The results of these early studies showed that at least eight centromere repositioning events occurred in the genus *Equus*. Surprisingly, at least five of these events arose in the donkey after its divergence from the zebra, which took place approximately 1 million years ago (Carbone L et al. 2006); subsequently the evolutionary history of horse chromosome 5q in seven species belonging to the genus Equus was investigated (Piras MF et al. 2009); two further centromere repositioning events were detected involving donkey chromosome 16 and Burchell's zebra chromosome 17, respectively.

Previous work from our laboratory demonstrated that centromere repositioning played an important role in the rapid karyotype evolution of the species belonging to the genus *Equus* (horses, asses and zebras) (Carbone L et al. 2006; Piras MF et al. 2009, 2010). The chromosomal distribution of satellites was investigated by one-color FISH in *Equus caballus*, *Equus asinus*, *Equus grevyi and Equus burchelli* (Figure I10, Piras MF et al. 2010). Metaphases were hybridized with genomic DNA. Due to the different hybridization kinetics between single copy and highly reiterated sequences, this procedure identified regions containing very abundant tandem repeats, such as centromeric satellite repeats. All centromeres of the horse, except for the one of chromosome 11, were labeled. In the other species, several chromosomes lacked visible satellite DNA at centromeres, which is instead present at non-centromeric sites.
The centromere of ECA11 was the only one in the horse lacking any hybridization signal in FISH experiments (absence of blue FISH signal in the horse chromosome 11 as shown Figure I10) in which the two major horse satellites or total horse genomic DNA were used as probes.

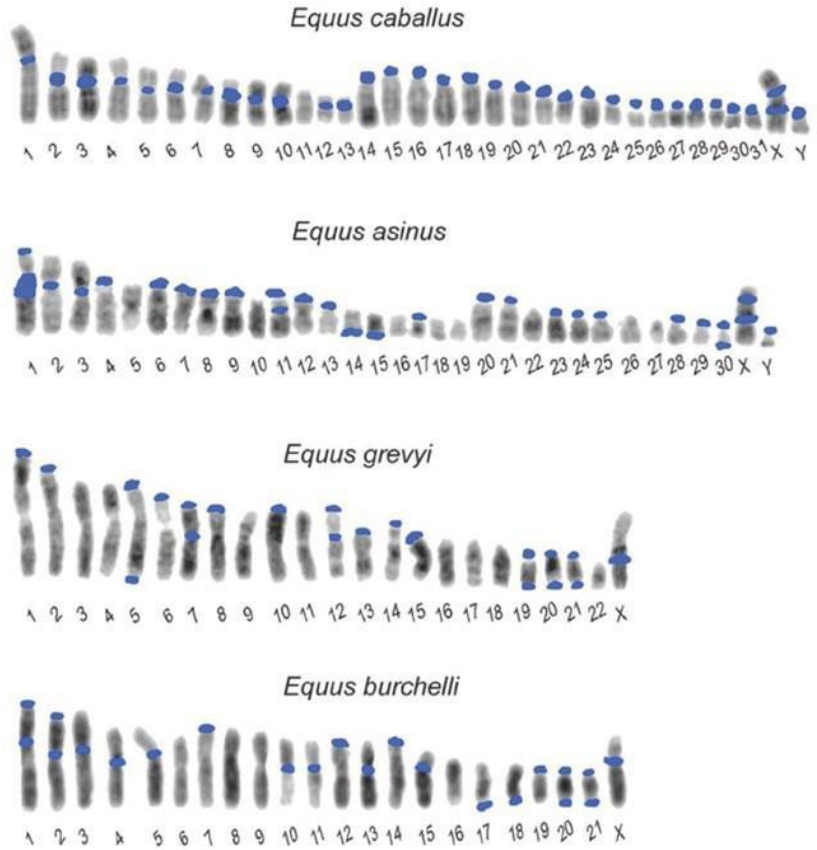


Figure I10 FISH distribution of total satellite repeats (blue spots) on horse, donkey, Grevy's and Burchell's zebras chromosomes (Piras MF et al. 2010).

As mentioned above, one centromere in the horse, and 16 centromeres in the donkey, 17 in Grevy's zebra and 9 in Burchell's zebra were devoid of satellite DNA. These data strongly suggested that, in equid species, centromere function is uncoupled from satellite DNA, and that the functional centromere coincides with the primary constriction in satellite-containing as well as in satellite-free centromeres (Wade CM et al. 2009; Piras MF et al. 2010). Thanks to the presence of such high number of satellite-free centromeres in the equids, this genus is a unique model for the study of centromere function, organization and evolution.

To test, at sequence level, whether satellite DNA was completely missing at this centromere, the primary constriction of ECA11 was localized by performing two and three color hybridization experiments on horse metaphase spreads with a panel of BAC clones. Taking advantage of the horse genome sequence assembly, a 2.7 Mb region predicted to contain the centromeric function was identified and an array covering this region was analyzed by ChIP-on-chip. The array was hybridized with DNA purified from chromatin immuno-precipitated with antibodies against CENP-A or CENP-C. With both antibodies, two peaks spanning about 135 and 100 kb, respectively, separated by a

165 kb region were identified whereas no hybridization was observed in the flanking sequences. The 400 kb region comprising the two peaks did not show protein coding genes, normal levels of interspersed repetitive elements, and no evidence of accumulation of L1 transposons, which were previously hypothesized to influence ENC formation (Chueh AC et al. 2009). The absence of extended tandem repeat arrays demonstrated that our initial hypothesis was correct and that a normal, stable and functional mammalian centromere can be totally deprived of satellite DNA. Using a similar approach, a satellite-free ENC was then identified in orangutan, where it is present in a heterozygous state (Locke DP et al. 2011). Interestingly, centromere of horse chromosome 11 is contained in a large conserved synthetic region within many mammalian species. The fact that this region is conserved within many species but it is centromeric only in the horse, supports the idea the centromeric function is not related to DNA sequence. It was proposed that the ECA11 centromere is evolutionarily young and, although functional and stable in all horses, did not yet acquire all the marks typical of mammalian centromeres. As a result of this work it was possible for first time to identify a mammalian ENC in an immature state, as suggested by our previously proposed model [Figure I11].

In the horse, the coexistence of satellite-based and satellite-less centromere makes this species a particularly useful model for studies on the role of centromeric repeats. The physical relations among the major horse satellite DNA families (37cen, 2PI, and EC137) at satellite-based centromeres were investigated (Nergadze SG et al. 2014) taking advantage of two color FISH on stretched chromosomes and on combed DNA fibers. The 37cen sequence consists of a 221 bp repeat, 2PI sequence is formed by 23 bp repeated units and EC137 satellite is composed of 137 bp long units. 37cen was demonstrated to be the most represented satellite DNA family in the horse genome. It colocalized the primary constriction on all chromosomes except ECA11 and it can spread in the pericentromere. On the

contrary, the 2PI and EC137 sequences are less abundant, EC137 being pericentromeric, and partially overlapping with 37cen.
Nergadze and co-workers (2018) analysed mechanically stretched chromosome preparations and suggested that 2PI, being often present in pericentromeric uncoiled regions, could have a role in driving the pericentromeric heterochromatin supercoiling which is needed for the correct architectural organization of the centromere core (Blower MD et al. 2002), .Other cytogenetics analyses revealed that small arrays the 2PI and EC137 satellites (ranging in size from 2–8 kb) are strictly intermingled and immerged within blocks of the 37cen sequence extending for hundreds of kilobases (reference?). This organization highlights the plasticity of satellite DNA. ChIP-seq and high resolution immune-FISH experiments demonstrated that, in the horse, 37cen is bound by CENP-A (Cerutti F et al. 2016). Sequence analysis showed that 37cen sequence associated to CENP-A is organized in a head-to-tail fashion and is GC- rich. Moreover the horse seems to share the same CENP-A blocks organization (within arrays of satellite DNA) with other species (Blower et al. 2002).
The satellite-less centromere of horse chromosome 11 was the first to be analyzed on a molecular level (Wade CM et al. 2009; Purgato S et al. 2015) by our laboratory, leading to the observation of a phenomenon named "centromere sliding" (see section 11 Centromere sliding).

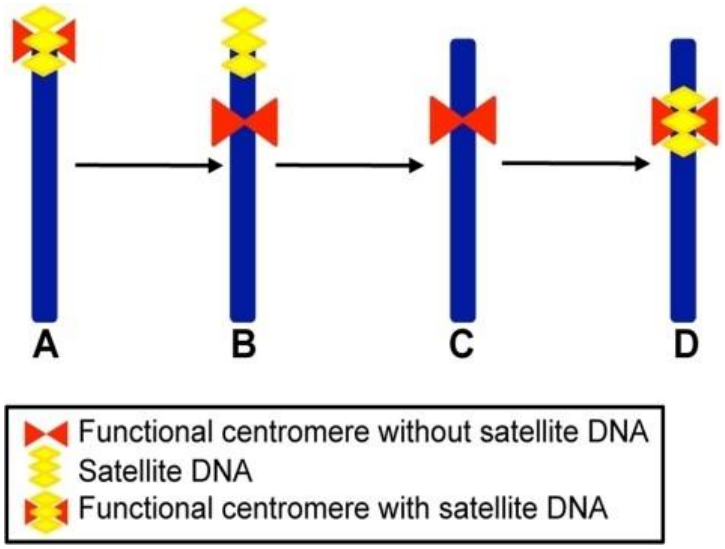


Figure I11: Schematic representation of the four-step mechanism for neocentromere formation during evolution*. A) Acrocentric ancestral chromosome carrying satellite DNA (yellow) at its terminal centromere (red). B) Sub-metacentric chromosome derived from centromere repositioning; the chromosome maintains satellite DNA sequences at the terminal position, coinciding with the old centromere site, while the neocentromere is devoid of repetitive sequences. C) Sub-metacentric chromosome derived from (B) in which the terminal satellite sequences have been lost. D) Sub-metacentric chromosome in its full maturation stage carrying satellite DNA at the neocentromere site (Piras MF et al. 2010).*

10 Centromere repositioning

A fundamental step in understanding centromere biology was the discovery that the ENC at horse Chromosome 11 is completely devoid of satellite DNA (Wade CM et al. 2009). This observation revealed, for the first time, that a satellite-free centromere can be present in all individuals of a vertebrate species as a normal karyotype component. This centromere is established on a segment of DNA, conserved in vertebrates, which is free of genes as well as of satellite DNA, providing an example of an evolutionarily "young" ENC which has not acquired repetitive sequences. Satellite-free centromeres were subsequently observed in chicken (Shang WH et al. 2010), orangutan (Locke DP et al. 2011) and potato (Gong Z et al. 2012). In this context, centromere repositioning demonstrates the evolutionary plasticity of centromeres (Montefalcone G et al. 1999). Comparison of the karyotypes among primate species showed that centromere position can change without a corresponding change in DNA structural organization (Montefalcone G et al. 1999; Cardone MF et al. 2006; Ventura M et al. 2007). These Evolutionarily New Centromeres (ENCs) suggest that centromere evolution seems to be driven by forces other than the surrounding DNA, such as epigenetic markers.

A relationship between ENCs and the analphoid neocentromeres observed in human clinical samples emerged from analysis of the positions in which these events occur. For example, human neocentromeres at Chromosome 3, 9 and 6 occur in the same genomic regions as ENCs observed in some primates, indicating that certain regions of the genome have a propensity to form centromeres (Ventura M et al. 2004; Capozzi O et al. 2008, 2009). Thus, regions of the genome may harbour 'latent' centromere potential (Voullaire LE et al. 1993). The observation that the primate ENCs possessed typical arrays of alpha satellite DNA led to the hypothesis that epigenetic marks can drive the movement of centromere function to new genomic sites which can subsequently mature through the acquisition of satellite DNA sequences (Amor DJ and Choo KH 2002; Piras MF et al. 2010; Kalitsis J and Choo KH 2012). Following their original discovery in primates, a surprisingly large number of ENCs were identified in the genus *Equus* (Carbone L et al. 2006; Piras MF et al. 2009) and some examples were also observed in other animals (Ferreri GC et al. 2005; Kobayashi T et al. 2008) and in plants (Han Y et al. 2009), indicating that centromere repositioning is a widespread force for karyotype evolution.

The horse karyotype was considered as the ancestral configuration to analyze chromosomal fusions across the equids. It was considered also the ancestral configuration for the analysis I reported in Results Part 2, particularly when centromere repositioning events were analyzed.

11 Centromere sliding

The initial description of two, well defined peaks of CENP-A at the satellite free centromere of ECA11 did not determine whether the observation reflected a single chromosome with two peaks or two chromosomes each with a peak at a different location (Wade CM et al. 2009) Centromeric regions were characterized in five additional individuals and their positional analysis showed that one or two CENP-A binding domains can be present in a single individual. The region comprising the localization of each peak varies within a 500 kb long region (Purgato S et al. 2015). The broader impact of these results suggests that the centromere function is not coupled to a specific sequence but can slide within a relatively wide region. SNP based approach and immune-FISH experiments on single chromatin fibers, demonstrated that the two CENP-A binding domains correspond to the localization of the centromeric function on the two homologs with high positional variation into the population giving rise to multi-allelic epigenetic polymorphism [Figure I12].

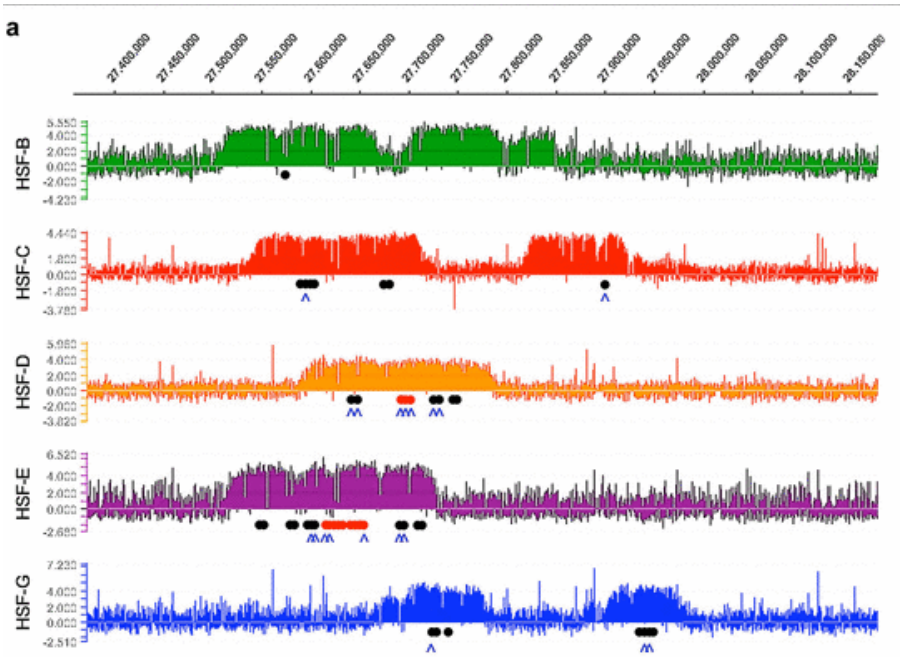


Figure I12 Positional shift of the horse centromere at chromosome 11. *Different CENP-A binding profile, visualized as different peaks on five horse individuals are shown. Results are presented as the log2 ratio of the hybridization signals obtained with immuno-precipitated DNA versus input DNA (y axis); genomic coordinates on ECA11 are plotted on the x axis (adapted from Purgato S et al. 2015).*

This analysis also suggested that CENP-A nucleosomes displays high mobility and instability, a property that could be related to the evolutionary mobility of centromeres. In this scenario, satellite DNA may provide positional stability to this domain along the chromosome. Thus, centromeres exhibit large scale relocalization

(centromere repositioning) during evolution as well as short range relocalization (centromere sliding) within a population (Giulotto E et al. 2017).
The mechanism underlying the centromere sliding remains unknown, as well as the timing of the movement. Our recent work (Nergadze SG et al 2018) proved that centromeric movement occurs in one generation. Taking advantage of equid hybrids (mule) we revealed that the centromeric domain can shift along the chromosome in one generation. Moreover, we did not detect any centromeric sliding when clonal cell lines were analyzed.

12 Satellite-less Centromeres in other equid species

Cytogenetic experiments (Figure I10) (Piras MF et al. 2010) suggested that chromosomes lacking satellite DNA are present in other equid species. Particularly, 17 Grevy's zebra chromosomes and 9 Burchell's zebra chromosomes were negative for centromeric satellite as reported for one horse chromosome and 16 donkey chromosomes. This peculiar feature (the presence of a high number of satellite-less centromeres) may be explained by the fact that equids recently diverged and underwent rapid karyotype evolution.
The karyotype of different equid species was studied using chromosomal painting probes. Several chromosomal rearrangements were detected in different species of the genus *Equus* (Richard F et al. 2001; Yang F et al. 2003; Trifonov VA et al. 2008; Musilova P et al. 2009, Figure I13-18). Following the reconstruction of a perissodactyl ancestral karyotype (Trifonov VA et al. 2008), its analysis suggested that the ancestral karyotype consisted of several acrocentric chromosomes, which underwent fusions during evolution, reducing the chromosome number. These events may be the force triggering the formation of a satellite-less centromere as the result of a tandem fusion of a chromosome with the centromeric region of an acrocentric chromosome, while satellite repeats are lost. Satellite-less centromere formation in equids is relatively frequent (Rocchi M et al. 2012) and has been well documented (Carbone L et al. 2006; Piras MF et al. 2009).
53 fusion events were identified. For this study, the horse karyotype was considered closest to the ancestral configuration to analyze chromosomal fusions across the equids. It was considered also the ancestral configuration for the analysis I reported in Results Part 2, particularly when centromere repositioning events were analyzed.
After centric fusion events some chromosomes maintained the centromere at the fusion site (and the satellite DNA), others underwent centromere repositioning (Musilova P et al. 2013).

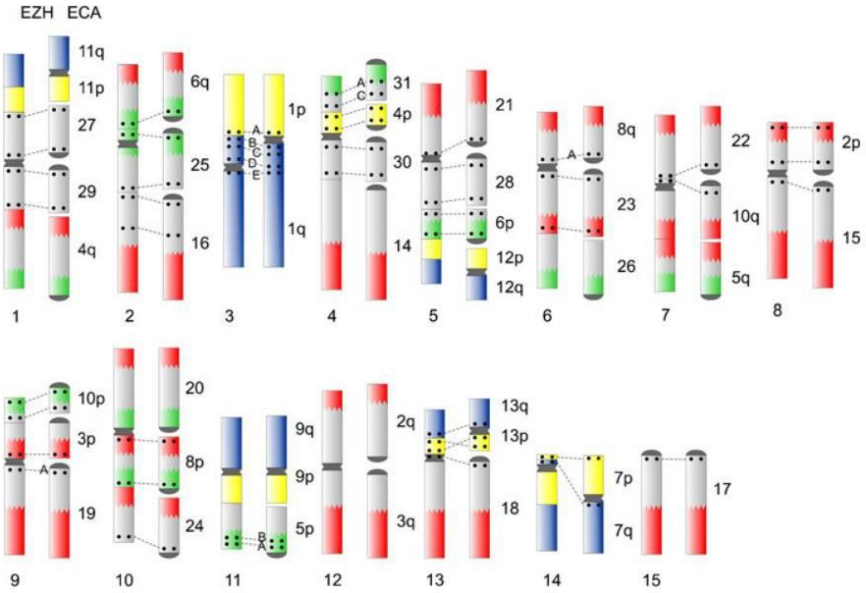


Figure I13: Diagram summarizing subchromosomal comparative mapping between E. zebra hartmannae and E. caballus. *Hartmann's chromosomes are shown on the left of each pair, while horse chromosomes are shown on the right. The hybridization signal of horse distal/proximal region-specific probes and arm-specific painting probes on orthologous chromosomes are depicted by different colors: distal, red; proximal, green; p-arm, yellow, q-arm, blue. BAC clone locations are represented by double dots. Dashed lines connect the corresponding BAC signals in orthologous chromosomes. Capital letters refer to the used BACs (Musilova P et al. 2013).*

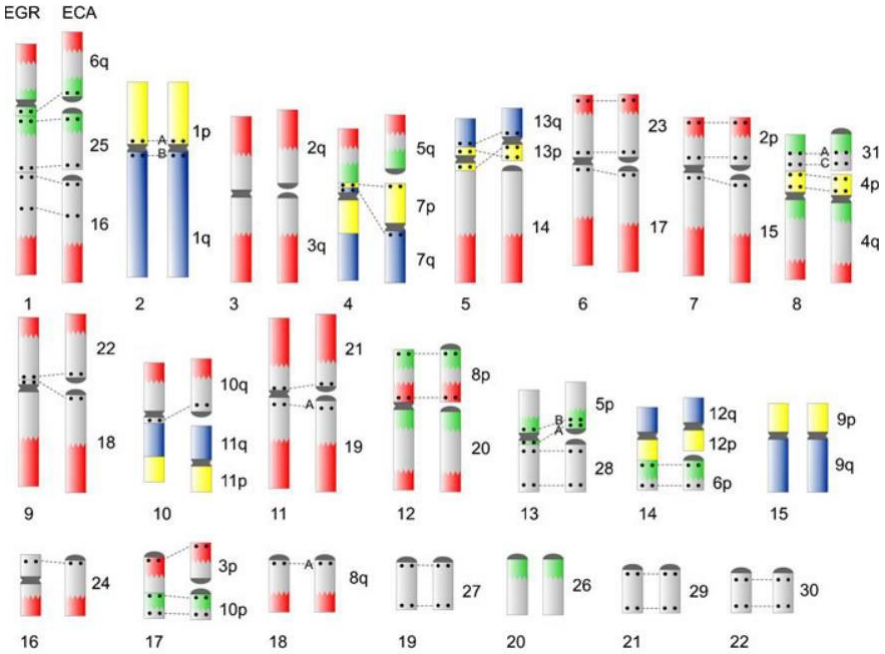


Figure I14: Diagram summarizing subchromosomal comparative mapping between E. grevyi and E. caballus. *Grevy's chromosomes are shown on the left of each pair, while horse chromosomes are shown on the right. The hybridization signal of horse distal/proximal region-specific probes and arm-specific painting probes on orthologous chromosomes are depicted by different colors: distal, red; proximal, green; p-arm, yellow, q-arm, blue. BAC clone locations are represented by double dots. Dashed lines connect the corresponding BAC signals in orthologous chromosomes. Capital letters refer to the used BACs (Musilova P et al. 2013).*

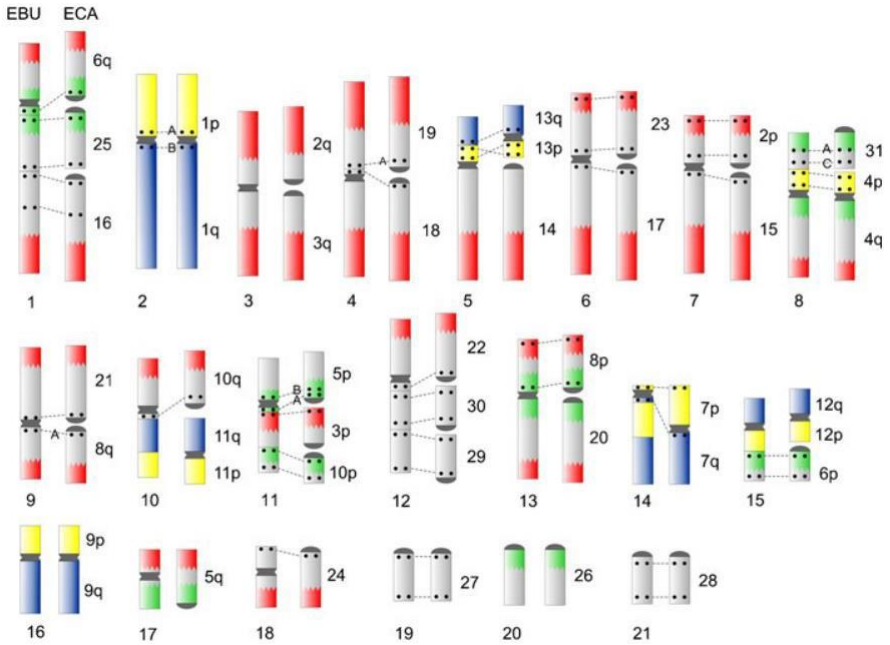


Figure I15: Diagram summarizing subchromosomal comparative mapping between E. burchelli and E. caballus. *Burchell's chromosomes are shown on the left of each pair, while horse chromosomes are shown on the right. The hybridization signal of horse distal/proximal region-specific probes and arm-specific painting probes on orthologous chromosomes are depicted by different colors: distal, red; proximal, green; p-arm, yellow, q-arm, blue. BAC clone locations are represented by double dots. Dashed lines connect the corresponding BAC signals in orthologous chromosomes. Capital letters refer to the used BACs (Musilova P et al. 2013).*

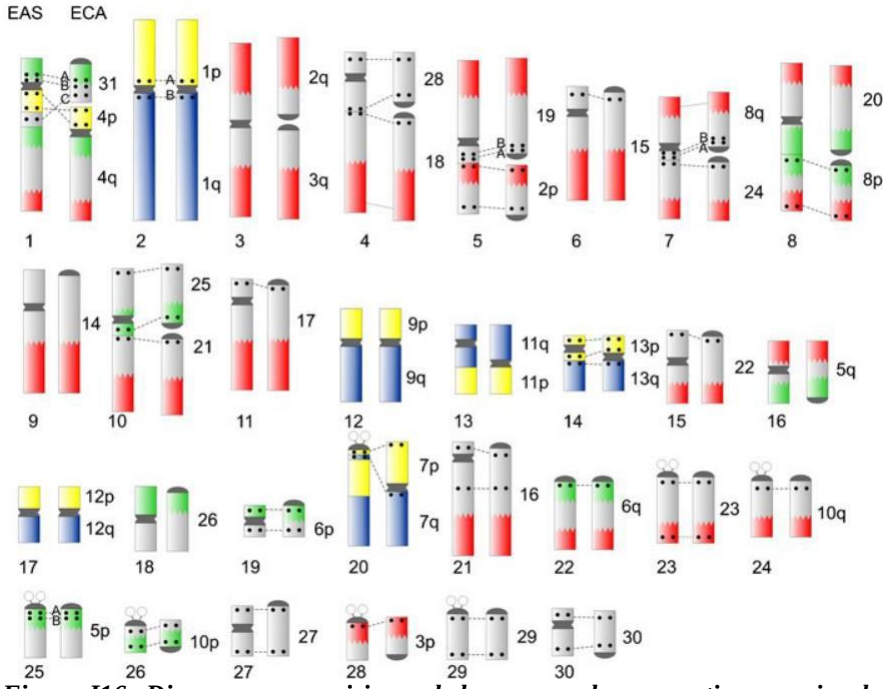


Figure I16: Diagram summarizing subchromosomal comparative mapping between E. asinus and E. caballus. *donkey's chromosomes are shown on the left of each pair, while horse chromosomes are shown on the right. The hybridization signal of horse distal/proximal region-specific probes and arm-specific painting probes on orthologous chromosomes are depicted by different colors: distal, red; proximal, green; p-arm, yellow, q-arm, blue. BAC clone locations are represented by double dots. Dashed lines connect the corresponding BAC signals in orthologous chromosomes. Capital letters refer to the used BACs (Musilova P et al. 2013).*

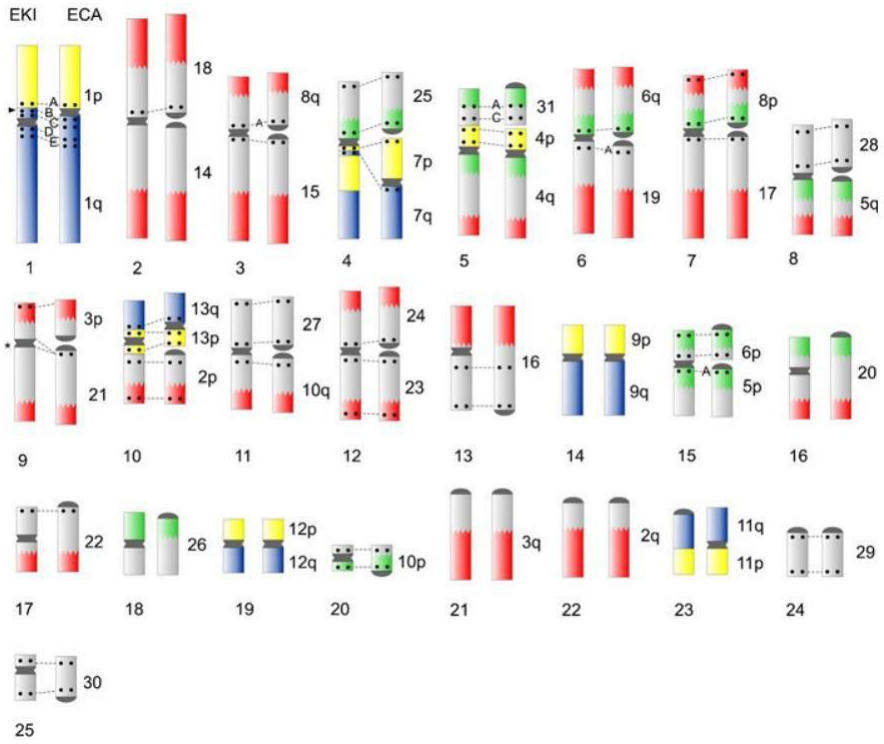


Figure I17: Diagram summarizing subchromosomal comparative mapping between E. kiang and E. caballus. *kiang's chromosomes are shown on the left of each pair, while horse chromosomes are shown on the right. The hybridization signal of horse distal/proximal region-specific probes and arm-specific painting probes on orthologous chromosomes are depicted by different colors: distal, red; proximal, green; p-arm, yellow, q-arm, blue. BAC clone locations are represented by double dots. Dashed lines connect the corresponding BAC signals in orthologous chromosomes. Capital letters refer to the used BACs (Musilova P et al. 2013).*

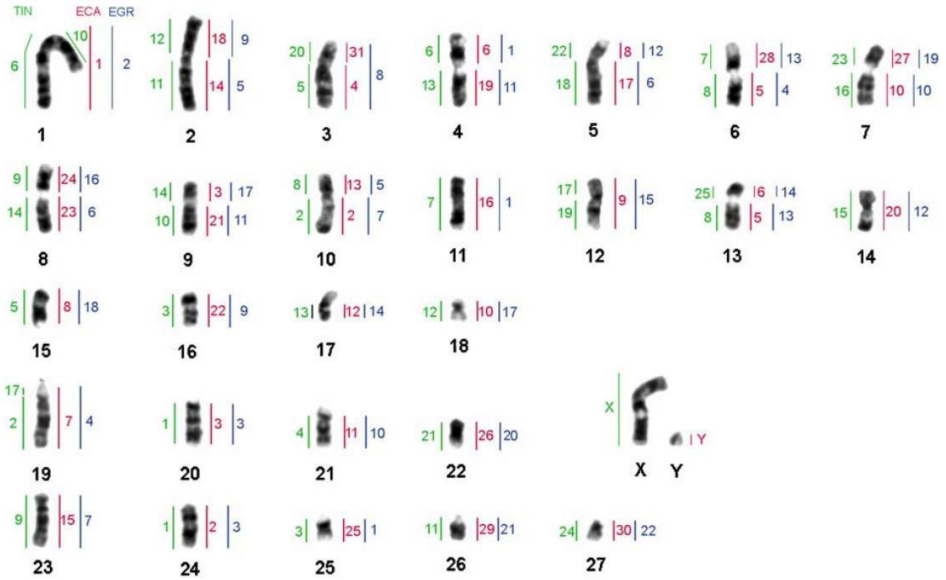


Figure I18 Karyotype of onager (EHE) with Malayan tapir (TIN), horse (ECA) and Grevy's zebra (EGR) homologies shown on either side of the EHE chromosomes (Trifonov VA et al. 2008).

AIMS OF THE RESEARCH

- Previous work showed that some equid chromosomes lack satellite DNA at the centromeric locus. In particular, through ChIP-seq, one satellite-less centromere in the horse and 16 satellite-less centromeres in the donkey were identified. Therefore we wanted to test whether other equid species do bear satellite-less centromeres. To this purpose we tested the presence of satellite-less centromeres in 5 other equid species. We also aim to analyze the phylogenetic relationship among *Equus* species using the centromere as epigenetic marker.

- As previously reported in the PhD thesis of Riccardo Gamba (2017), we analyzed the epigenetic and transcriptional profile of the horse and donkey satellite-less centromeres. In this work we further investigated on the epigenetic state of the centrochromatin analyzing H4k20me1 and H3k4me3 histone markers on the centromeric regions of the horse and the donkey which were not previously analyzed.

- We know that CENP-A is one of the most important determinant of centromere function but it is also present on non-centromeric loci throughout the genome. Very little or none is known about the possible role of this protein on loci other than the centromere. However it may play a role on the epigenetic control of some molecular pattern, as for gene expression. To this aim we planned to identify CENP-A secondary binding domains, characterize its distribution on functional genomic elements and possibly find new epigenetic regulatory functions of the Centromere Protein A.

- Transcription at centromeric level was first reported for mouse satellite-based centromeres but evolutionarily new centromeres and human neocentromeres were found to arise in gene desert regions. Thanks to the equid model system, we can evaluate the transcriptional profile of centromeres which are lacking satellite DNA and, unlike the clinical neocentromeres, are fixed in a species. To this aim we wanted to evaluate the transcriptional status of the satellite-less centromere of horse chromosome 11 in different tissues of two horse individuals.

MATERIAL AND METHODS

1 Cell cultures and animals used in this work

Primary fibroblast cell lines from HorseS and DonkeyA were previously established in our laboratory from the skin of slaughtered animals and used in Part 1-2-3-4. The DonkeyB, HorseA and HorseC, cell lines used in Part 1 kindly provided to us by Professor Douglas F. Antczak from the Cornell University (Ithaca, NY, USA), were obtained from skin samples collected using a 2mm punch biopsy tool. Tissue was digested in 0.25% Trypsin-EDTA (Gibco, Grand Island, NY) for 10 minutes at 37° C, so the epidermal layer could be scraped off with a scalpel and discarded. Remaining tissue was minced and digested with 0.25% Trypsin-EDTA at 37 o C for 10 minutes. Digested tissue was gently triturated and passed through a cell strainer to yield a single cell suspension. Cells were washed in DMEM (Gibco) and plated in tissue culture flasks with DMEM supplemented with 10% FBS (Hyclone, Logan, UT) and Penicillin-Streptomycin (Gibco).

The hybrid cell lines were obtained from embryos derived from in vitro fertilization and used in Part 1. To this purpose sperm cells from a single male donkey were used to fertilize oocytes from three different female horses. The resulting hybrid embryos were implanted and, after 32-34 days, three mule conceptuses were obtained via uterine lavages, as described in (Adams and Antczak 2001). Conceptuses were dissected under a microscope into discreet tissues, including the tail/hind quarter region which is comprised of primarily fibroblasts at this stage of development. The tail region was minced, spun down, washed in 1x Phosphate Buffered Saline (PBS) and resuspended in DMEM (Invitrogen) plus 10% FBS for culture.

Mule immortalized fibroblasts were grown in the same medium of primary fibroblasts supplemented with G418 sulphate (Invivogen) at the final concentration of 0.4 mg/ml.

Burchell's and Grevy's zebra fibroblasts were purchased from Coriell Repositories. Hartmann's zebra, kiang and onager fibroblasts, derived from skin biopsies, were kindly provided by Dr. Oliver Ryder (Center for Reproduction of Endangered species, Zoological Society of San Diego). These cell lines were used for Part 2. The fibroblasts used for all the experiments were isolated and established from skin biopsies under sterilized conditions. Primary fibroblasts were cultured in high-glucose DMEM (EuroClone) medium supplemented with: 20% fetal calf serum (EuroClone), 2x NEAA (non-essential amino acids, EuroClone), 2mM L-glutamine (SIGMA), 1x penicillin/streptomycin (SIGMA).

Here is a list of the Latin and common names of species used in this work. *Equus caballus*, horse. *Equus kiang*, kiang. *Equus hemionus onager*, onager, *Equus asinus*, donkey. *Equus grevyi*, Grevy's zebra. *Equus burchelli*, Burchell's zebra. *Equus zebra hartmannae*, Hartmann's zebra.

All cells were maintained in a humified atmosphere of 5% CO2 at 37°C.

2 Chromatin Immunoprecipitation (ChIP)

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

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

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

Pre-clearing was performed with A/G beads (Protein A SepharoseTM 4 Fast Flow/Protein G SepharoseTM 4 Fast Flow, GE Healthcare), previously treated with a blocking buffer (phosphate-buffered saline containing sonicated E. coli genomic DNA 500 ng/ul and BSA 10 mg/ml) for 1 hour at 4°C on shaking. Then, after centrifugation at 4000 rpm for 5 minutes at 4°C, the supernatant was recovered and beads discarded. 240 ul of the supernatant were saved as Input (20% of the total chromatin used for each IP). The remaining part was divided into aliquots and incubated first with the selected antibody, followed by incubation with previously treated A/G beads for 3 hours at 4°C on the rocket. Samples were then centrifuged for 2 minutes at 1200 g at 4°C and the supernatant was removed. The beads were washed 5 times with cold ChIP wash buffer (0,25% SDS, 1% TritonX-100, 2 mM EDTA pH 8, 150 mM NaCl, 20 mM Tris-HCl pH 8) and the last wash with cold ChIP final wash buffer (0,25% SDS, 1% TritonX-100, 2 mM EDTA pH 8, 500 mM NaCl, 20 mM Tris-HCl pH 8). After completely discarding the last wash, the immunocomplexes were eluted adding ChIP elution buffer (1% SDS, 100 mM NaHCO3, 40 ug/ml RNase A). Samples were incubated at RT for 15 minutes, then at 37°C for 1 hour and finally reverse cross-linked at 65°C, over-night. The day after the DNA was purified and eluted using the kit Promega (Wizard SV Gel and PCR Clean-up System) according to the manufacturer's instructions.

After purification, the DNA was quantified using the QuantusTM Fluorometer

(Promega) with the QuantiFluor® ONE dsDNA System (Promega) according to manufacturer's instruction.
These experiments (Cell cultures and ChIPs) were carried out by a collaborator of mine, Dr Francesca Piras (University of Pavia).
Cytogenetic experiments and analysis were carried out by Dr Francesca Piras and the PhD student Eleonora Cappelletti.

3 Antibodies

Chromatin from the Immortal Parental MuleA cell line and from its clones was immuno-precipitated with a human CREST serum, provided by Dr. Claudia Alpini (health institute "Fondazione I.R.C.S.S.-Policlinico San Matteo"), whose CENP-A specificity was previously demonstrated (Purgato S et al. 2015).
ChIP-seq experiments of Part 1 and 5, targeted at the protein CENP-A, were performed with a purified polyclonal antibody against human CENP-A protein, kindly provided by Prof. Mariano Rocchi (University of Bari).
ChIP-seq experiments of Part 2 and 5, targeted at the protein CENP-A, were performed with an antibody against the horse CENP-A raised in sheep (called Bleed3), which was kindly provided by Prof Kevin Sullivan (University of Galway).
The ChIP-seq experiments described in Part 3 were performed with commercially available antibodies: anti-H3K9me3 (Abcam ab8898), anti-H3K4me3 (Abcam ab12209), anti-H3K4me2 (Abcam ab32356), anti-H4K20me1 (Novus NBP1-30091PCP).

4 ChIP sequencing and bioinformatic analysis

Immuno-precipitated and input DNAs were paired-end sequenced through an Illumina HiSeq2000 or HiSeq2500 platform by IGA Technology Services, Udine, Italy. Sequence reads were aligned to the horse reference genome (EquCab2.0, Equcab3.0) or to the EquCabAsi references with Bowtie 2.0 (Langmead B and Salzberg SL 2012). The method used to assemble the EquCabAsi reference genome is described in the attached paper.
To identify contigs that are significantly enriched after immuno-precipitation with anti-CENP-A, peak-calling was performed through the software MACS version 2.0.10 (Zhang Y et al. 2008), using default parameters. Then, stringent criteria (Bailey T et al. 2013) were applied to identify significantly enriched regions: fold enrichment > 5, pile-up > 100, -log10(p-value) > 100, -log10(q-value) > 100.
To define the start and end position of the peaks and compare their location between different datasets, peak-calling was performed with MACS software version 2.0.10 (Zhang Y et al. 2008). When the peak-calling identified a single region underlying one centromeric peak, the coordinates of that region were used

as coordinates of the peak. When more than one region was identified corresponding to one centromeric peak, the start coordinate of the firs region was used as start coordinate of the peak. The end coordinate of the last region was used as end coordinate of the peak.
All ChIP-seq plots of Parts 2-5 were obtained with the Integrative Genome Viewer (IGV) software (Robinson JT et al. 2011).
Is it important to remind that the reads used in Part 2 were mapped on the horse reference genome (Equcab3.0), so species specific sequences are absent from the analysis. For this reason some reads obtained from the different species(asses and zebras) could have been not mapped, therefore some information from the ChIP-seq datasets could have been missed. Also different reference genome assembly (Equcab2.0 and Equcab3.0) may give different results when reads are mapped. For example, in Burchell's zebra chromosome 9 no FISH signals, corresponding to satellite DNA, were detected (Piras MF et al. 2010) and no CENP-A binding domains were mapped when EquCab2 was used as reference. However, when we mapped the same reads on the new version of the genome, EquCab3, we found a ChIP-seq signal for that centromere, whose horse orthologous sequence was in ECA21 (Figure R2-20).

5 RNA bioinformatic analysis

RNA-seq datasets presented in Part 5 were obtained from the horse community within the FAANG consortium in which we are associated with. Datasets were downloaded from a cloud storage and analyzed. RNA-seq datasets were analyzed for quality check and trimmed through TrimGalore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), then reads were mapped on the horse reference genome Equcab3 through TopHat software version 2.1.0 (Kim D et al. 2013). Mapped read files were converted using BedTools (https://github.com/pezmaster31/bamtools) and Samtools (Li H et al. 2009). Bigwig files were loaded and visualized on the Integrative Genome Viewer (IGV) software (Robinson JT et al. 2011).
MiRNA-seq datasets presented in Part 5 were also obtained from the horse community within the FAANG consortium. Datasets were downloaded from a cloud storage and analyzed. MiRNA-seq datasets were analyzed for quality check and trimmed through TrimGalore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), then reads were mapped on the horse reference genome Equcab3 through Bowtie2 (Langmead B and Salzberg SL 2012). Mapped read files were converted using BedTools(https://github.com/pezmaster31/bamtools) and Samtools (Li H et al. 2009). Bigwig files were loaded and visualized on the Integrative Genome Viewer (IGV) software (Robinson JT et al. 2011).

6 Peak calling annotation

Since we are interested in secondary CENP-A binding sites, we deleted all the peaks mapping on neocentromeres and Unplaced Chromosome of the different datasets (the Unplaced chromosome is a fictional one that contains all that contigs that were located on actual physical chromosomes

The analysis performed on the secondary CENP-A binding sites was done through HOMER (Hypergeometric Optimization of Motif Enrichment)(Heinz S et al. 2010), which is a suite for *de novo* Motif Discovery and next-generation sequencing analysis. Through this software, we can annotate peak-calling results using species-specific transcriptomes, to get the genomic context of each peak (e.g. if it is within promoter-TSS, exon, intron, TTS or intergenic regions). HOMER needs the output of the MACS software and a list of genes present in the genome as input; we retrieved the lists of the genes present in the horse, mouse and human genes from UCSC Genome Browser: Ensembl Genes for the horse and donkey, GENECODE VM9 (Ensembl 84) for the mouse and NCBI RefSeq genes for the human.

7 Gene expression analysis

Gene expression analysis was performed using the human transcriptome (Human Protein Atlas available from www.proteinatlas.org, Uhlen M et al. 2015) and mouse transcriptome(Söllner JF et al. 2017).

The RNA-seq data was used to classify all genes according to their tissue specific or cell line specific expression into one of six different categories, defined based on the total set of all TPM values:

- Tissue/Cell line enriched (expression in one tissue at least five-fold higher than all other tissues/cell lines)
- Group enriched (five-fold higher average TPM in a group of two to seven tissues/cell lines compared to all other tissues/cell lines
- Tissue/Cell line enhanced (five-fold higher average TPM in one or more tissues/cell lines compared to the mean TPM of all tissues/cell lines)
- Expressed in all ($\geq$ 1 TPM in all tissues/cell lines)
- Not detected (< 1 TPM in all tissues/cell lines)
- Mixed (detected in at least one tissue/cell line and in none of the above categories)

Tissue specificity for each gene was assigned only if a single gene was found into the first three categories. Once obtained all this data, the table of gene-associated peaks was crossed with the one contained tissue-specificity for each gene, in order to assess the tissue profile expression of each gene-associated peak. For the cancer

related gene's analysis, we obtained a list of 1,571 manually curated protein-coding cancer genes(An O et al. 2016) that we crossed with the list of gene-associated CENP-A peaks obtained through HOMER.

8 Motif analysis

Motif analysis was performed using HOMER as a *de-novo* motif discovery tool. It uses the starting and ending point of the sequences that we want to analyze as input, formatted in a .bed file. We looked for motif in all the secondary CENP-A binding sequences (target sequences) divided in each genomic region (e.g. promoter-TSS, intron, exon, TTS and intergenic regions). Standard parameters were used for the motif discovery analysis (http://homer.ucsd.edu/homer/ngs/peakMotifs.html): the motif should be present within a 500 bp region around the center of ectopic CENP-A peaks. The frequencies of target sequences containing the motif were calculated as well as the frequencies of random sequences (background, chosen matching C+G-content of target sequences) containing the motif. The p-value was calculated as the probability that the frequency observed in the target sequences was significantly different from the one observed in the background sequences by chance. We used as statistically significant threshold a p-value $< 1.0x10^{-50}$.
These experiments (Peak calling annotation, Gene expression analysis and Motif analysis) were carried in collaboration with the master student Antonio Rausa.

PART 1 BIRTH, EVOLUTION AND TRANSMISSION OF SATELLITE-LESS MAMMALIAN CENTROMERIC DOMAINS

Here I briefly report the results and the discussion of the attached manuscript in which I am co-first author.

The study of the chromatin domain of mammalian centromeres has so far been hindered by the presence of satellite DNA, highly repetitive DNA. The histone H3 variant CENP-A is the main centromeric determinant, therefore, centromeres are epigenetically specified.

In previous work from our laboratory (Piras MF et al. 2010) 16 donkey centromeres were reported of lacking satellite DNA sequences through cytogenetic analysis. The first example of a natural satellite-less centromere was identified in the *Equus caballus* on Chromosome 11 (Wade CM et al. 2009). To assess the exact CENP-A centromeric binding domain and to confirm the absence of repetitive sequence at molecular level we performed ChIP-seq experiments on donkey primary skin fibroblasts. We investigated the satellite-less centromeres of *Equus asinus* by using ChIP-seq with anti-CENP-A antibodies. Through bioinformatic analyses, we identified an extraordinarily high number of centromeres lacking satellite DNA: 16 out of 31.

First, we mapped the ChIP-seq reads obtained by the donkey cell lines to the horse reference genome, Equcab2, since it was the only well assembled genome at chromosomal level among equids. Figure 1 of the paper shows the graphical representation of the enrichments profile of peaks corresponding to the satellite less centromeres in the donkey and the peak corresponding to the centromere of horse chromosome 11. Peaks showed different type of enrichment profile. While some peaks showed a Gaussian-like regular shape (such as EAS4 and EAS30), other peaks were irregular (such as EAS8 and EAS14), contained gaps (such as EAS7 and EAS14) or exhibited a narrow, spike-like distributions, such as EAS9 and EAS19. The 16 donkey centromeric regions spanned 54-345 kb and contained one or two CENP-A binding domains.

Since we mapped donkey reads on the horse reference genome we wondered if the peak profile heterogeneity was due to differences in DNA sequence between the two species. We sequenced the donkey centromeres by assembling Illumina reads and carrying out Sanger sequencing of regions amplified by PCR to resolve gaps in the assembly. For each centromeric region, genomic segments ranging in size between 157 and 358 kb were assembled. To avoid bias due to read mapping on short sequences (our assembled centromeres), instead of an entire reference genome, we constructed a chimeric reference genome by inserting the assembled centromeric donkey contigs in EquCab2.0 to replace their orthologous horse

sequences. This reference genome was named EquCabAsiA. Reads were then mapped on the newly assembled genome and a comparison of the peak profiles obtained with the two reference genomes was performed. The comparison analysis showed that large gaps and irregular peak profiles disappeared on the chimeric genome, proving that CENP-A binds uninterruptedly the centromeric region of the satellite-less donkey centromeres similarly to horse Chromosome 11 (Wade CM et al. 2009).

We performed a sequence analysis to detect what kind of rearrangements occurred between the donkey and horse sequence of the 16 satellite-less centromeres. Analysis of the centromeric domains of EAS8, EAS9, EAS16, EAS18 and EAS19 suggested the presence of donkey specific tandem repetitions that are in single copy in the horse orthologous non-centromeric regions. Specific experiments were done to confirm the presence of sequence amplification at these five loci.

Southern blotting, qPCR and read count experiments were carried out on four individuals: one horse (HorseS), two donkeys (DonkeyA and DonkeyB) and a mule (MuleA), offspring of DonkeyB were used to prove that a subset of these centromeres is associated with DNA amplification. Figure 2 of the paper shows the results of this analysis done on EAS9 EAS18 and EAS19. All these three independent experiments proved the presence of tandem sequence amplification at a subset of centromeres in the donkey, with evidence for marked inter-individual variation in copy number at some of these loci

We analyzed DNA sequence features within the satellite-free donkey assembled centromeres in comparison with the corresponding regions in the horse genome. SINEs, LINEs, LTR-derived sequences, transposable DNA elements and GC content at the donkey centromeric domains did not differ from the orthologous horse sequences, thus indicating that no significant modification of transposable elements and GC content occurred after centromere formation. We then compared the percentage of transposable elements at this loci compared to the average genome wide values using previously published data on the donkey (Huang J et al. 2015). Surprisingly the assembly of donkey centromeres showed to be poor in SINE, LINE rich and with same abundance of LTR and DNA elements when compared to the rest of the genome. GC content analysis reveals instead that the satellite-less centromeres are AT rich.

The double peaks observed on several chromosomes (EAS5, EAS10, EAS12, EAS14 and EAS18) suggested the presence of different epialleles on the two homologs in the donkey as previously reported for horse Chromosome 11 (Purgato S et al. 2015). To detect and prove the presence of epialleles we used a Single Nucleotide Variation (SNV) based approach. Thanks to both ChIP and input reads, we were able to discriminate between nucleotide variants present in the two alleles. Figure 3 of the paper shows the result of this analysis. Red and green dots under the peaks represent nucleotide variants present only in one of the two homologs, while yellow dots indicate variants present in both alleles. We were able to prove that in 8-total satellite-less centromeres analyzed (the only informative ones), extensive positional allelism occurs at most donkey centromeres with specific CENP-A binding pattern on each homolog.

To further investigate the individual variability of the donkey satellite-free centromeric domains, we analyzed an additional unrelated donkey (DonkeyB) by ChIP-seq with the same anti CENP-A antibody used for the first donkey. After read mapping on the horse reference genome EquCab2, we identified 15 satellite-less centromeres in this donkey individual. Centromere on EAS8 was not detected, maybe because it is positioned on a satellite-containing sequence, possibly highlighting some interindividual polymorphism. We carried out a positional allelism analysis between the two-donkey individual to better characterize how centromeric CENP-A binding domains shift along the chromosome.
A marked variability in the position of CENP-A binding domains between the two individuals was observed at the six chromosomes. No positional variability was detected on the other 9 chromosomes.
For instance we reported the presence of one peak in EAS4 and EAS7 only in the DonkeyA individual. Two peaks were detected instead on the DonkeyB individual (Figure 3 of the paper).
In conclusion, comparison of the satellite-free centromeres between the two donkeys showed that CENP-A binding domains can shift within regions of up to 600 kb.
We then investigated the germ-line and somatic transmission of the centromeric domains since we wondered how and when such centromeric movement occurs. The stability of centromeres across generations was examined by crossing DonkeyB with three mares (HorseA, HorseB and HorseC) by *in vitro* fertilization. Embryonic fibroblasts were established from the resultant mule concepti (MuleA, MuleB and MuleC). Adult skin fibroblast cell lines were established from DonkeyB and from two of the three mares (HorseA and HorseC; cells from HorseB were not available). Since we proved sequence positional variability between the centromeric domains of DonkeyA and DonkeyB, we thus constructed a new chimeric genome (EquCabAsiB) containing the assembled satellite-less centromeres of DonkeyB, using the same approach previously applied to DonkeyA. We ChIP-sequenced all the established cell lines using the same anti- CENP-A antibody. We mapped all the reads on the newly assembled reference genome and carried out a positional variability analysis of the centromeric loci among the above-mentioned individuals. Figure 4 in the paper shows the family trees of the individuals and the analysis done on EAS4 and EAS7. Both centromeres have two distinct peaks in DonkeyB while each mule inherited only one, revealing independent assortment of epialleles and normal monoallelic transmission. No variations were observed among replicates indicating absence of experimental variability. Regarding EAS4, MuleA inherited the left peak in the same position, MuleB inherited the right peak but shifted by about 50 kb and MuleC inherited the left peak with little to no movement. In chromosome 7 all three mules inherited the left with a major shift of about 50 kb in MuleB. Analyzing 30 segregation events in three mules (donkey centromeres with sequence amplification were discarded from the analysis) we observed clear positional movement in 5 out of 33 (including 3 horse centromeres) transmission events. In the remaining cases, little or no movement was detected. We then wondered if this centromere sliding occurs

during propagation in culture, so we examined centromeric positional stability in six clonal cell lines isolated from immortalized fibroblasts derived from MuleA. Applying the same ChIP-seq technique and after mapping the reads on the chimeric genome (EquCabAsiB), we localized the satellite-less centromeres and performed a positional variability analysis. Figure 5 in the paper shows that no relevant changes in peak position were detected among the clones and between the clones and the immortal parental cell line; this analysis proves that centromeric position in the immortal cell population was homogeneous despite the high number of cell and sliding did not occur. We were able to prove that centromere sliding, observed in the families, does not occur *in vitro* cell culturing.

The analysis of epiallele transmission in hybrids (three mules and one hinny) showed that centromeric domains are inherited as Mendelian traits but their position can slide in one generation. Conversely, centromere location is stable during mitotic propagation of cultured cells. Our results demonstrate that the presence of more than half centromeres devoid of satellite DNA is compatible with genome stability and species survival. The presence of amplified DNA at some centromeres suggests that these arrays may represent an intermediate stage towards satellite DNA formation during evolution. Figure 6 of the paper shows a possible model for maturation of centromeres during evolution. The observation of the presence of sequence amplification in five donkey centromeres may represent an intermediate stage toward satellite DNA. According to the model, the presence of amplified sequences at a neocentromere is an indication of its more mature stage compared to non-amplified centromeres although it remains to be demonstrated whether amplification is a necessary step towards centromeric satellite DNA formation.

PART 2 COMPARATIVE ANALYSIS OF CENP-A BINDING DOMAINS IN 7 EQUID SPECIES

RESULTS

At the conclusion of the study reported in part 1 (Nergadze SG et al. 2018), we investigated the satellite-free centromeres of *Equus asinus* and identified 16 centromeres lacking satellite DNA in the donkey. We also proved that the location of CENP-A binding domains can vary in different individuals giving rise to epialleles. However, we did not characterize the situation for other equid species. Previous studies in this laboratory (Piras MF et al 2010) already demonstrated that satellite-less centromeres were found also in Grevy's zebra and Burchell's zebra. For this reason we carried out ChIP-seq experiments with the anti-CENP-A antibody in 5 other equid species: *E. zebra hartmannae, E. grevyi, E. burchelli, E. kiang* and *E. hemionus onager*. We then compared the position and sequence of all satellite-less centromeres in the different species. The previously analyzed ChIP-seq datasets from *E. caballus* and *E. asinus* (Nergadze SG et al. 2018) were used as part of this comparative analysis. The molecular characterization of several satellite-less centromeres could help us to unravel when, and possibly how, centromere repositioning occurred throughout evolution. Furthermore, the analysis of the centromeric loci could suggest how such important events (seeding and formation of satellite-less centromeres) may have influenced speciation in the genus.

We performed a chromatin immunoprecipitation on these other equid species using anti-CENP-A antibodies, which target the histone H3 variant at centromeres on chromatin. Purified DNA from each sample (immunoprecipitated DNA and control DNA) was sent to the IGA Technologies Service to perform library preparation and sequencing through Illumina HiSeq2500 platform. Raw read datasets were collected and processed through a bioinformatic pipeline to check reads quality and to map them to the horse reference genome. We used the horse genome as reference (Equcab3) mainly for two reasons:

- Complete reference genomes, assembled at chromosome level, of the equid species, other than *E. caballus* are not available yet
- •Comparative analysis of satellite-less centromeres of equids is done in respect to the horse, which is considered the species with the karyotype configuration closest to the perissodactyla among the equid species here studied

The karyotypes of the extant *Equus* species are characterized by the presence of a variable number of meta- and submetacentric chromosomes derived from fusions between ancestral acrocentric elements (Trifonov VA et al. 2008). Their

chromosome number goes from 32 in *E. zebra* to 66 in *E. przewalskii*. Following the reconstruction of a perissodactyl ancestral karyotype (Trifonov VA et al. 2008), its analysis suggested that the ancestral karyotype consisted of acrocentric chromosomes (2n=74), which underwent fusions during evolution, reducing the chromosome number. For this reason, the horse karyotype (2n=64) was considered the closest to the ancestral configuration.
After having mapped reads from the ChIP and input datasets obtained from 7 equids, we identified and localized (as described in Material and Methods) on the horse genome, satellite-less centromeres in each one of the species as reported in figures R1-1 through R1-28.
By comparing the ChIP dataset and the input dataset we were able to identify an exceptionally high number of satellite-less centromeres in all the analyzed species.

We detected, though the ChiP-seq approach a total of 82 satellite-less centromeres assigned as follows:

- 1 in *Equus caballus* (Wade et al. 2009; Nergadze et al. 2018)
- 16 in *Equus asinus* (Nergadze et al. 2018)
- 10 in *Equus zebra hartmannae*
- 11 in *Equus grevyi*
- 14 in *Equus burchelli*
- 15 in *Equus kiang*
- 15 in *Equus hemionus onager*

With such large number of centromeres devoid of tandemly repeated DNA sequences, a molecular detailed dissection of the centromeric loci is possible.
In the PhD thesis of Francesco Gozzo (2018), sequences of 44 neocentromeres from 4 equid species (Hartmann's zebra, Grevy's zebra, Burchell's zebra and kiang) were assembled and sequence rearrangement analysis was performed. That analysis was done using Equcab2 as reference sequence. In my thesis, reads from all the species were remapped on Equcab3 reference sequence and onager was added to this analysis. Several neocentromeres which were not detected in the previous analysis were identified following the remapping process (EZH14, EBU18, EGR8, EGR11, EBU9, EBU17).
In this thesis we report some features already recognized in the donkey satellite-less centromeres (Nergadze et al. 2018), specifically sequence rearrangements or amplification, epiallelism and repositioning respect to the horse centromeres.
Centromeric loci were identified by comparing the ChIP dataset to the control; peaks represent the read distribution of the ChIP datasets across the reference genome.
Satellite-less centromeres cover a wide spectrum of peak shapes, but they are mainly divided in three groups: Gaussian-like peaks, irregular peaks and spike-like peaks (as shown in Figure 1 of the attached paper).
From the donkey study (Nergadze SG et al 2018) we now know the biological reason behind these differences in terms of reads distribution and consequently, of

peak shape. Gaussian-like peak is the typical distribution of reads at the satellite-less centromeres in which the beneath sequence did not exhibit major rearrangements compared to the horse orthologues non centromeric regions (e.g. Hartmann's zebra centromere mapped on ECA4 showed in Figure R2-2A). Irregular peaks are a consequence of sequence rearrangements compared to the horse reference genome (e.g. Donkey centromere mapped on ECA11 showed in Figure R2-10A). Spike-like peaks may reflect a sequence amplification of the centromeric domain (e.g. Burchell's zebra centromere mapped on ECA15 in Figure R2-14A) as previously demonstrated in the donkey (Nergadze SG et al 2018).

Another feature we identified throughout the centromeric domain analysis, as previously seen in the donkey and in the horse, is the presence of epialleles, different CENP-A binding domains in the two homologs [e.g. Burchell's zebra centromere in Figure R2-5A]. Read mapping was performed on all ChIP and input datasets, using the alignment software Bowtie2.0 (Langmead B and Salzberg SL 2012). We then executed a bioinformatic pipeline previously used in our work (Nergadze et al. 2018) and described in Material and Method section. We loaded the read coverage file (bigwig) obtained by converting the BAM file of each of the six equids (zebras and asses) on the EquCab3 reference genome on the IGV software (Robinson et al. 2011).

We then compared the presence and position of neocentromeres in the different equids trying to infer when neocentromeres formation occurred during equid speciation. To this purpose we created three different phylogenetic tree models [Figure R2-29-31] and positioned the putative events of neocentromere formation on the appropriate nodes of each model. It is important to point out that this analysis is based only on ChIP-seq data without taking into account chromosome rearrangements that may have occurred during the evolution of the different lineages. A detailed comparison of ChIP-seq and cytogenetic data is under way in the laboratory.

The following Figures (R1-1 through R1-28) show the position of the CENP-A binding sites for each of the different equid species in context of the homologous region on the horse genome assembly. The top line in panel A identifies the horse chromosome region which is homologous to the centromeric region in the other species based on DNA sequence. Panel A also shows the shape of the CENP-A binding peaks for each of the species obtained from the ChIP-seq read mapping. Panel B includes a blue bar at the bottom which represents the chromosome for the domestic horse and identifies its length and the relative position of the horse centromere. The positions of the neocentromeres of the other species are identified with colored shapes and with numbers designated the position with respect to the sequences in Equcab3.0

Mapping of CENP-A binding domains on the orthologs of horse chromosome 1

In the orthologs of horse chromosome 1 [Figure R2-1A] only one species shows the presence of CENP-A binding domain: *E. kiang.*

Kiang satellite-less centromere localizes almost 14 Mb away [Figure R2-1B] from the horse satellite-based centromere, and shows a Gaussian-like peak shape, suggesting no sequence rearrangements.

The presence of a satellite-less centromere in the donkey ortholog only suggests that this Centromere Repositioning event occurred in the kiang lineage after its separation from the other equids.

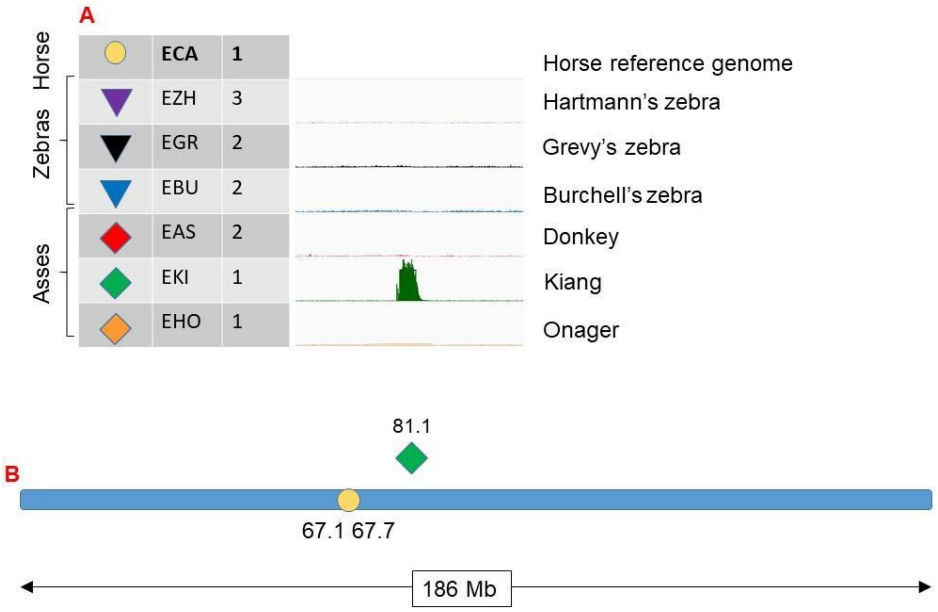


Figure R2-1: Comparative analysis of neocentromeres in the orthologs of horse chromosome. *A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 1. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

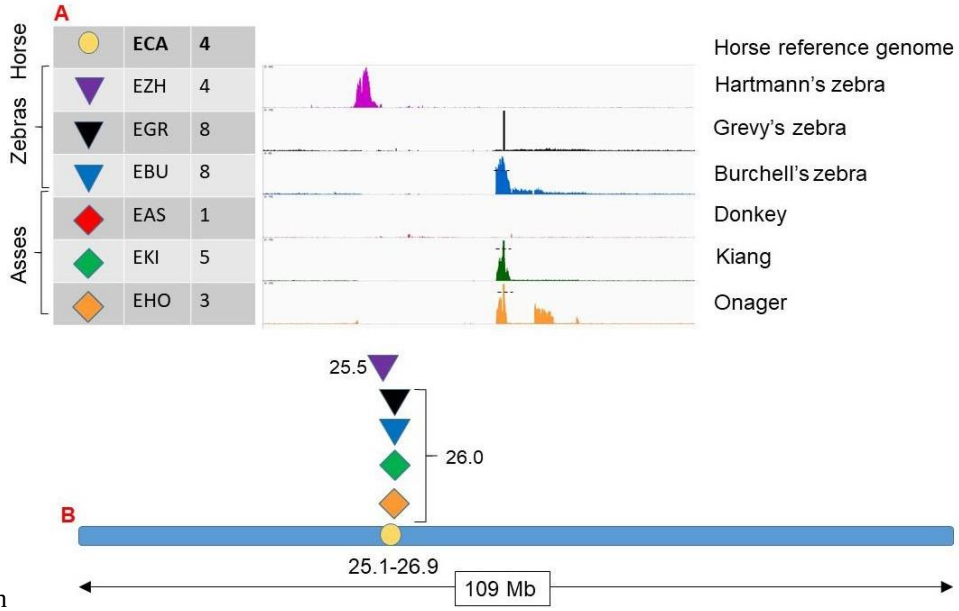
Mapping of CENP-A binding domains on the orthologs of horse chromosome 4

In the orthologs of horse chromosome 4 [Figure R2-2A] four species show the presence of CENP-A binding domains on satellite-less sequences: *E. zebra hartmannae, E. burchelli, E. kiang* and *E. hemionus onager*.

The satellite-less centromeres of Burchell's zebra, kiang and onager are located on the same region relative to the horse reference genome [Figure R2-2B], adjacent to the genomic region of the horse satellite-based centromere. The satellite-less centromere of Hartmann's zebra is about 500 kb away [Figure R2-2B] from the other neocentromeres. In Figure R2-2B a sketch of horse chromosome 4 is shown and different symbols are used to indicate the position, on the horse reference sequence, of the satellite-less centromeres in the different species.

In this chromosome, CENP-A binding domains of Hartmann's zebra and Kiang have a Gaussian-like shape, reflecting little to no sequence rearrangements, while in Burchell's zebra the CENP-A peak displays a spike-like shape suggesting that sequence amplification is present at this locus compared to the horse reference sequence, as previously demonstrated in the donkey (Nergadze SG et al 2018). The peak shape in Onager shows a gap possibly reflecting a sequence deletion in respect to the horse genome.

In Grevy's zebra a narrow spike-like peak was observed (7kb in length). This peak is present also in the input sample (data not shown) but shows a 10-fold enrichment in the CENP-A ChIP. Based on our previous observations from the donkey (Nergadze SG et al 2018), this peak possibly corresponds to a neocentromere located on a 7kb sequence that is amplified only in the Grevy's zebra genome.



hh

***Figure R2-2**: **Comparative analysis of neocentromeres in the orthologs of horse chromosome 4.** A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 4. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 5

In the orthologs of horse chromosome 5 [Figure R2-3A] only two species show the presence of CENP-A binding domains: *E. burchelli* and *E. asinus.*

The Burchell's zebra neocentromere colocalizes with the donkey centromere almost 28 Mb away [Figure R2-3B] from the horse satellite-based centromere. The CENP-A binding peak shows a spike-like shape, suggesting sequence amplification in respect to the horse orthologous sequence. The donkey centromere localizes almost 28 Mb away [Figure R2-3B] from the horse centromere. As previously reported (Nergadze et al. 2018) and discussed in Part 1 of this thesis, this centromere displays a sequence amplification in respect to the horse orthologous sequence although CENP-A partially binds a non-repetitive sequence flanking the 3' end of the amplified region. The amplified sequence structure of this centromere was confirmed by NGS and specific experiments (Nergadze et al. 2018). The spike like shape of this peak is due to the fact that it is amplified in the donkey and is single copy in the horse.

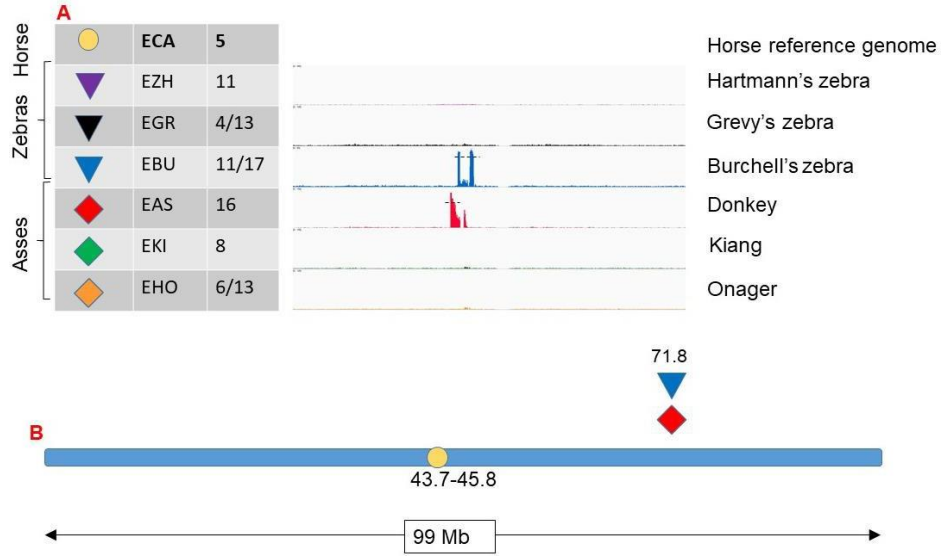


***Figure R2-3*: Comparative analysis of neocentromeres in the orthologs of horse chromosome 5.** *A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 5. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 6

In the orthologs of horse chromosome 6 [Figure R2-4A] only one species shows the presence of CENP-A binding domain: *E. asinus.*

It localizes almost 10 Mb away [Figure R2-4B] from the horse centromere. As previously reported (Nergadze et al. 2018) and discussed in Part 1 of this thesis, this centromere displays a sequence amplification in respect to the horse orthologous sequence. This was confirmed by NGS and specific experiments. The presence of a satellite-less centromere in the donkey ortholog only suggests that this centromere repositioning occurred in the donkey lineage after its separation from the other asses.

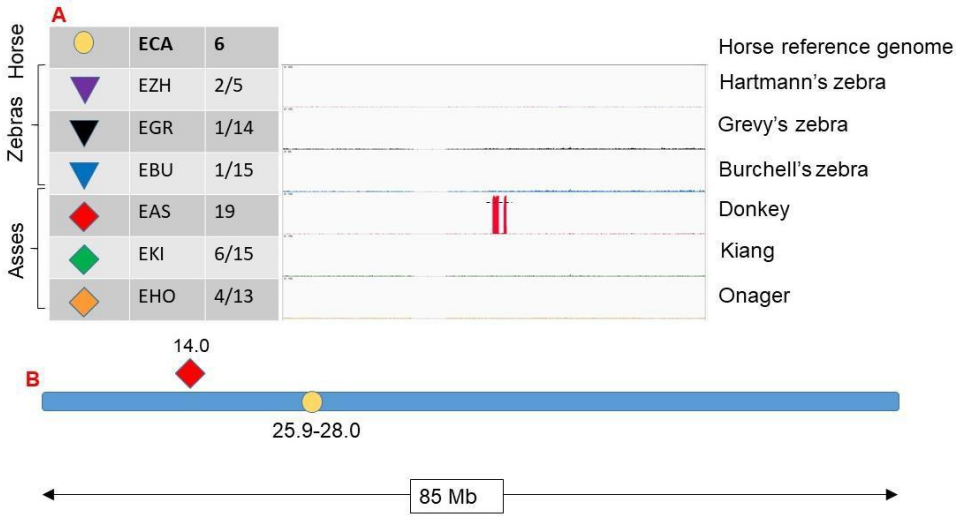


***Figure R2-4**: **Comparative analysis of neocentromeres in the orthologs of horse chromosome 6.** A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 6. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 7

In the orthologs of horse chromosome 7 [Figure R2-5A] three species show the presence of CENP-A binding domain: *E. grevyi, E. burchelli*, and *E. hemionus onager*.

The satellite-less centromere of Burchell's zebra and one of the domains of Grevy's zebra are located in the same position relative to the horse reference sequence. A second peak of Grevy's zebra is located in the same position as the peak of Hartmann's zebra, while in onager a satellite-less centromere is positioned 700 kb away [Figure R2-5B].

Hartmann's zebra and onager neocentromeres, and one of the peaks of Grevy's zebra are located relatively near the horse satellite-based centromere.

Burchell's zebra centromere and the second Grevy's zebra peak are located almost 8 Mb apart [Figure R2-5B] from the horse satellite-based centromere. Burchell's zebra neocentromere shows two different CENP-A binding domains with a Gaussian like peak shape, indicative of epiallelism. The localization of the two peaks of the two CENP-A binding domains in Grevy's zebra is peculiar: the two different peaks are located 8 Mb from each other [Figure R2-5B]; this particular localization of the peaks may be due to sequence rearrangements in Grevy's zebra compared to the horse reference genome.

In Hartmann's zebra a narrow spike-like peak was observed (23kb in length). This peak is present also in the input sample (data not shown) but shows a 10-fold enrichment in the CENP-A ChIP. According to our previous data from the donkey (Nergadze SG et al 2018), this peak probably corresponds to a neocentromere located on a 23kb sequence that is amplified only in the Hartmann's zebra genome. This situation is similar to the one described above for Grevy's zebra neocentromere in the ortholog of horse chromosome 4 (Figure R2-2A).

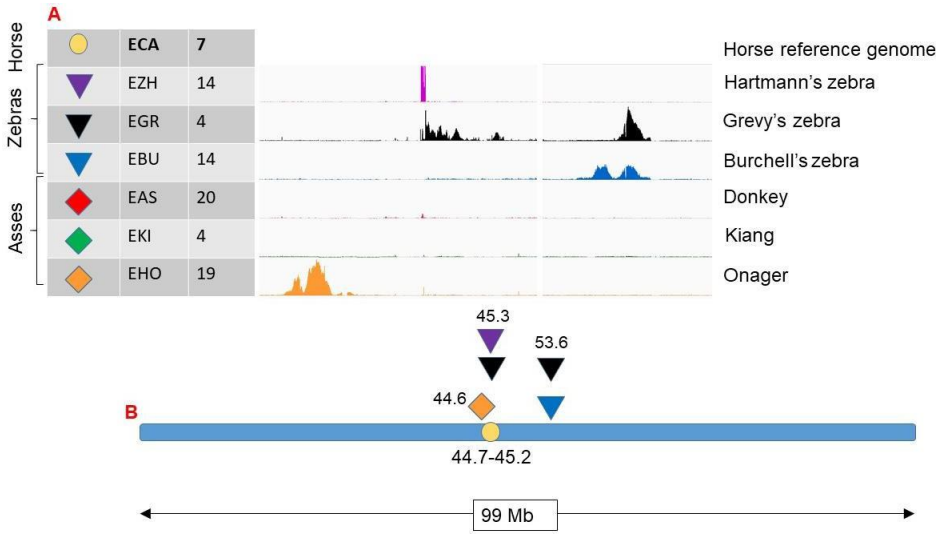


***Figure R2-5*: Comparative analysis of neocentromeres in the orthologs of horse chromosome 7.** *A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 7. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 8

In the orthologs of horse chromosome 8 [Figure R2-6A] only one species shows the presence of CENP-A binding domain: *E. asinus.*

It is localized 12 Mb away [Figure R2-6B] from the horse centromere. This centromere has a Gaussian-like peak shape, indicating no rearrangements between the horse and the donkey sequence (Nergadze et al. 2018).

The presence of a satellite-less centromere in the donkey ortholog only suggests that this centromere repositioning occurred in the donkey lineage only.

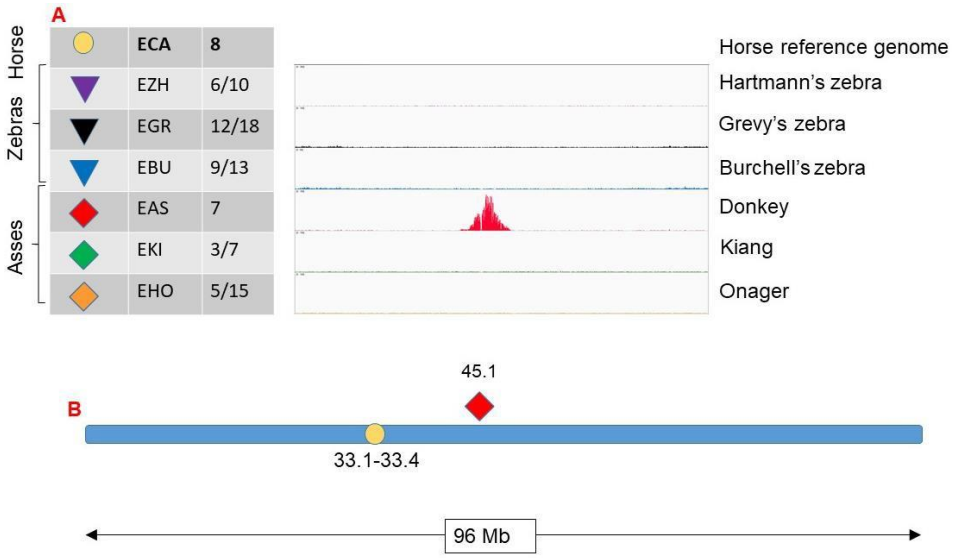


***Figure R2-6**: **Comparative analysis of neocentromeres in the orthologs of horse chromosome 8.** A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 8. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 9

In the orthologs of horse chromosome 9 [Figure R2-7A] six species show the presence of CENP-A binding domain: *E. zebra hartmannae*, *E. grevyi*, *E. burchelli*, *E. asinus*, *E. kiang* and *E. hemionus onager*.

All satellite-less centromeres are in a different position respect to the horse satellite-based centromere. In Hartmann's zebra two CENP-A binding domains were observed; in the horse reference genome, one of these domains is located near the satellite-based centromere in the p-arm, the other domain in the q-arm. Therefore, the horse centromere lays in between these two peaks. Consequently, the distance between the two peaks may be due to the fact that the horse centromere is not assembled and that this region in Hartmann's zebra underwent extensive rearrangement. Grevy's and Burchell's zebra have their CENP-A binding domains located almost 2 Mb away [Figure R2-7B] from the horse centromere on the p-arm of the horse reference genome. Donkey centromere is located 4 Mb away [Figure R2-7B] from horse centromere in the p-arm. The two Gaussian-like peaks indicate no sequence rearrangements and epiallelism.

Kiang and onager centromeres are located 1 Mb away [Figure R2-7B] from the horse centromere in the q-arm. In both cases the peak shape suggests some sequence rearrangement (as shown previously in our laboratory for kiang centromere in Francesco Gozzo PhD thesis 2018) and possibly epiallelism.

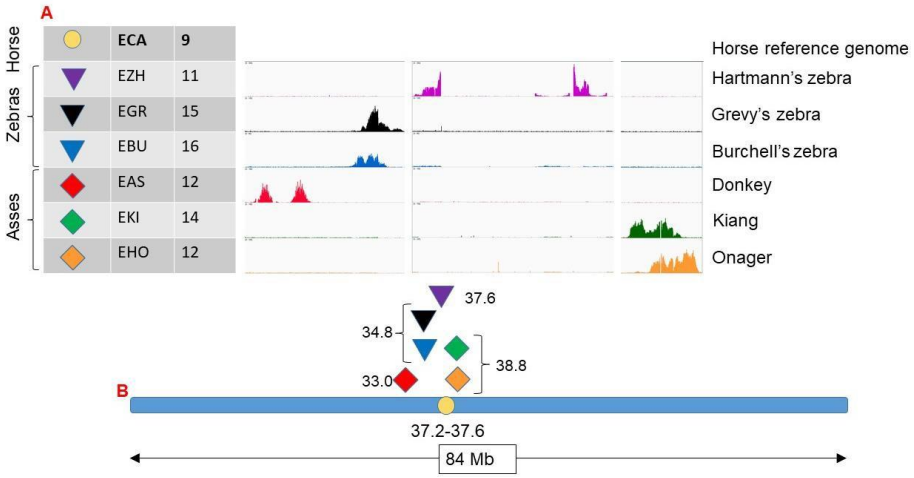


***Figure R2-7*: Comparative analysis of neocentromeres in the orthologs of horse chromosome 9.** *A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 9. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 10p

For horse chromosome 10 the analysis is presented separately for the p and for the q-arm since they have a different evolutionary history (Musilova P et al. 2013) and both bear satellite-less centromeres in different species. In the orthologs of horse chromosome 10p [Figure R2-8A] two species show the presence of CENP-A binding domain: *E. kiang* and *E. hemionus onager*.

These satellite-less centromeres are about 17 Mb away [Figure R2-8B] from the horse satellite-based centromere.

The presence of two domains in both species suggests epiallelism.

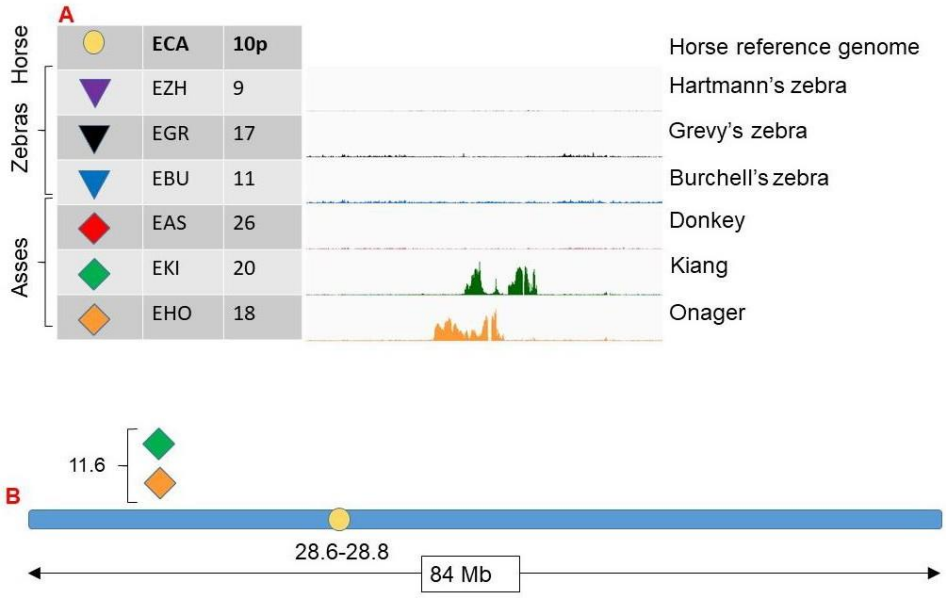


***Figure R2-8**: **Comparative analysis of neocentromeres in the orthologs of horse chromosome 10.** A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 10p. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 10q

In the orthologs of horse chromosome 10q [Figure R2-9A] four species show the presence of CENP-A binding domain: *E. zebra hartmannae*, *E. grevyi*, *E. burchelli* and *E. kiang.*

All satellite-less centromeres are in a different position respect to the horse satellite-based centromere. Hartmann's zebra centromere has a CENP-A binding domain with a regular peak shape located about 3 Mb away [Figure R2-9B] from the horse centromere.

Grevy's and Burchell's zebra have their CENP-A binding domains co-localizing almost 5 Mb away from the horse centromere [Figure R2-9B]. The peak shape of both neocentromeres suggests no sequence rearrangements and the presence of two peaks in Burchell's zebra indicates epiallelism.

Kiang chromosome 11 has a neocentromere located about 1 Mb away [Figure R2-9B] from the horse centromere, and has a peak shape suggesting the presence of two epialleles partially overlapping.

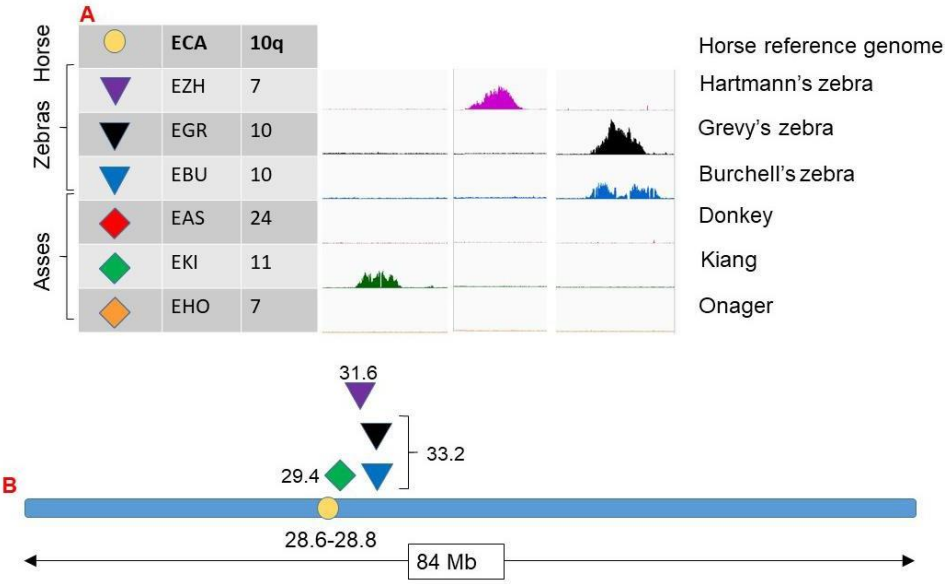


***Figure R2-9**: Comparative analysis of neocentromeres in the orthologs of horse chromosome 10q. A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 10q. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 11

In the orthologs of horse chromosome 11 [Figure R2-10A] only three species show the presence of CENP-A binding domain: *E. caballus*, *E. asinus* and *E. hemionus onager*.

In this chromosome, the horse has a satellite less centromere as previously reported in Part 1 and in previous work (Wade CM et al. 2009; Nergadze et al. 2018).

Neocentromeres of donkey and onager both colocalize in the same genomic region, which is 20 Mb away [Figure R2-10B] from the horse centromere.

The CENP-A binding domain in the donkey has a peculiar irregular shape that is due to sequence rearrangements in the donkey compared to the horse reference sequence as previously described in Part 1 and in the attached paper (Nergadze et al. 2018).

The onager centromere is represented as a Gaussian-like peak, suggesting no sequence rearrangements.

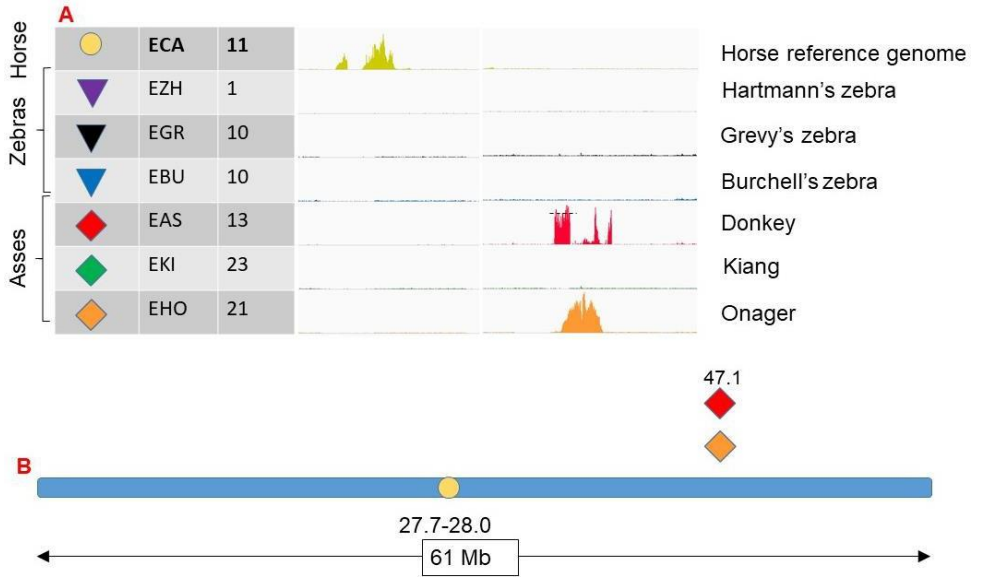


***Figure R2-10**: **Comparative analysis of neocentromeres in the orthologs of horse chromosome 11.** A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 11. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 12

In the orthologs of horse chromosome 12 [Figure R2-11A] only one species shows the presence of CENP-A binding domain: *E. burchelli.*

It localizes 2 Mb away from the horse centromere [Figure R2-11B]. This centromere has two Gaussian-like peaks, highlighting no sequence rearrangements compared to the horse sequence. Moreover, the two peaks are indicative of the presence of epialleles.

The presence of a satellite-less centromere in the Burchell's zebra ortholog only suggests that this Centromere Repositioning event occurred in this zebra lineage after the separation from the other zebras.

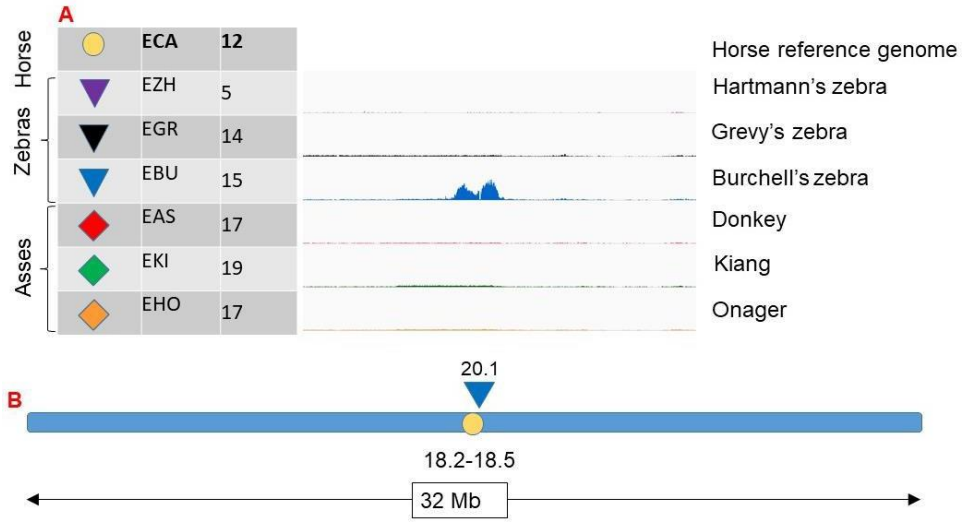


Figure R2-11: Comparative analysis of neocentromeres in the orthologs of horse chromosome 12. *A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 12. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 13

In the orthologs of horse chromosome 13 [Figure R2-12A] four species show the presence of CENP-A binding domain: *E. grevyi*, *E. asinus*, *E. kiang* and *E. hemionus onager*.

All satellite-less centromeres detected colocalize in the same genomic region which is 5 Mb away [Figure R2-12B] from the horse centromere.

Grevy's zebra centromere has an irregular CENP-A binding domain due to sequence rearrangements as shown previously in our laboratory (Francesco Gozzo PhD thesis 2018). Donkey centromere has two peaks with minor sequence deletions and displays epiallelism.

Kiang centromere has a regular Gaussian-like peak shape indicative of no sequence rearrangements while onager centromere has a more irregular peak shape suggesting some sequence rearrangements. Centromere repositioning analysis based only on ChIP-seq data of this chromosome, inferred over three different equid phylogenetic trees, is reported below in Figures R2-29-31. A detailed analysis to describe the formation and evolution of these neocentromeres, taking into account cytogenetic and ChIP-seq data, as shown in Figure D2-1 for ECA4, is underway.

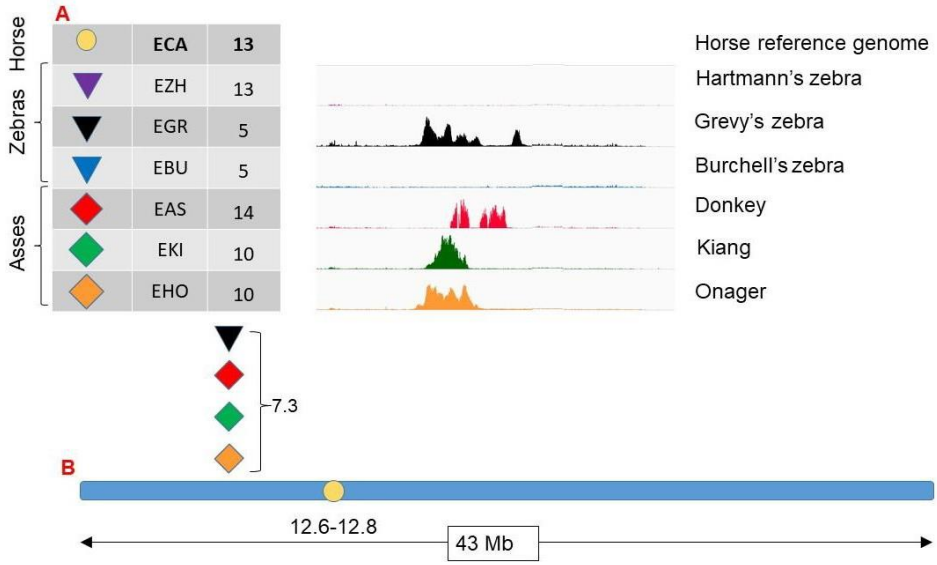


***Figure R2-12**: **Comparative analysis of neocentromeres in the orthologs of horse chromosome 13.** A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 13. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 14

In the orthologs of horse chromosome 14 [Figure R2-13A] only two species show the presence of CENP-A binding domain: *E. asinus* and *E. kiang.*

Donkey centromere is localized about 29 Mb away [Figure R2-13B] from the horse centromere. The presence of the spike-like shaped peak indicates sequence amplification as previously reported (Nergadze et al. 2018).

Kiang centromere is 1 Mb away [Figure R2-13B] from the horse centromere. It has a Gaussian-like peak shape with no major sequence rearrangements.

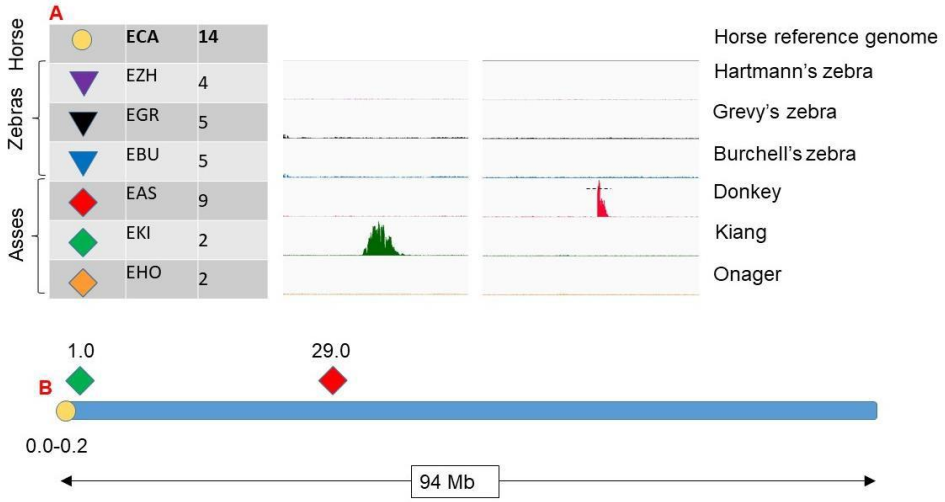


***Figure R2-13**: **Comparative analysis of neocentromeres in the orthologs of horse chromosome 14.** A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 14. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 15

In the orthologs of horse chromosome 15 [Figure R2-14A] only two species show the presence of CENP-A binding domain: *E. burchelli* and *E. hemionus onager*.
Burchell's zebra centromere is localized about 2 Mb away [Figure R2-14B] from the horse centromere. The presence of the spike-like shaped peak indicates partially sequence amplification at the 3' end of this centromere as shown previously in our laboratory (Francesco Gozzo PhD thesis 2018).
Onager centromere is 25 Mb away [Figure R2-14B] from the horse centromere. Two Gaussian-like peaks, located about 600 kb apart from each other, are indicative of epiallelism.

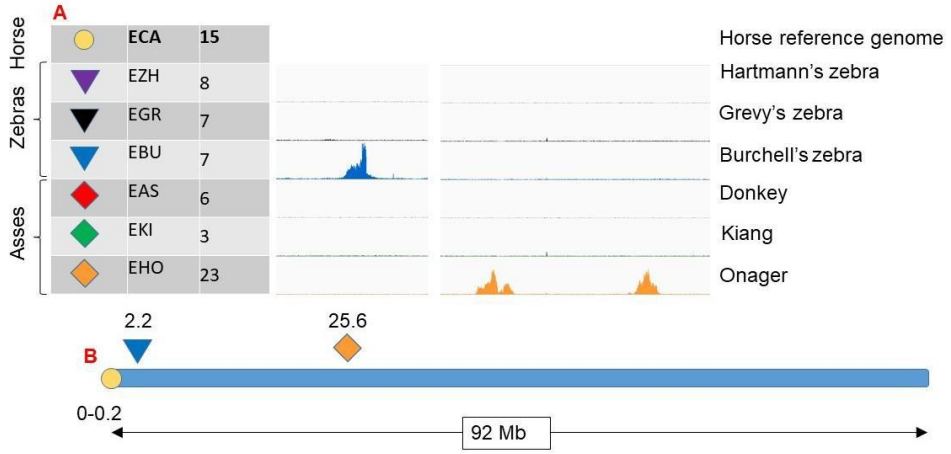


***Figure R2-14**: **Comparative analysis of neocentromeres in the orthologs of horse chromosome 15.** A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 15. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 16

In the orthologs of horse chromosome 16 [Figure R2-15A] only two species show the presence of CENP-A binding domain: *E. kiang* and *E. hemionus onager*.
Both neocentromeres colocalize in the same genomic region, which is 51 Mb away [Figure R2-15B] from the horse centromere.
Kiang centromere in this chromosome displays a CENP-A binding domain of two peaks partially overlapping; no major sequence rearrangements are present.
Onager centromere, although being present in the same genomic region of kiang, displays two different CENP-A binding domains positioned 400 kb apart from each other [Figure R2-15B] in the horse genome. We do not know whether this large distance between the epialleles is related to centromere sliding only or to some rearrangement that occurred in the onager genome.

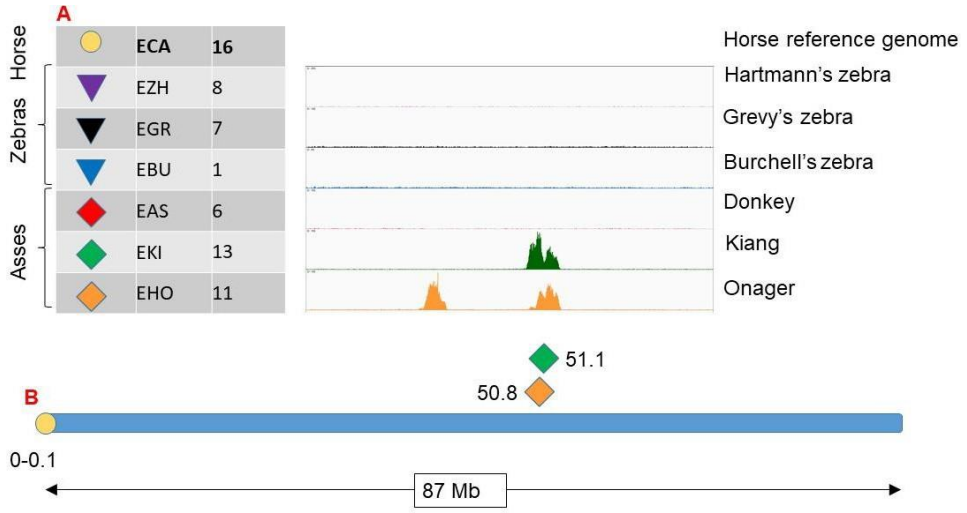


***Figure R2-15**: **Comparative analysis of neocentromeres in the orthologs of horse chromosome 16.** A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 16. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 17

In the orthologs of horse chromosome 17 [Figure R2-16A] only one species shows the presence of CENP-A binding domain: *E. asinus.*

It localizes almost 18 Mb away [Figure R2-16B] from the horse centromere. This centromere has the peculiar Gaussian-like peak shape, highlighting (Nergadze et al. 2018) little to no rearrangements between the horse and the donkey sequence.

The presence of a satellite-less centromere in the donkey ortholog only suggests that this centromere repositioning event occurred in the donkey lineage after the separation from all other equids.

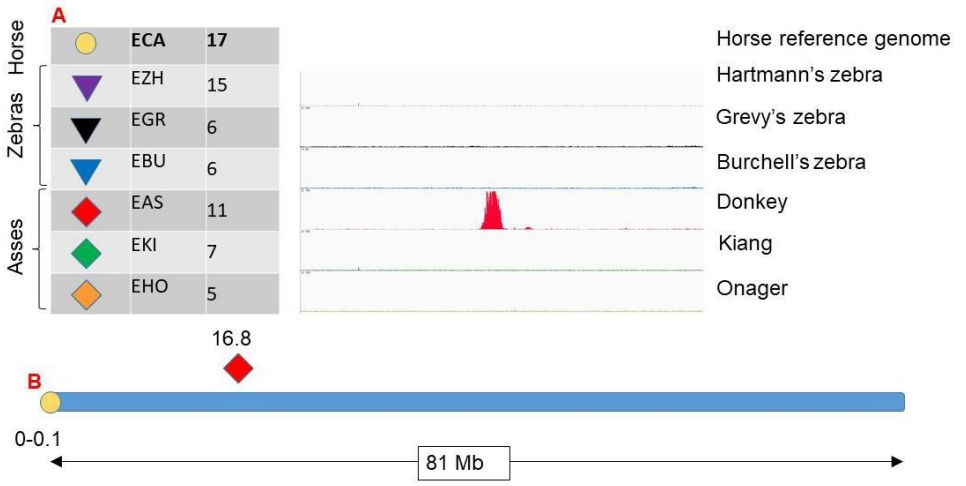


***Figure R2-16**: **Comparative analysis of neocentromeres in the orthologs of horse chromosome 17.** A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 17. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 18

In the orthologs of horse chromosome 18 [Figure R2-17A] four species show the presence of CENP-A binding domain: *E. zebra hartmannae*, *E. grevyi*, *E. burchelli* and *E. hemionus onager*.

All satellite-less centromeres are in a different position respect to the horse satellite-based centromere. Hartmann's zebra centromere shows a CENP-A binding domain with a regular peak shape, thus displaying no sequence rearrangement, and is distant about 1.5 Mb [Figure R2-17B] from the horse centromere. Grevy's and Burchell's zebra have their CENP-A binding domain located almost 7 Mb away [Figure R2-17B] from the horse centromere and they both show no major sequence rearrangements.

Onager centromere is located 8 Mb away [Figure R2-17B] from the horse centromere. Two CENP-A binding domains are present at this locus, which correspond to two epialleles separated by about 100 kb; no major sequence rearrangements are present.

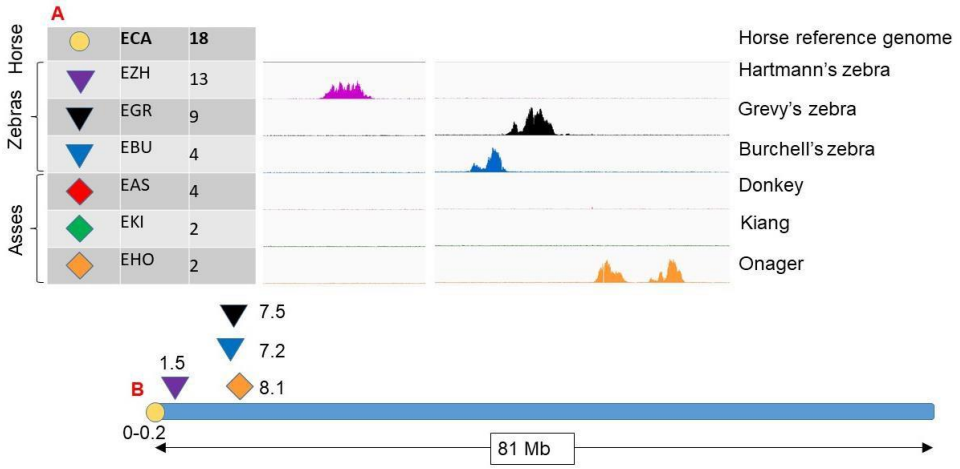


***Figure R2-17**: **Comparative analysis of neocentromeres in the orthologs of horse chromosome 18.** A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 18. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 19

In the orthologs of horse chromosome 19 [Figure R2-18A] only one species shows the presence of CENP-A binding domain: *E. asinus.*

It localizes almost 7 Mb away [Figure R2-18B] from the horse centromere. This centromere shows two Gaussian-like peaks, highlighting (Nergadze et al. 2018) little to no rearrangements between the horse and the donkey sequence and epiallelism.

The presence of a satellite-less centromere in the donkey ortholog only suggests that this Centromere Repositioning event occurred in the donkey lineage after its separation from the other equids.

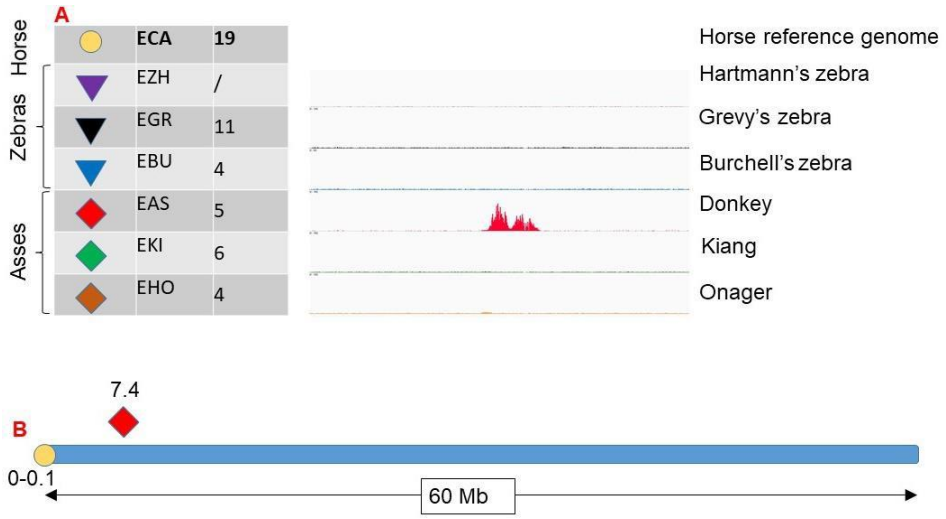


***Figure R2-18**: **Comparative analysis of neocentromeres in the orthologs of horse chromosome 19.** A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 19. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 20

In the orthologs of horse chromosome 20 [Figure R2-19A] four species show the presence of CENP-A binding domain: *E. zebra hartmannae, E. grevyi, E. asinus, E. kiang* and *E. hemionus onager*.

All satellite-less centromeres detected localize in a different genomic region compared to the horse centromere.

Hartmann's zebra centromere localizes almost 0.7 Mb [Figure R2-19B] from the horse centromere and displays a spike-like peak configuration which suggests that sequence amplification occurred in the Hartmann's zebra genome as shown previously in our laboratory (Francesco Gozzo PhD thesis 2018).

Donkey and kiang centromeres and one peak of onager centromere reside in the same genomic region which is 27 Mb away [Figure R2-19B] from the horse centromere. Donkey centromere displays an irregular peak shape and is located into an amplified region which is single copy in the horse as previously demonstrated (Nergadze SG et al 2018).

Kiang centromere, although being present in the same genomic locus, shows a Gaussian-like peak shape, indicative of little to no sequence amplification.

Onager centromere presents a peculiar situation: two Gaussian-like peaks are present. One epiallele is located on the same genomic region of the donkey and kiang centromeres, while the other epiallele is 2 Mb away [Figure R2-19B] from the first one. Being mapped on the horse reference genome and not on the proper kiang reference genome (not available), we do not know whether this distance is related to centromere sliding only or also to sequence rearrangement.

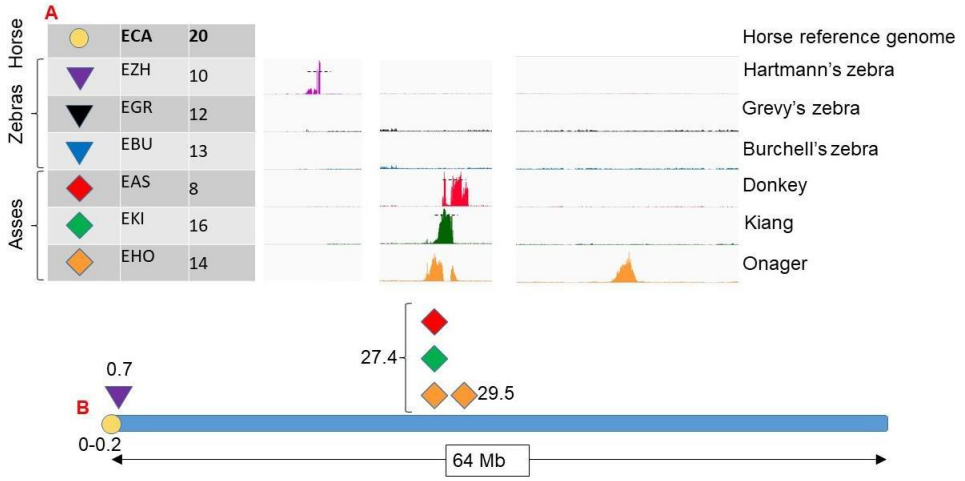


***Figure R2-19**: Comparative analysis of neocentromeres in the orthologs of horse chromosome 20. A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 20. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 21

In the orthologs of horse chromosome 21 [Figure R2-20A] five species show the presence of CENP-A binding domain: *E. zebra hartmannae, E. grevyi, E. burchelli, E. kiang* and *E. hemionus onager*. All satellite-less centromeres are in a different position respect to the horse satellite-based centromere. Hartmann's, Grevy's and Burchell's zebra localize almost 1 Mb away [Figure R2-20B] from the horse centromere. Hartmann's zebra centromere has a CENP-A binding domain with an irregular peak shape, with a gap between the two CENP-A enriched regions, reflecting a large sequence deletion compared to the horse reference genome (Francesco Gozzo PhD thesis). Grevy's zebra centromere displays an irregular peak shape, reflecting some sequence rearrangements, while Burchell's zebra centromere has a Gaussian-like shape, indicative of no sequence rearrangements. Kiang centromere presents a peculiar situation: two Gaussian-like peaks are present. One epiallele is located on the same genomic region of the zebras' centromeres, while the other epiallele is 4 Mb distant [Figure R2-20B] from the first one. Being mapped on the horse reference genome and not on the proper kiang reference genome (not available), we do not know whether this distance is related to centromere sliding only or also to sequence rearrangement. Onager centromere displays two Gaussian-like peaks, two epialleles, and is located 10 Mb away [Figure R2-20B] from the horse centromere. No major sequence rearrangements are present.

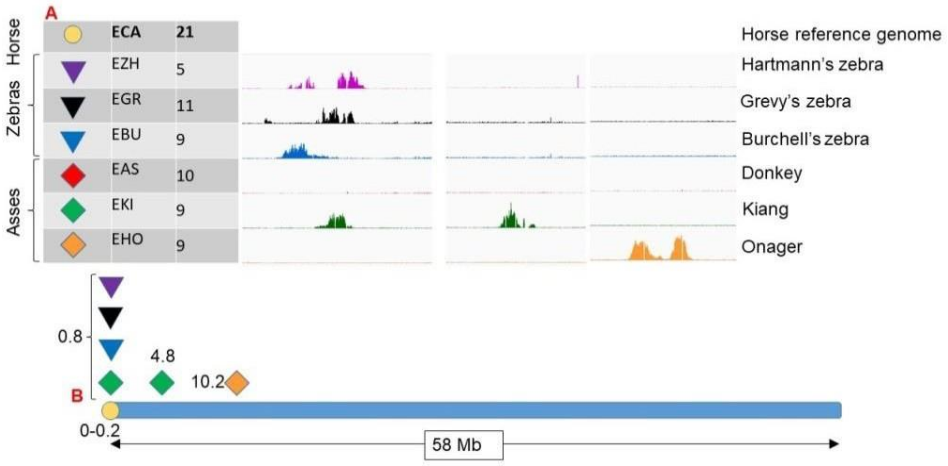


Figure R2-20: Comparative analysis of neocentromeres in the orthologs of horse chromosome 21. *A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 21. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 22

In the orthologs of horse chromosome 22 [Figure R2-21A] three species show the presence of CENP-A binding domain: *E. burchelli*, *E. kiang* and *E. hemionus onager*.

Burchell's zebra centromere is localized almost 12 Mb away [Figure R2-21B] from the horse centromere and displays a Gaussian-like peak which suggests no rearrangements compared to the horse reference sequence.

Kiang centromere is present almost 24 Mb away [Figure R2-21B] from the horse centromere and displays two CENP-A binding domains with an irregular peak shape indicative of epiallelism and sequence rearrangements as shown previously in our laboratory (Francesco Gozzo PhD thesis 2018).

Onager centromere displays two Gaussian-like peaks: one epiallele is located on the same genomic region of the kiang centromere. The other epiallele is 1.5 Mb distant [Figure R2-21B] from the first one. Being mapped on the horse reference genome and not on the proper kiang reference genome (not available), we do not know whether this distance is related to centromere sliding only or also to sequence rearrangement.

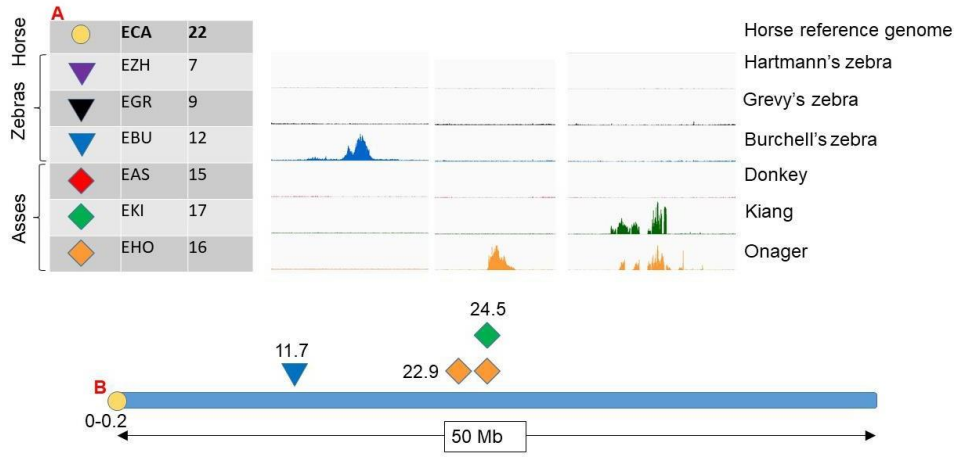


***Figure R2-21**: **Comparative analysis of neocentromeres in the orthologs of horse chromosome 22.** A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 22. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 24

In the orthologs of horse chromosome 24 [Figure R2-22A] only two species show the presence of CENP-A binding domain: *E. grevyi* and *E. burchelli*

Grevy's zebra centromere localizes almost 17 Mb away [Figure R2-22B] from the horse centromere. This centromere has an irregular peak shape, highlighting several sequence rearrangements between Grevy's zebra sequence and the horse sequence as shown previously in our laboratory (Francesco Gozzo PhD thesis 2018).

Burchell's zebra centromere colocalizes with Grevy's zebra centromere. Similarly to donkey chromosome 19 (Nergadze SG et al 2018), the spike-like peak suggests that sequence amplification occurred in the Burchell's zebra lineage after its separation from Grevy's zebra.

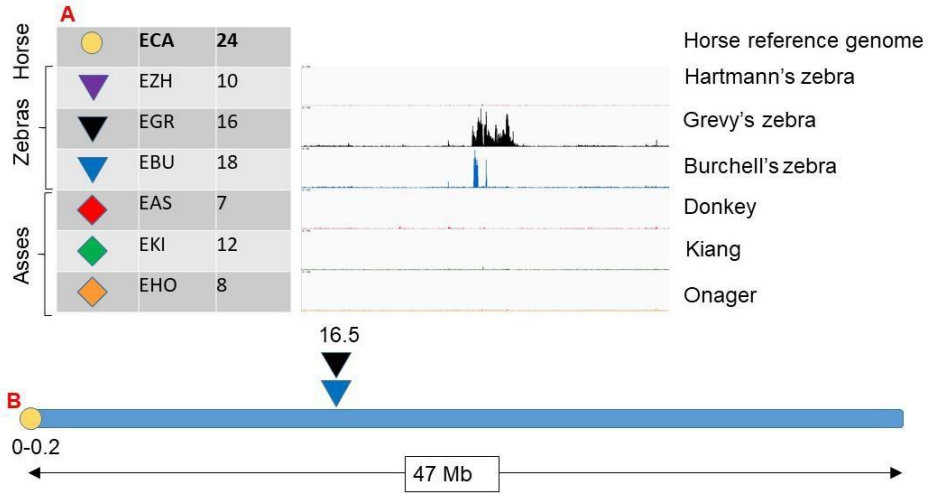


***Figure R2-22**: **Comparative analysis of neocentromeres in the orthologs of horse chromosome 24.** A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 24. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 25

In the orthologs of horse chromosome 25 [Figure R2-23A] only two species show the presence of CENP-A binding domain: *E. zebra hartmannae* and *E. asinus.*
Hartmann's zebra centromere displays two major peaks (epialleles): one peak is present almost 9 Mb away [Figure R2-23B] from the horse centromere, the other one 10 Mb away [Figure R2-23B]; the two epialleles are separated by 1 Mb [Figure R2-23B] from each other, however being mapped on the horse reference genome and not on the proper Hartmann's zebra reference genome (not available), we do not know whether this distance is related to centromere sliding only or also to sequence rearrangement. Both peaks present a Gaussian-like shape, indicative of little to no sequence rearrangements
Donkey centromere is located 9 Mb away [Figure R2-23B] from the horse centromere, co-localizing also with one peak of Hartmann's zebra centromere. It displays two adjacent Gaussian-like peaks, indicative of epiallelism, and no sequence rearrangements between the donkey and the horse genome.

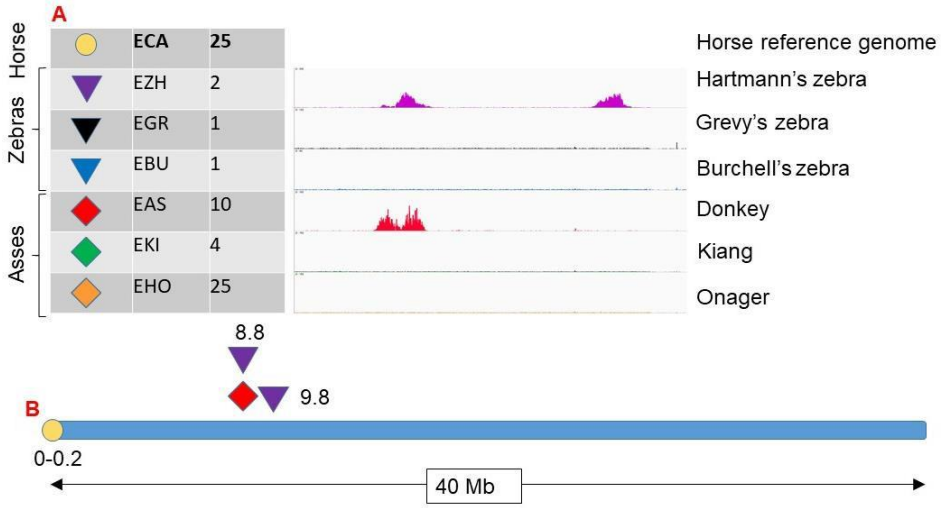


***Figure R2-23**: **Comparative analysis of neocentromeres in the orthologs of horse chromosome 25.** A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 25. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 26

In the orthologs of horse chromosome 26 [Figure R2-24A] five species show the presence of CENP-A binding domain: *E. grevyi, E. burchelli, E. asinus, E. kiang* and *E. hemionus onager*.

All satellite-less centromeres are in a different position respect to the horse satellite-based centromere.

Grevy's zebra centromere is located 6.4 Mb away [Figure R2-24B] from the horse centromere. It comprises two CENP-A binding domains with a Gaussian-like peak shape, suggesting epiallelism.

Burchell's zebra centromere colocalizes with Grevy's zebra centromeric position [Figure R2-24B] but it displays only one epiallele in this individual. Donkey and kiang centromeres colocalize in the same genomic region, 23 Mb away [Figure R2-24B] from the horse centromere. As previously reported (Nergadze et al. 2018), this donkey centromere partially resides on an amplified sequence; donkey peak showed in figure R2-23A is composed by a Gaussian like peak flanked by a spike-like peak, so CENP-A partially binds an amplified region. Kiang centromere displays two Gaussian-like peaks, indicating epiallelism and no major sequence rearrangements. Onager centromere is localized almost 22 Mb away [Figure R2-24B] from the horse centromere showing one Gaussian-like peak, indicative of no major sequence rearrangements.

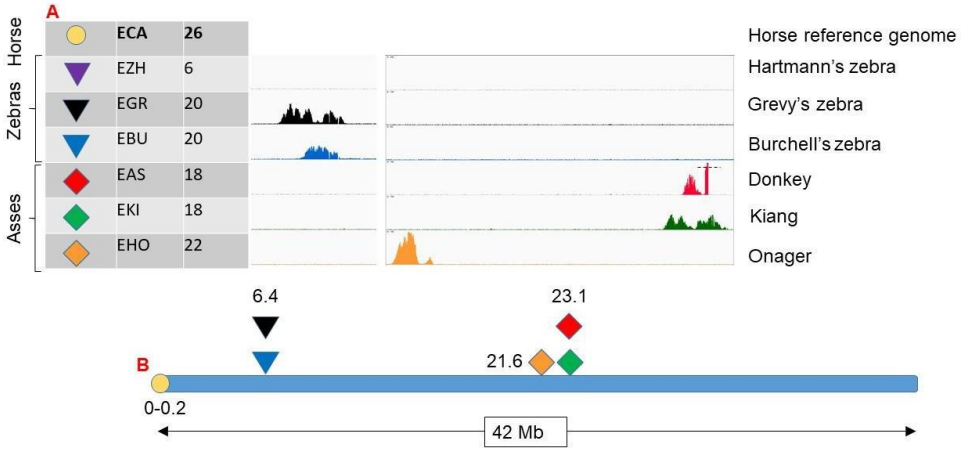


Figure R2-24: Comparative analysis of neocentromeres in the orthologs of horse chromosome 26. *A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 26. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 27

In the orthologs of horse chromosome 27 [Figure R2-25A] four species show the presence of CENP-A binding domain: *E. zebra hartmannae, E. burchelli, E. asinus* and *E. kiang.*

All satellite-less centromeres are in a different position respect to the horse satellite-based centromere.

Hartmann's and Grevy's zebra centromeres are co-localized 1 Mb away [Figure R2-25B] from the horse centromere. Hartmann' zebra centromere comprises one Gaussian-like peak in which a large sequence deletion occurred compared to the horse sequence as shown previously in our laboratory (Francesco Gozzo PhD thesis 2018). Grevy's zebra centromere shows an irregular peak shape indicative of major sequence rearrangements (also reported in Francesco Gozzo PhD thesis).

Donkey centromere is localized in 20 Mb away [Figure R2-25B] from the horse centromere. As previously reported (Nergadze et al. 2018) this centromere shows two Gaussian-like peaks partially overlapping, indicative of epiallelism and no major sequence rearrangements. Kiang centromere localized almost 1 Mb away [Figure R2-25B] from the horse centromere, in the same genomic region of Hartmann's and Burchell's centromeres and displays one Gaussian-like peak; minor to no sequence rearrangements are present in this locus.

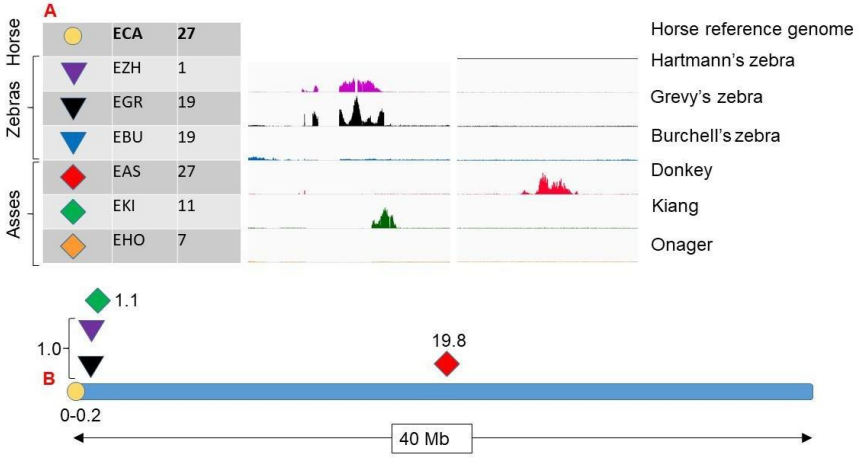


Figure R2-25: Comparative analysis of neocentromeres in the orthologs of horse chromosome 27. *A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 27. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 28

In the orthologs of horse chromosome 28 [Figure R2-26A] only two species show the presence of CENP-A binding domain: *E. burchelli* and *E. asinus.*

Burchell's zebra centromere is located 3 Mb away [Figure R2-26B] from the horse centromere and displays two adjacent Gaussian-like peaks, indicative of epiallelism and no sequence rearrangements.

Donkey centromere localizes 14 Mb away [Figure R2-26B] from the horse centromere and displays one Gaussian-like peak, indicative of no sequence rearrangements.

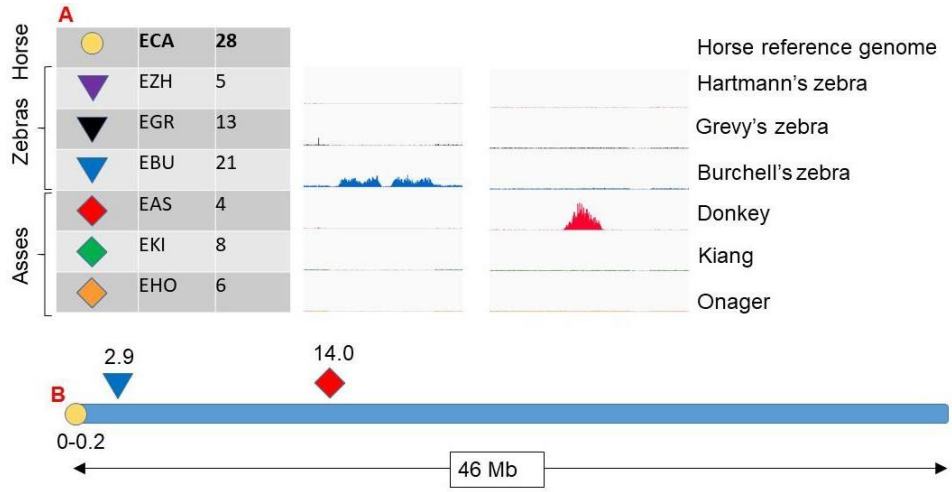


Figure R2-26: Comparative analysis of neocentromeres in the orthologs of horse chromosome 28. *A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 28. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome 30

In the orthologs of horse chromosome 30 [Figure R2-27A] three species show the presence of CENP-A binding domain: *E. asinus, E. kiang* and *E. hemionus onager*. All satellite-less centromeres detected colocalize in the same genomic region which is about 18 Mb away [Figure R2-27B] from the horse centromere.

Donkey centromere has one Gaussian-like peak indicative of no sequence rearrangements.

Kiang and onager centromeres share the same features: two colocalized Gaussian-like peaks, indicative of epiallelism and no major sequence rearrangements.

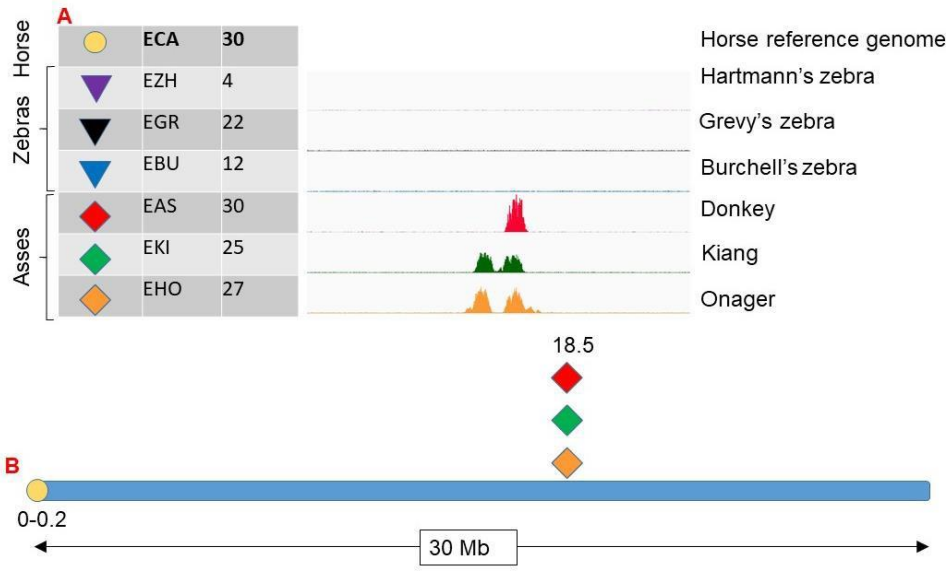


***Figure R2-27**: **Comparative analysis of neocentromeres in the orthologs of horse chromosome 30.** A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome 30. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Mapping of CENP-A binding domains on the orthologs of horse chromosome X

In the orthologs of horse chromosome X [Figure R2-28A] six species show the presence of CENP-A binding domain: *E. zebra hartmannae*, *E. grevyi*, *E. burchelli*, *E. asinus*, *E. kiang* and *E. hemionus onager*. Satellite-less centromeres of Hartmann's, Grevy's, Burchell's zebra, kiang and onager are localized adjacent to the genomic region of the satellite-based centromere of the horse [Figure R2-28B]. Hartmann's zebra centromere displays two peaks. One is irregular, indicating genomic rearrangements as shown previously in our laboratory (Francesco Gozzo PhD thesis 2018) and the other one has a Gaussian-like shape. Grevy's zebra centromere displays one Gaussian-like peak, indicative of no major sequence rearrangements. Burchell's zebra centromere shows three irregular peaks; two localizing in the same genomic region of the other zebras' centromeres [Figure R2-28B], the other one 2 Mb away, suggesting major sequence rearrangements (also reported in Francesco Gozzo PhD thesis). Donkey centromere localizes 20 Mb away [Figure R2-28B] from the horse centromere, thus not colocalizing with the other neocentromeres; its Gaussian-like peak indicates no major sequence rearrangements. Kiang and onager centromeres share the exact genomic location [Figure R2-28B] and peak conformation. Peak shape of both neocentromeres suggests sequence rearrangements respect to the horse genome as shown previously in our laboratory (Francesco Gozzo PhD thesis 2018).

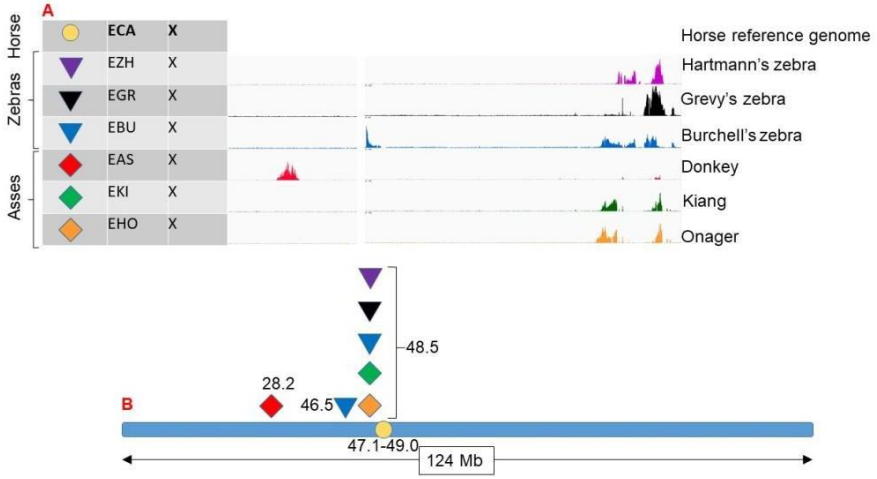


***Figure R2-28**: **Comparative analysis of neocentromeres in the orthologs of horse chromosome X.** A) Table showing colors and symbols used to identify chromosomes in different species compared to the orthologous region on horse chromosome X. The central panel shows CENP-A binding sites obtained by ChIP-Seq using the horse genome (EquCab3) as sequence reference; B) Schematic representation of the horse chromosome 1 (blue bar) with symbols and numbers indicating the position (Mb) of the neocentromeres in the other species with respect to the sequences in Equcab3.0.*

Analysis of putative centromere repositioning events during the evolution of the genus *Equus*

Changes in genome organization, insertions of DNA sequences, deletions, duplications or other types of rearrangements are frequently used as phylogenetic markers. Genomic rearrangements involving chromosomes of a common ancestor are usually conserved and inherited in the orthologous chromosomes of the descendent species. The analysis of the presence/absence of such rearrangements is a potential approach to determine the phylogeny of evolutionarily related species. For instance, a comparative study previously applied to retrotransposons allowed to reconstruct the mammalian evolutionary tree (Usdin K et al. 1995; Kriegs JO et al. 2006). Centromere repositioning has been proposed as useful phylogenetic marker (Rocchi M et al. 2012, Piras MF et al. 2009, Piras MF et al. 2010, Nergadze SG et al. 2018). Here we dated CR events during the evolution of the genus *Equus* considering the sole presence of satellite-less centromeres identified through ChIP-seq.

Figures R2-29-31 show three different putative phylogenetic trees in which the centromeric repositioning events for each of the chromosomes are inferred; based on the presence of ChIP-seq signals and species relationship, we suggest a date for each repositioning event during the genus *Equus* radiation. These three phylogenetic models have been proposed in order to identify the evolutionary relationships giving rise to the lowest possible number of repositioning events compatible with our ChIP-seq data.

It is important to underline that this analysis does not take into account chromosomal rearrangements that may have modified the neocentromere position. Lineage sorting may have also played a role in the inheritance of the neocentromeres in the different equid lineages as described in the discussion.

Orthologs of horse chromosomes 1, 6, 8, 12, 17, 19, are not reported since a neocentromere was observed only in one species suggesting that a repositioning event occurred only in that lineage.
Here I report some examples of the results of the analysis on the first proposed tree. As shown in Figure R2-29, the neocentromeres observed in the orthologs of ECA 9 were probably generated in the common ancestor of all asses and zebras as a single repositioning event. A similar situation can be proposed to explain the origin of the neocentromeres identified on the X chromosomes. These events should have occurred after the separation of the asses and zebras ancestor from the horse lineage.
For the neocentromeres observed on the orthologs of ECA7, the absence of a neocentromere in the donkey and kiang, suggests that two repositioning events may

have occurred. The first one in the common ancestor of zebras and the second one in the onager lineage.
In the orthologs of ECA22 neocentromeres are present in Burchell's zebra, kiang and onager while no neocentromeres were observed in Hartmann's zebra, Grevy's zebra and donkey. These results suggest that two independent repositioning events may have occurred. One in the common ancestor of kiang and onager and the other one only in Burchell's zebra, after its divergence from the other zebras.
The presence of neocentromeres on the orthologs of ECA26 in all species, except for Hartmann's zebra, suggests that two repositioning events may have occurred. One in the common ancestor of the asses and the other one in the common ancestor of Grevy's and Burchell's zebra. An alternative hypothesis is that chromosomal rearrangements took place in the Hartmann's zebra lineage displacing the satellite-less centromere.
Neocentromeres identified in the orthologs of ECA24 in only two species (Grevy's zebra and Burchell's zebra) suggest that only one repositioning event occurred in the common ancestor of the two zebras.
According to this evolutionary tree and taking into account only ChIP-seq data, in the orthologs of ECA27 four repositioning events may have occurred. One in Hartmann's zebra, one in Grevy's zebra, one in the donkey and one in kiang. If this was true, this hypothesis suggests that this genomic region may act as hotspot for the centromere formation. It is important to underline that the evolution of these centromeres is particularly complex because in the two zebras and in kiang the CENP-A binding domains colocalize in the same position while in the donkey the domain lays 19 Mb away [Figure R2-25].
The total number of centromere repositioning events [Table R2-1] according to this model is 40. In many cases more than one repositioning event occurred suggesting that neocentromeres may have been seeded at "centromerization" hotspots.

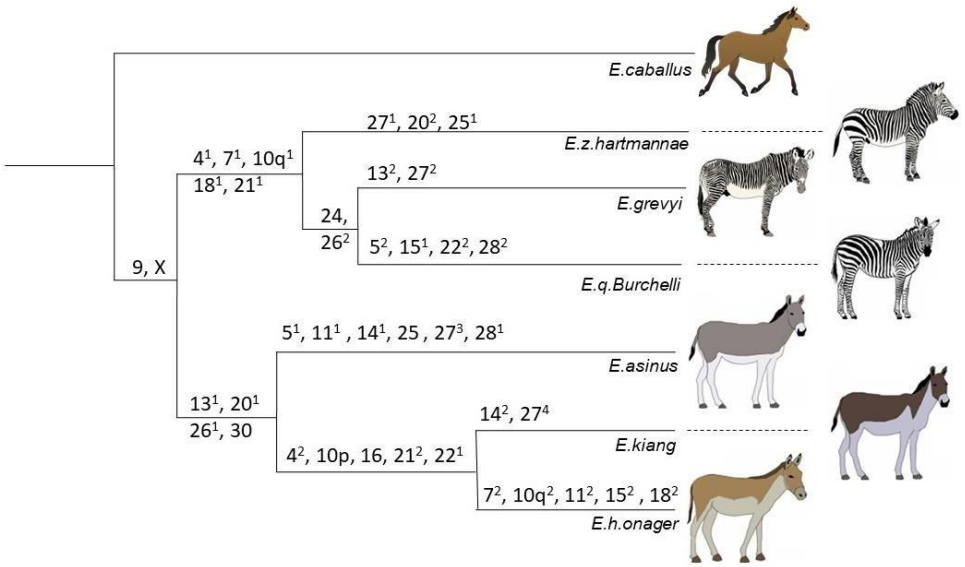


***Figure R2-29**: Analysis of centromere repositioning events, based on the presence of satellite-less centromeres, in the lineages of the classical Equus phylogenetic tree.* *This phylogenetic tree was previously proposed by Jónsson H et al. (2014). E. caballus is considered as the outgroup. Each chromosome in which a satellite-less centromere was mapped is reported over the corresponding evolutionary lineage. Chromosome numbers correspond to those of the horse reference genome. Superscript numbers indicate neocentromeres observed in more than one lineage. According to this phylogenetic tree and taking into account only the results of ChIP-seq analysis, these neocentromeres should derive from multiple repositioning events.*

CHR	N° of CR events	CHR	N° of CR events
ECA4	2	ECA18	2
ECA5	2	ECA20	2
ECA7	2	ECA21	2
ECA9	1	ECA22	2
ECA10q	2	ECA24	1
ECA10p	1	ECA25	2
ECA11	2	ECA26	2
ECA13	2	ECA27	4
ECA14	2	ECA28	2
ECA15	2	ECA30	1
ECA16	1	ECAX	1
TOTAL NUMBER OF CR EVENTS: 40			

Table R2-1: Number of putative centromere repositioning events based only on ChIP-seq data and deduced from the phylogenetic tree in Figure R2-29.

Also, in the second tree [Figure R2-30] proposed for this work, like the first tree [Figure R2-28], the satellite-less centromeres observed in the orthologs of ECA 9 were probably originated in the common ancestor of all asses and zebras as a single repositioning event, early in the equid speciation. Neocentromeres present on the orthologs of ECAX suggest a similar scenario: a single repositioning event that occurred in the common ancestor of asses and zebras.

Here some examples of other repositioning events are reported.

Neocentromeres identified in all the orthologs of ECA4 except for the donkey, suggest that two repositioning events may have occurred. One in the common ancestor of the zebras and asses and the other one in Burchell's zebra.

According to this tree, the presence of satellite-less centromeres only in the donkey and Burchell's zebra orthologs of ECA5, suggests that only one repositioning event occurred in their common ancestor.

Within this model, three repositioning events are necessary to explain the neocentromere formation in the orthologs of some horse chromosomes such as ECA13 and ECA27.

The total number of centromere repositioning events [Table R2-2] according to this model is 45.

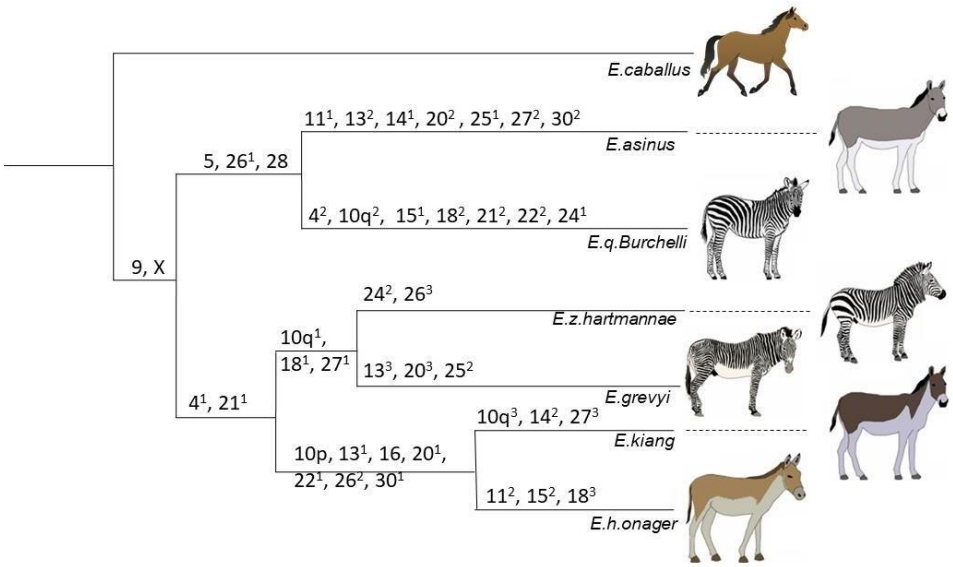


***Figure R2-30**: Analysis of centromere repositioning events, based on the presence of satellite-less centromeres, in the lineages of the Equus phylogenetic tree proposed by Piras MF et al. (2009).* *E. caballus is considered as the outgroup. Each chromosome in which a satellite-less centromere was mapped is reported over the corresponding evolutionary lineage. Chromosome numbers correspond to those of the horse reference genome. Superscript numbers indicate neocentromeres observed in more than one lineage. According to this phylogenetic tree, these neocentromeres should derive from multiple repositioning events.*

CHR	N° of CR events	CHR	N° of CR events
ECA4	2	ECA18	3
ECA5	1	ECA20	3
ECA7	3	ECA21	2
ECA9	1	ECA22	2
ECA10q	3	ECA24	2
ECA10p	1	ECA25	2
ECA11	2	ECA26	3
ECA13	3	ECA27	3
ECA14	2	ECA28	1
ECA15	2	ECA30	2
ECA16	1	ECAX	1
Total number of CR events: 45			

Table R2-2: Number of centromere repositioning events deduced from the phylogenetic tree in Figure R2-30.

Finally, the last model proposed for the repositioning event analysis is reported [Figure R2-31]. Unlike the other two proposed models, in the orthologs of ECA9 and ECAX repositioning events are not single events. Since the donkey, in this tree, separated early in the equid speciation, for these two chromosomes, two repositioning events can be proposed. One in the common ancestor of asses and zebra, and the other one in donkey. On chromosome X, the donkey neocentromere was observed in a different position compared to all other species. This observation supports this phylogenetic tree in which the donkey lineage separated early during the evolution of the genus Equus [Figure R2-28B].

Here some examples of other repositioning events are reported.

Presence of neocentromeres in the orthologs of ECA7 in all the zebras and in kiang, suggests that two repositioning events may have occurred. One in the common ancestor of the zebras and the other one only in the onager lineage.

Satellite-less centromeres identified in the orthologs of ECA22 in kiang, onager and Burchell's zebra, and absence in the other species suggest that two independent repositioning events occurred. One before the divergence of kiang and onager, and the other one after the zebras' divergence and only in the Burchell's zebra lineage.

This model better explains the repositioning event for the orthologs of ECA4 and ECA21 since the absence of the donkey neocentromere in the orthologs of these two chromosomes. In both cases only one repositioning event occurred in the common ancestor of zebras and asses.

The total number of centromere repositioning events [Table R2-3] according to this model is 44.

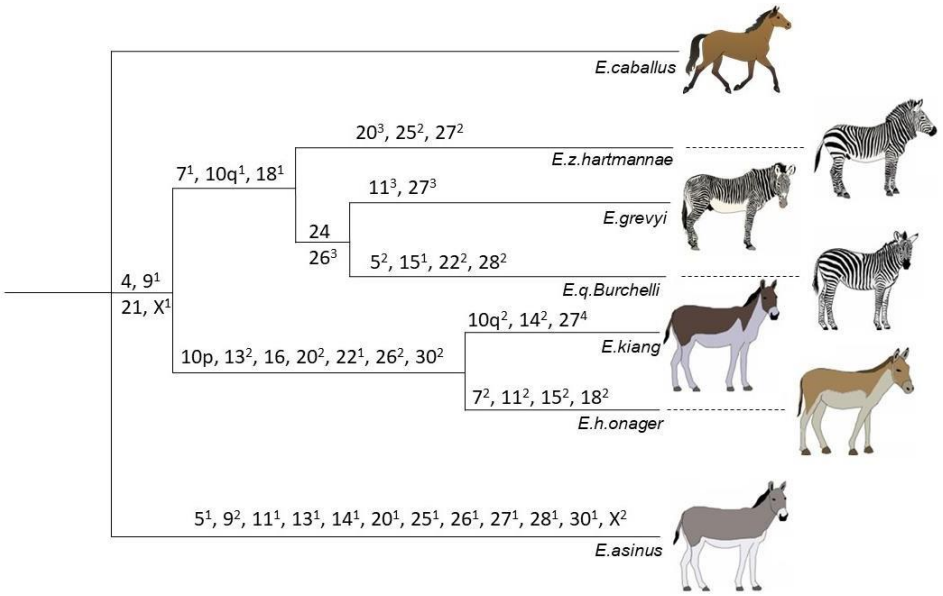


***Figure R2-31**: **Analysis of centromere repositioning events, based on the presence of satellite-less centromeres, in the lineages of a newly proposed Equus phylogenetic tree.** This phylogenetic tree was newly proposed to better explain some centromere repositioning event. E. caballus is considered as the outgroup. Each chromosome in which a satellite-less centromere was mapped is reported over the corresponding evolutionary lineage. Chromosome numbers correspond to those of the horse reference genome. Superscript numbers indicate neocentromeres observed in more than one lineage. According to this phylogenetic tree, these neocentromeres should derive from multiple repositioning events.*

CHR	N° of CR events	CHR	N° of CR events
ECA4	1	ECA18	2
ECA5	2	ECA20	3
ECA7	2	ECA21	1
ECA9	2	ECA22	2
ECA10q	2	ECA24	1
ECA10p	1	ECA25	2
ECA11	2	ECA26	3
ECA13	3	ECA27	4
ECA14	2	ECA28	2
ECA15	2	ECA30	2
ECA16	1	ECAX	2
TOTAL NUMBER OF CR EVENTS: 44			

Table R2-3: Number of centromere repositioning events deduced from the phylogenetic tree in Figure R2-31.

DISCUSSION

Following the identification of 16 evolutionarily new centromeres in *E. asinus* through a ChIP-seq approach we wanted to further investigate the molecular details of this enigmatic locus. In fact, lack of satellite sequences at these loci allowed us to dissect at molecular level this important site which is of vital importance for cell segregation and survival.
A first cytogenetic analysis conducted in our laboratory revealed the presence of several centromeres uncoupled from satellite DNA in *E. grevyi* and in *E. burchelli* (see Introduction). We wanted to extend the study to other equid species and decided to use the molecular approach applied to horse and donkey. In particular we included in our work three zebras and two wild asses: *E. zebra hartmannae, E. grevyi, E. burchelli, E. kiang and E. hemionus onager.*
We demonstrated that an incredible number of satellite-less centromeres is present in a different genomic location compared to the horse centromere in the above-mentioned species: a total of 82 satellite-less centromeres.

In *E. zebra hartmannae* we detected 10 evolutionarily new centromeres. Four out of
10 neocentromeres [EZH1-5-11-X] displayed some sequence rearrangements relative to the horse reference sequence, as shown previously in our laboratory (Francesco Gozzo PhD thesis 2018); two neocentromeres displayed sequence amplification [EZH10-14]; the remaining four neocentromeres displayed no major sequence rearrangements [EZH2-4-7-13]. We detected only one clear case of epiallelism [EZH2].

In *E. grevyi* 11 satellite-less centromeres are present. Five out of 11 [EGR4-5-11-16- 19] showed some sequence rearrangements relative to the horse reference sequence (Francesco Gozzo PhD thesis 2018), one displayed sequence amplification [EGR8], while the remaining four showed no sequence rearrangements [EGR9-15- 20-X]. We identified only one clear case of epiallelism, on Grevy's zebra centromere mapped on horse chromosome 26.

In *E. burchelli* we identified 14 evolutionarily new centromeres. 9 did not show any major sequence rearrangements [EBU4-9-10-12-14-15-16-20-21], four are characterized by sequence amplification [EBU7-8-17-18] similarly to some donkey centromeres. Finally, the *E. burchelli* centromere in the ortholog of ECAX displayed a peculiar sequence rearrangement. Three main peaks where identified, one of which is located almost 2 Mb away the others. This could suggest a sequence translocation compared to the horse reference sequence. Four satellite-less centromeres clearly showed the presence of epialleles [EBU10-14-15-21].
E. asinus centromeres were deeply discussed at molecular level in the previous chapter and in the attached paper (Nergadze SG et al 2018). Genomic positioning of these centromeres respect to the horse and to centromeres of other equid species will be discussed later.

In *E. kiang* 15 satellite-less centromeres were identified. In three cases [EKI14-17-X] we found sequence rearrangements relative to the horse reference sequence, as shown previously in our laboratory (Francesco Gozzo PhD thesis 2018). Twelve centromeres showed a Gaussian-like peak shape thus reflecting no sequence rearrangements [EKI1-2-4-5-9-10-11-13-16-18-20-25], while 5 of these centromeres showed epiallelism [EKI9-11-18-20-25]. Horse chromosome 10 is a peculiar case, since two satellite-less centromeres were identified: one in EKI11, ortholog to ECA 10q and one in EKI20 corresponding to ECA 10p. Moreover, the positions of the two satellite-less centromeres are really far away respect to each other and in the two different horse chromosome arms, which is explained by the fact that the two neocentromeres belong to two different kiang chromosomes.

In *E. hemionus onager* a total of 15 satellite-less centromeres localized in a different position respect to the horse centromeres. Five satellite-less centromeres showed major sequence rearrangements [EHO3-10-12-16-X], while the remaining showed a Gaussian-like peak shape [EHO2-9-11-14-18-19-21-22-23-27], therefore no major sequence rearrangements. Onager is characterized by the presence of many satellite-less centromeres in which two clearly distinct epialleles are present: 8 out of 15 showed epiallelism [EHO2-9-11-14-18-19-23]. This may be a peculiarity of the onager lineage since it's the equid species with the highest number of neocentromeres clearly displaying positional epialleles.

The cytogenetic and the molecular approaches gave some different results in the identification of the satellite-less centromeres although the majority of them are void of satellite with both methods. 9 out of the 11 Grevy's zebra neocentromeres and 7 out of the 14 Burchell's zebra neocentromeres here detected were demonstrated to be devoid of satellite DNA also at cytogenetic level (Piras et al 2010). We can divide the remaining cases in two groups: neocentromeres identified by ChIP-seq but marked with satellite DNA at FISH level (EBU4, 10, 12, 14, 15, 20, 21, X. EGR19, 20) and centromeres uncoupled from satellite DNA at cytogenetic level but not revealed by ChIP-seq (EBU6, 9. EGR1, 2, 3, 6, 13, 4, 17, 18, 22). These discrepancies are probably caused by a different resolution power of FISH experiments compared to ChIP-seq. In the first case, a possible explanation is that satellite sequences are not really centromeric but reside very close to the centromere even if they are detectable at the primary constriction of metaphasic chromosomes at FISH level. In the second case, the absence of ChIP-seq signal for FISH negative centromeres may be due to the fact that relatively short stretches of satellite repeats, not detectable by FISH, may impair the detection of centromeres through Next generation sequencing. Alternatively, species-specific sequences not present in the horse reference genome may be present at these centromeres or are not well assembled on chromosomes. This case was proved at least for one satellite- less centromere. In Burchell's zebra chromosome 9 no FISH signals, corresponding to satellite DNA, were detected (Piras MF et al. 2010) and no CENP-A binding domains were mapped when EquCab2 was used as reference.
However, when we mapped the same reads on the new version of the genome,

EquCab3, we found a ChIP-seq signal for that centromere, whose horse orthologous sequence was in ECA21 (Figure R2-20). This example suggests that some discrepancies between the two experimental approaches will be overcome when an accurate assembly of all equids genomes will be available.
It is important to keep in mind that the centromere repositioning analysis discussed in this thesis, was carried out on the basis of the sole presence of the satellite-less centromeres detected through ChIP-seq. An integration with the observation of the major cytogenetic rearrangements that contributed to the karyotype of these species will be needed to unravel the details of the centromere repositioning events occurred during the equid speciation.
The results presented in Figures R2-19-31 show, in several cases, that a neocentromere can be present in different lineages while absent in related lineages. To explain this puzzling observation, three different hypotheses can be proposed**.**

1) Chromosome rearrangement hypothesis: Following the formation of a neocentromere, various types of rearrangements (such as fusions, fissions, inversions, translocations) gave rise to chromosomes in which the centromere function moved to a different position.
2) Centromerization hotspot hypothesis: The formation of the same neocentromere in different lineages is due to independent centromere repositioning events driven by a genomic sequence acting as preferential seeding site for centromeric formation.
3) Incomplete lineage sorting: It is known that the equid species diverged recently and that their populations encountered bottle-necks during their evolution. This type of population structure and evolution may have favored the transmission of neocentromeres to some lineages and their disappearance in other lineages.

We think that these hypotheses are not mutually exclusive.

In Figure D2-1 the analysis of the orthologs of horse chromosome 4 is shown, in which ChIP-seq data from our laboratory and karyotype analysis from the literature (Musilova P et al. 2013, Trifonov VA et al. 2008, Figures I13-18) are taken into account. The absence of the neocentromere on the donkey ortholog of ECA4 may be explained by an inversion event that specifically occurred in the donkey lineage, which moved the centromere in a satellite-containing region (Figure I10). This information suggests that the centromere repositioning event occurred in the common ancestor of the genus *Equus*, and then a specific rearrangement occurred in the donkey lineage, displacing its centromere.

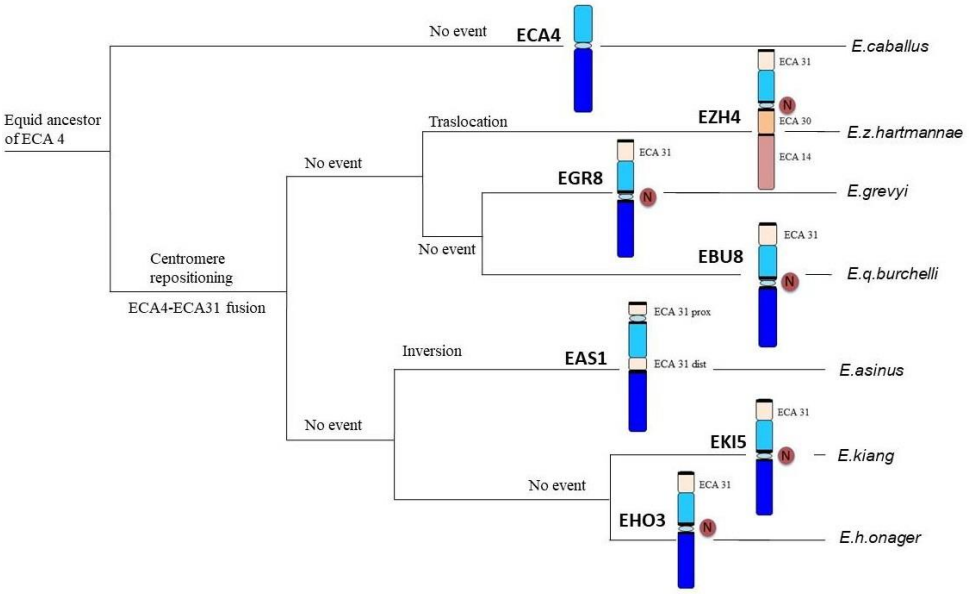
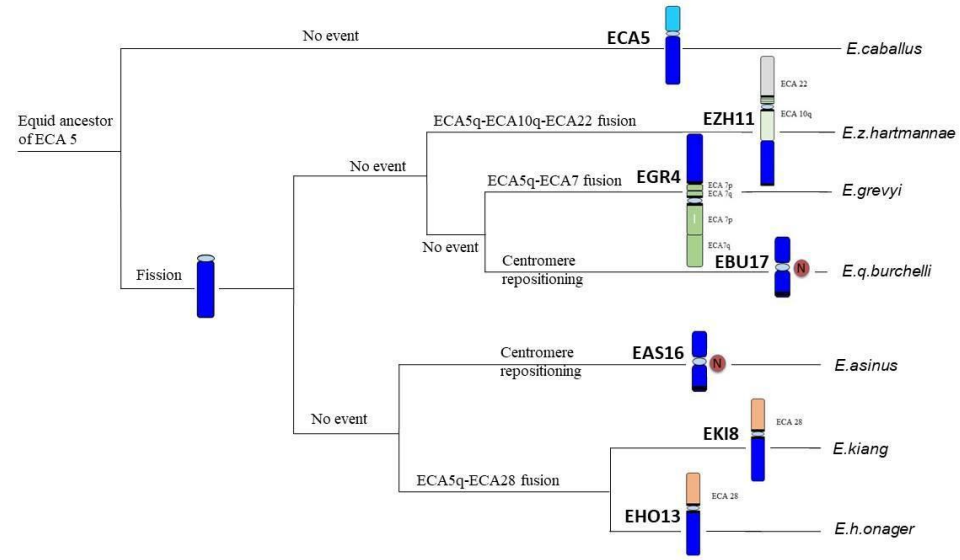


Figure D2-1: Phylogenetic tree of ECA 4 and its orthologous chromosomes. *This image shows the different orthologs of horse chromosome 4 in the different equid lineages; N indicates the presence of a neocentromere detected through ChIP-seq. Different colors indicate different arm-specific painting probes used to characterize orthologies between different species (Musilova et al 2013).*

The same type of analysis was carried out on the orthologs of ECA5. Figure D2-2A shows that, following a fission, two independent centromere repositioning events may have occurred: one generating a neocentromere on Burchell's zebra chromosome 17 and the other one on donkey chromosome 16. As reported in figure R2-3, the two neocentromeres colocalize in the same genomic region, suggesting that this is a preferential centromerization hotspot. Also previous data (Ventura M et al. 2004; Rocchi M et al. 2012) showed that some genomic regions have the potential of being latent centromere hotspots, and can be activated and used as seeding sites for neocentromeric formation. An alternative hypothesis is that a neocentromere was formed in the ancestor of all zebras and asses on the sequence corresponding to horse chromosome 5q. Several rearrangements then occurred in all lineages except *E .asinus* and *E. burchelli*, where the neocentromere was maintained (Figure D2-2B). A third hypothesis, previously proposed by our group (Piras MF et al. 2009), is that a single centromere repositioning event occurred in a putative common ancestor of *E. asinus* and *E. burchelli*, according to the phylogenetic tree shown also in Figure R2-30.

A



B

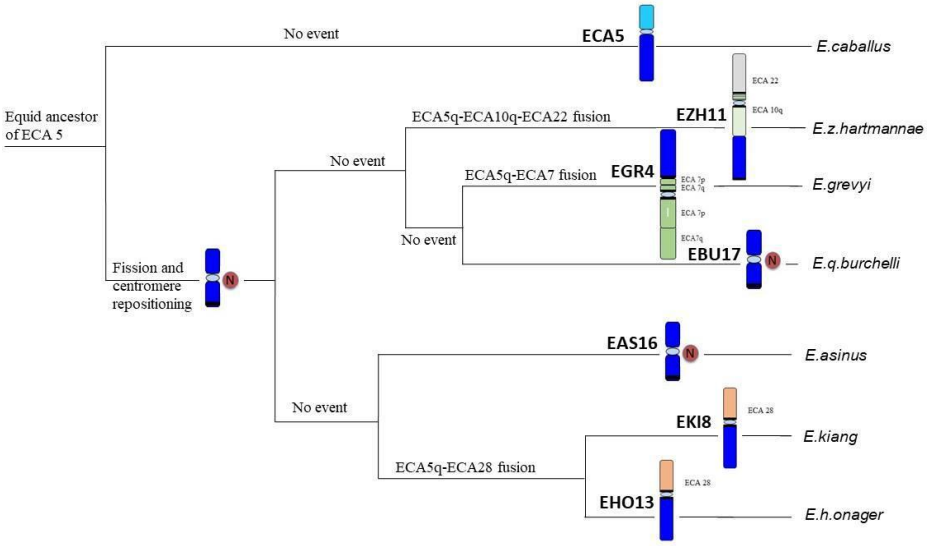


Figure D2-2: Phylogenetic tree of ECA 5 and its orthologous chromosomes. *The two panels show the different orthologs of horse chromosome 5 in the different equid lineages based on two different hypotheses. A) phylogenetic tree of ECA5 and its orthologs based on centromerization hotspot hypothesis. B) phylogenetic tree of ECA5 and its orthologs based on the chromosome rearrangement hypothesis. N indicates the presence of a neocentromere detected through ChIP-seq. Different colors indicate different arm-specific painting probes used to characterize orthologies between different species (Musilova et al 2013).*

These examples point out that an analysis integrating different approaches (ChIP-seq and karyotype analysis) is the correct experimental way to proceed to shed light on the formation of the satellite-less centromeres. A comparative analysis of all *Equus* chromosomes in which we observed satellite-less neocentromeres is under way in our laboratory.

Some of the neocentromeres here detected formed upon fission or fusion events that occurred in some equid lineage, as previously shown in our laboratory (Francesco Gozzo PhD thesis 2018). Many equid chromosomes deriving from an ancestral fusion event (such as EBU4-10-12, EGR4-9, EZH1-2-4-7-10-13, EKI2-9-; Musilova et al 2013) had gone through centromere repositioning, since a satellite-less centromere was identified by ChIP-seq. Moreover, since these neocentromeres map on the orthologous sequence of one of the two chromosomes involved in the fusion event, we can conclude that satellite-less centromeres do not reside exactly on the fusion point but they tend to be repositioned on one side of the fusion point.

As shown in Figure D2-2 a fission event may have induced the formation of the neocentromere in donkey and Burchell's zebra orthologs of horse chromosome 5. This case suggests that after a fission event, centromere repositioning can occur, generating a neocentromere in a really distant position compared to the ancestral centromere, as previously reported (Ventura M et al. 2003; Ventura M et al. 2004).

The reason behind the centromere repositioning upon fusion or fission events could be that the old centromere is functionally destroyed and, to rescue this damaging event, a neocentromere will form in a different position. Molecular data suggest that, events in which DNA double-strand breaks occur (like chromosomal rearrangements) are resolved and rescued also by recruiting CENP-A (Zeitlin SG et al. 2009). It was demonstrated that CENP-A and other centromeric proteins are involved in DNA repair mechanisms, and since chromosomal rearrangements involve DNA breaks, loci in which DNA repair mechanisms are active may act as seeding genomic regions for centromeric formation due to the presence of CENP-A which bears this dual functional nature. However this last hypothesis must be more investigated since other studies demonstrate no correlation between neocentromeres and breakpoints (Warburton PE et al. 2000).

We also observed several centromere repositioning events uncoupled from obvious chromosome rearrangements, as for the donkey ortholog of ECA13 (Figure R2-12, Nergadze SG et al 2018).

During the identification of the neocentromeres formed by CR we were able to detect some features already recognized in the donkey satellite-less centromeres (Nergadze et al. 2018), sequence rearrangements and/or amplification and epiallelism.

Many neocentromeres displayed sequence rearrangements, relative to the horse reference sequence (e.g. Grevy's zebra centromere of the ortholog of ECA21,

Figure R2-20A), as shown previously in our laboratory (Francesco Gozzo PhD thesis 2018).
Other neocentromeres exhibited sequence amplification (e.g. Burchell's zebra centromere of the ortholog of ECA15, Figure R2-14A) which was previously identified in the donkey by NGS and specific experiments (Nergadze SG et al 2018). We propose that sequence amplification may be the first step towards the establishment of satellite DNA.
We also detected many cases in which two distinct epialleles were clearly visible within the same neocentromere (e.g. EZH2, EBU14, EKI20, EHO19). As previously reported (Purgato S et al. 2015, Nergadze SG et al. 2018) the centromeric position can vary among individuals but also on the two homologs of a chromosome in the same individual. Interestingly, the onager was the species in which we identified the highest number of epialleles. We don't know whether this peculiarity is to be ascribed to the onager species itself or to that individual. Our data suggest that some species differences in epiallelism may be present. We go from the lowest number of epialleles in the Grevy's zebra (1 out of 11 neocentromeres clearly displays epiallelism) to the highest number in onager (7 out of 15 neocentromeres clearly display epiallelism). It would be necessary to analyze several individuals from each species to test this hypothesis.
With the results shown in this thesis we can also correlate two different phenomena. The first one is the centromere repositioning, described as events of centromere movement during species evolution. The second is centromere sliding, a phenomenon of centromeric repositioning at smaller scale, which may occur from one generation to another (Nergadze SG et al. 2018). In this study we were able to detect both of these phenomena.
So, the distance of the neocentromeres respect to the ancestral centromere (here represented by the horse centromeres) may have been influenced by centromere repositioning only but also by centromere sliding as well. Some neocentromeres repositioned very close to the horse centromeric regions such as in ECA4 or ECA21 orthologs (less than 1 Mb), while other repositioned very far away from horse centromeres such as in ECA16 (almost 50 Mb away) or ECA30 orthologs (almost 19 Mb away). However chromosomal rearrangements may have further influenced the distance of the neocentromeres which are always measured on the horse reference sequence.
To establish a phylogenetic history of those chromosomes carrying repositioned centromeres we inferred multiple models of equid evolutionary tree using neocentromeres as markers. The lineage of the horse diverged from the ancestor of asses and zebras, about 2 MYA (Oakenfull EA et al. 2000). Other equid species recently and rapidly evolved and their karyotype went through a great number of chromosomal rearrangements (Trifonov VA et al. 2008). In particular, CR importantly contributed to their evolution (Carbone L et al. 2006, Giulotto E et al 2017). It has been suggested that centromere repositioning may be a major cause of

the genome plasticity observed in this genus. Taking into account the criterion of maximum parsimony, a lower number of CR events is more probable respect to a higher number of the same event occurring multiple times. In our case, the classical phylogenetic tree model revealed the lowest number of CR events as 40 in total versus the 44 of the new model and the 45 of the model proposed with this study. This suggests that the classical tree could represent a best-fit evolutionary reconstruction giving the lowest number of centromere repositioning events. However, since each chromosome has its own unique evolutionary history, also the other two models fit quite well if not better with some case.

In conclusion, thanks to the analysis of so many ENCs in the equids, we report data and suggest interpretations and insights about the evolution of centromeric loci that had been so far hindered by the presence of satellite DNA. Moreover, we can use the satellite-less centromeres as markers to investigate the genus *Equus* evolution, tracking their speciation history.

PART 3 GENOME-WIDE EPIGENETIC ANALYSIS OF SATELLITE-LESS CENTROMERES IN HORSE AND DONKEY

RESULTS

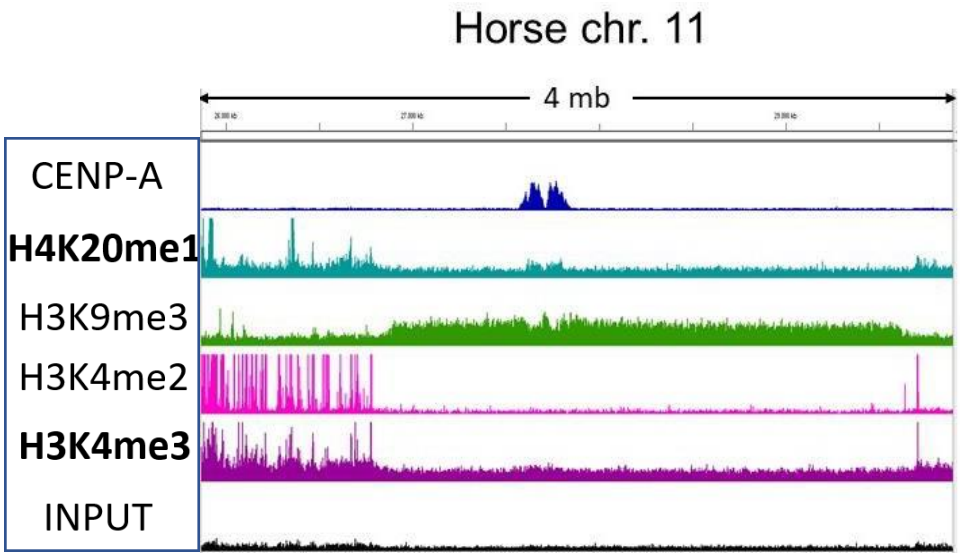
As previously reported in the PhD thesis of Riccardo Gamba (2017), we analyzed the epigenetic and transcriptional profile of the horse and donkey satellite-less centromeres. We immunoprecipitated chromatin of fibroblast cell lines from horse and donkey with antibodies against different histone modifications and, through Next generation sequencing (NGS) approach, we retrieved information about the centromeric loci by analyzing the genomic regions surrounding the CENP-A binding domains.
In particular we investigated the presence of H3K9me3 (histone H3 trimethylated at Lysine 9), H3K36me2 (histone H3 dimethylated at Lysine 36), H3K4me2, H3K4me3 (histone H3 dimethylated or trimethylated at Lysine 4) and H4K20me1 (histone H4 monomethylated at Lysine 20). These results were reported in the PhD thesis of Riccardo Gamba. In my thesis I report the newly analyzed markers in the donkey (H4K20me1 and H3K4me3) and, for comparison, I report again the results of CENP-A, H3K4me2 and H3K9me3.
Thanks to our model system and exploiting the high homology between the horse and the donkey genome, we were able to analyze and compare orthologous genomic regions, which are identical at DNA level, but different at functional level, being centromeric in one species and not centromeric in the other species.
After a ChIP-seq assay, horse reads were mapped on the EquCab2.0 reference, while donkey reads were mapped on the EquCabAsiA chimeric reference genome as described in Material and Method section.
The results of these experiments are summarized in figures R3 1-17. Each image comprises the CENP-A dataset, the newly analyzed markers (H4K20me1 and H3K4me3) and a couple of the previously analyzed markers (H3K9me3 and H3H4me2). Each figure is divided in two panels. In Figure R3-1, Panel A shows the epigenetic status of the centromeric domain region of the horse chromosome 11 (Equcab2). Panel B shows its orthologous non centromeric region on the donkey (EquCabAsiA). Figure R3-2 through Figure R3-17 are also divided in two panels. Panel A shows the epigenetic status of the centromeric domain region of the donkey (EquCabAsiA). Panel B shows its orthologous non centromeric region on the horse (Equcab2.0).

The blue tracks represent CENP-A peaks and allow the identification of the region in which the centromeric function resides.
The cyan track shows the results of the ChIP-seq assay with the H4K20me1

marker. In vertebrates, the H4K20me1 histone modification was previously identified in the centromeric nucleosomes that contain the CENP-A histone variant (Hori T et al. 2014; Bailey AO et al. 2015).
The green and the pink tracks show the H3K9me3 (an histone modification associated to constitutive heterochromatin Barski A et al. 2007) and H3K4me2 (an histone modification associated to transcriptionally competent, "open" chromatin Vakoc CR et al. 2006) datasets, respectively, already described in the thesis of Riccardo Gamba.
The purple track shows the H3K4me3 dataset, a histone modification typically associated to transcriptionally active chromatin in vertebrates and, in particular, to transcription starts sites of active genes (Ruthenburg AJ et al. 2007).
Finally, the black one is the Input track, used as control. Since the input sequencing represents the sequencing of non-immuno-precipitated DNA (like a whole genome sequencing), it helps us to unravel possible bioinformatic biases. This is particularly useful when we analyze sequencing data mapped on the EquCabAsiA reference, since only the centromeric portions of the donkey genome are assembled, while the rest of the genome is essentially the horse sequence. Since unnormalized reads are reported we use the input dataset as control. The profile of the input dataset should be flat, and when a peak is present in both ChIP and in the input, it should not be considered enriched.
The two newly generated markers were firstly preliminarily analyzed at genome-wide level and then the analysis focused on the centromeric regions as here described.

The centromere of horse chromosome 11

A



B

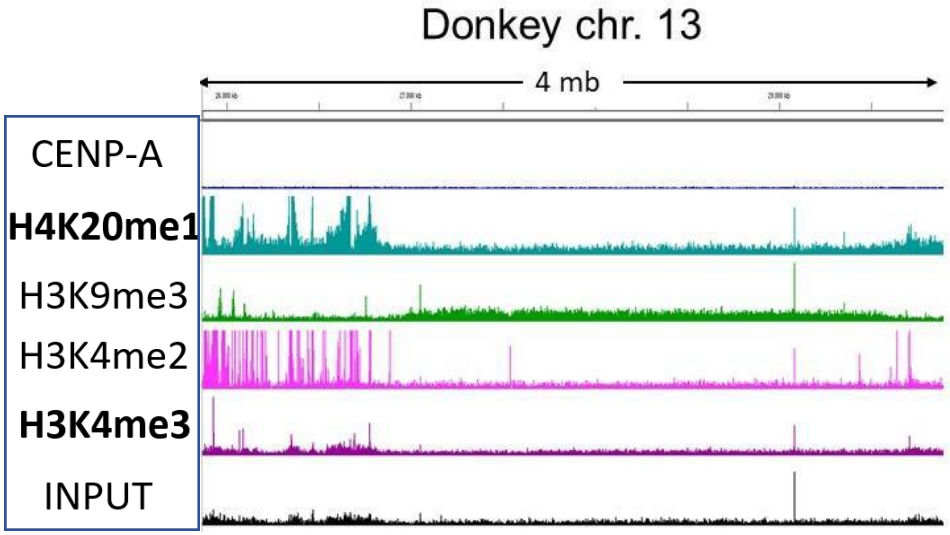
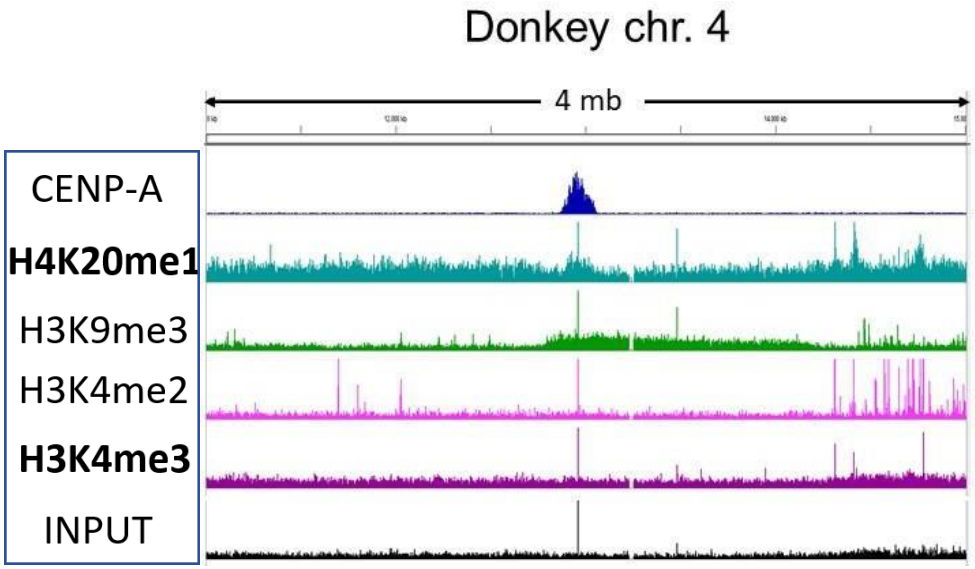


Figure R3-1: Epigenetic profile of the centromeric locus of horse chromosome 11 on Equcab2.0 (panel A) and its orthologous non centromeric region on donkey chromosome 13 on EquCabAsi (panel B). *Colored tracks correspond to the different sequencing datasets obtained with the antibodies listed on the left. For all the datasets, unnormalized read count is plotted.*

Figure R3-1 shows the analysis of the epigenetic markers at the centromeric locus of the horse chromosome 11. A region of about 4 Mb is displayed. Analysis on the H4K20me1 marker reveals the presence of two peaks exactly at the same location of the CENP-A domain (about 200 kb), confirming the association between CENP-A and H4 monomethylated in Lysine 20. Interestingly, a visible signal depression of this marker seems to surround the centromeric locus, which corresponds in length to the heterochromatic domain.
On the contrary, the marker for open chromatin, H3K4me3, is absent in the region corresponding to the heterochromatin (H3K9me3), confirming once again the transcriptional silent state of the centromeric domains in our model system.
Figure R3-1A shows that the centromeric peak is embedded in a ~2.8 Mb wide heterochromatic region marked by an abundant enrichment in H3K9me3 which is also present in the donkey (Figure R3-1B).
Transcriptionally silent status of this region is also supported by the absence of ChIP-seq signal of H3K4me2 both in the centromere of horse chromosome 11 and in the orthologous region on donkey chromosome 13.
Figures R3-2-17 report the results of the epigenetic analysis of the 16 satellite-less centromeres of the donkey (panels A) and at their orthologous regions on the horse (panels B).

The centromere of donkey chromosome 4

A



B

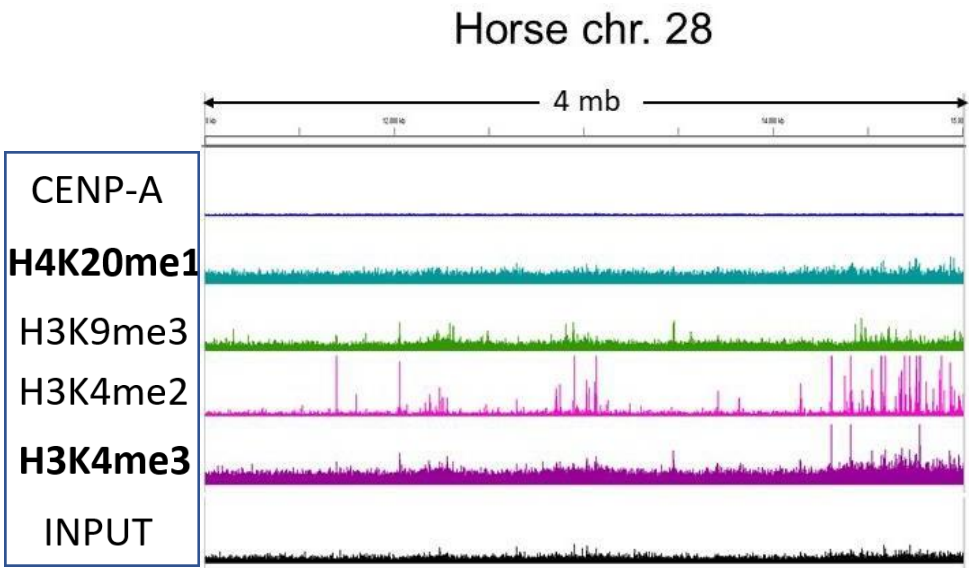
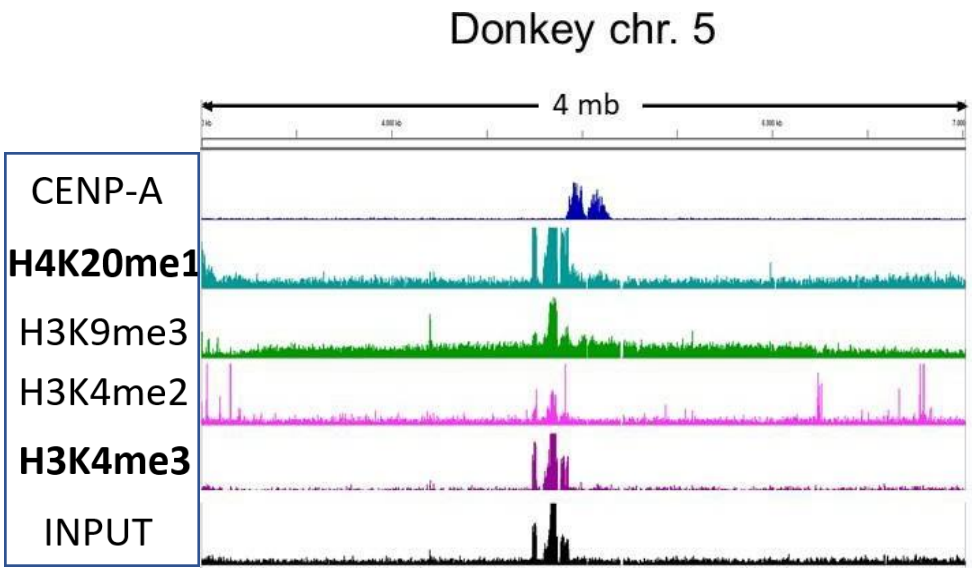


Figure R3-2: Epigenetic profile of the centromeric locus of donkey chromosome 4 on EquCabAsi panel A) and its orthologous non centromeric region on horse chromosome 28 on Equcab2.0 (panel B). *Colored tracks correspond to the different sequencing datasets obtained with the antibodies listed on the left. For all the datasets, unnormalized read count is plotted.*

Figure R3-2shows the results of the analysis at the satellite-less centromere of EAS 4 (panel A) and its orthologous region on ECA 28 (panel B). A ChIP-seq signal (about 180kb) of the H4K20me1 marker is present in correspondence of the CENP-A binding domain. A signal depression of the marker is not clearly visible. H3K4me3 signal is absent from the centromeric locus confirming the transcriptional silent status of this centromeric region.
Only in the donkey sample, we detected a ~ 1.5 Mb wide heterochromatic region defined by the H3K9me3 peak (panel A), asymmetrically surrounding the CENP-A binding domain and transcriptional activity absence.
In the horse, H4K20me1 marker is absent from the corresponding orthologous non-centromeric region, while H3K4me3 marker is equally absent in the two species.

The centromere of donkey chromosome 5

A



B

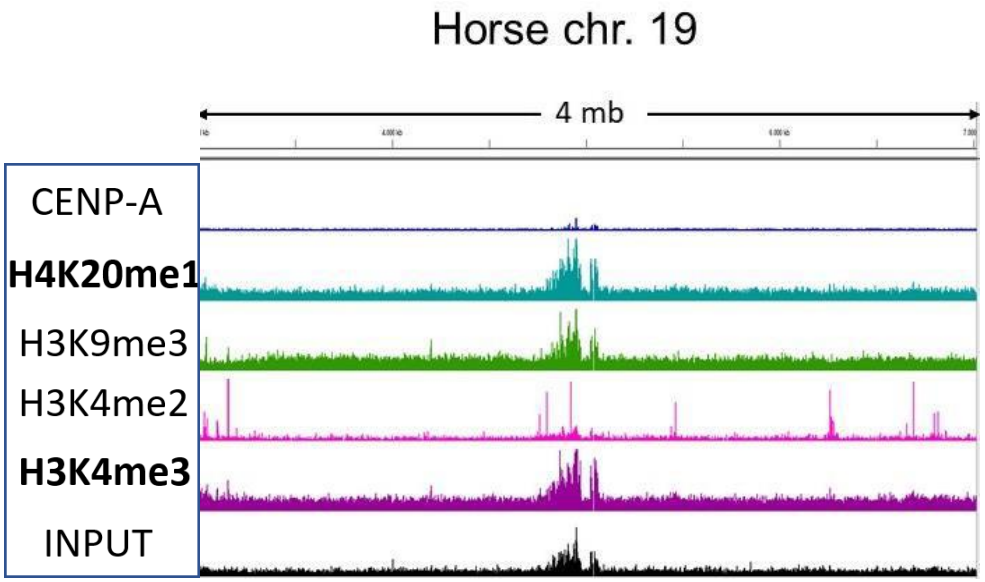
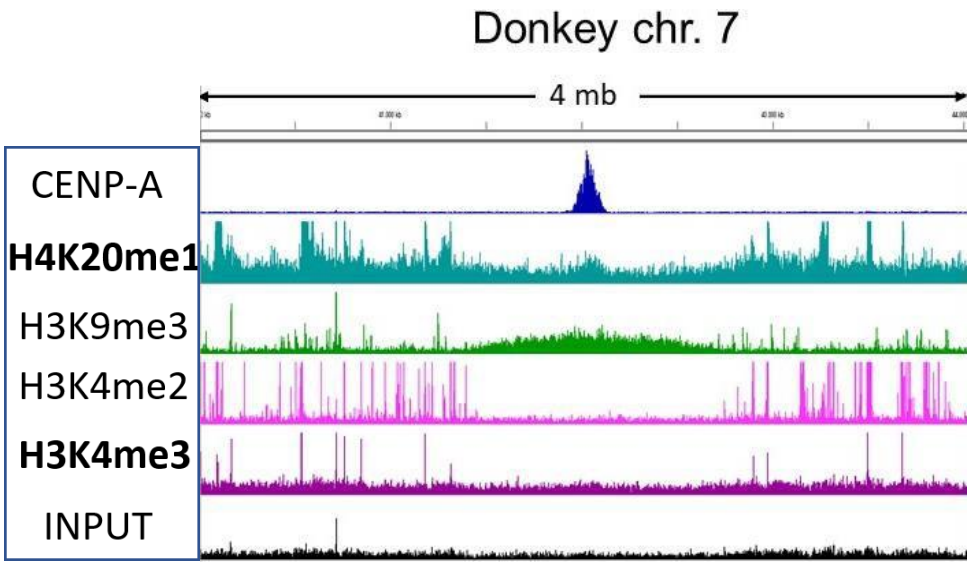


Figure R3-3: Epigenetic profile of the centromeric locus of donkey chromosome 5 on EquCabAsiA (panel A) and its orthologous non centromeric region on horse chromosome 19 on Equcab2.0 (panel B). *Colored tracks correspond to the different sequencing datasets obtained with the antibodies listed on the left. For all the datasets, unnormalized read count is plotted.*

Figure R3-3 shows the analysis of the epigenetic markers at the centromere of EAS 5 (panel A) and its orthologous region on ECA 19 (panel B). In both species, the input sample showed a complex peak localized very close to the centromeric site of donkey chromosome 5. This peak, which may be due to sequence duplication present in both species, is displayed also in all the ChIP-seq datasets of the epigenetic markers. This could be indicative of a bioinformatic bias, so H4K20me1 signal is not easily detectable, despite a slight enrichment in a ~ 100 kb region corresponding to the right centromeric epiallele may be present. H3K4me3 shows the same enrichment pattern of the other transcriptional marker.
Only in the donkey sample, we detected a ~ 3.5 Mb wide heterochromatic region defined by an H3K9me3 peak, surrounding the CENP-A binding domain, and absence of transcriptional activity.
Also, in the horse we observe the similar pattern of the newly analyzed datasets (H4K20me1 and H3K4me3), although being the analysis impaired by the duplication. Similarly to the donkey, we observed lack of transcriptional activity in the horse but also lack of heterochromatin signal.

The centromere of donkey chromosome 7

A



B

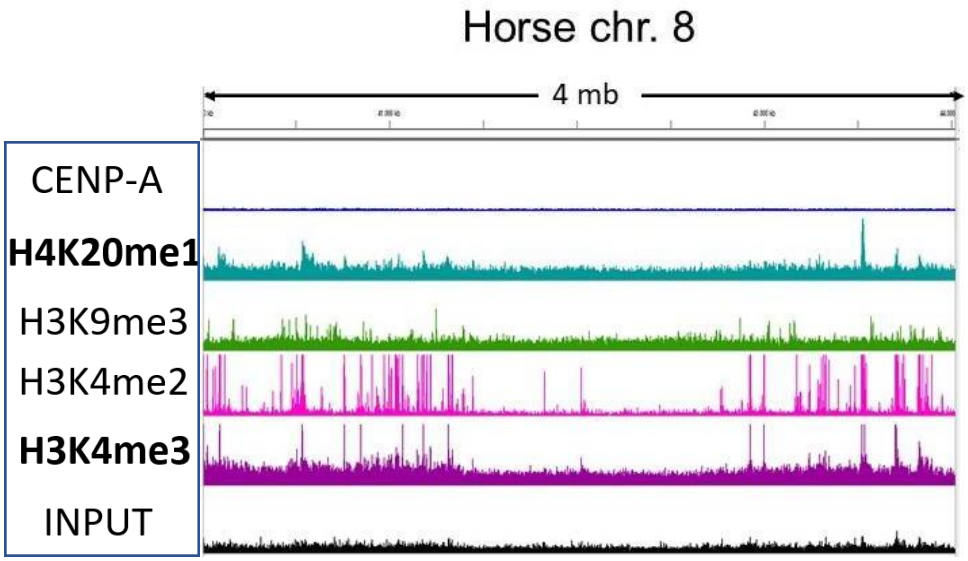
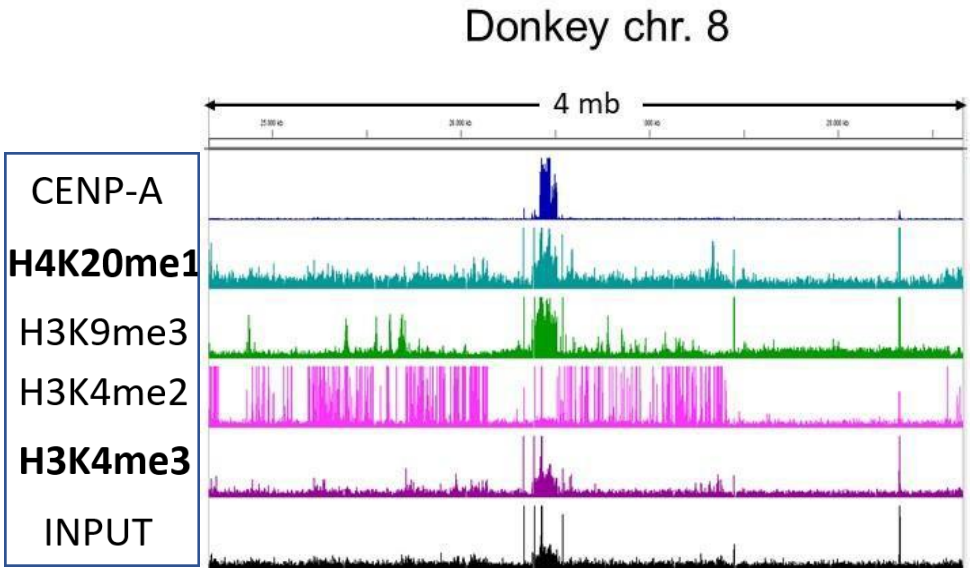


Figure R3-4: Epigenetic profile of the centromeric locus of donkey chromosome 7 on EquCabAsiA (panel A) and its orthologous non centromeric region on horse chromosome 8 on Equcab2.0 (panel B). *Colored tracks correspond to the different sequencing datasets obtained with the antibodies listed on the left. For all the datasets, unnormalized read count is plotted.*

Figure R3-4 shows the epigenetic profile of the EAS 7 centromere (panel A) and its orthologous region on ECA 8 (panel B). An H4K20me1 peak (~ 200 kb) is present in correspondence of the CENP-A binding domain, while a signal depression of this marker, corresponding to the heterochromatin domain is visible surrounding the centromeric locus. H3K4me3 signal is absent from the centromeric region confirming the transcriptional silent status of this centromeric region.
Only in the donkey sample, a ~1.2 Mb wide heterochromatic region is observed (panel A, green track) surrounding the CENP-A binding domain (panel A, blue track), while lack of signals of H3K4me2 reflects absence of transcriptional activity.
In the horse we observe lack of signals of both H4K20me1 and H3K4me3 markers. Moreover, we observed lack of transcriptional activity but also lack of heterochromatin signal in the horse.

The centromere of donkey chromosome 8
A



B

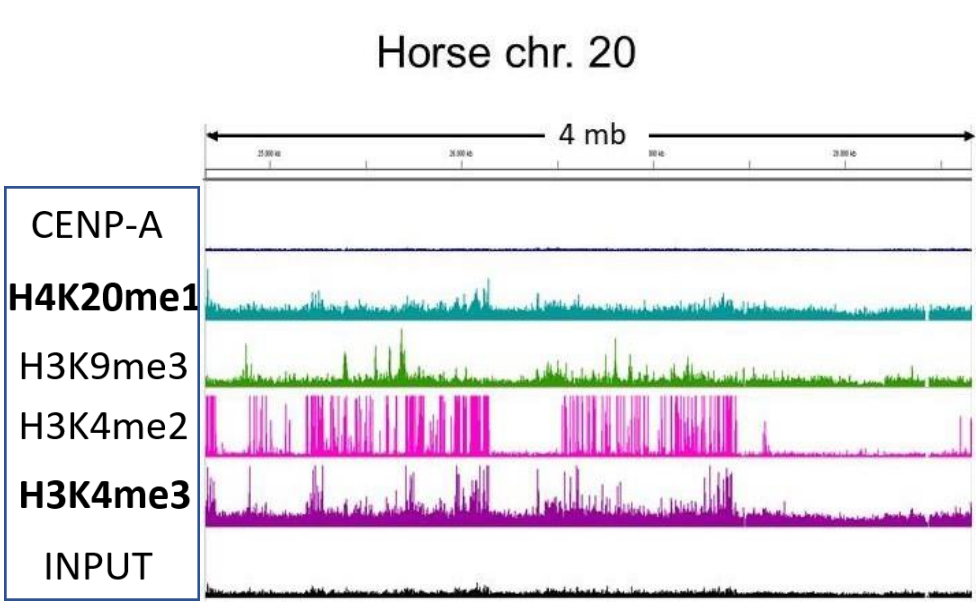
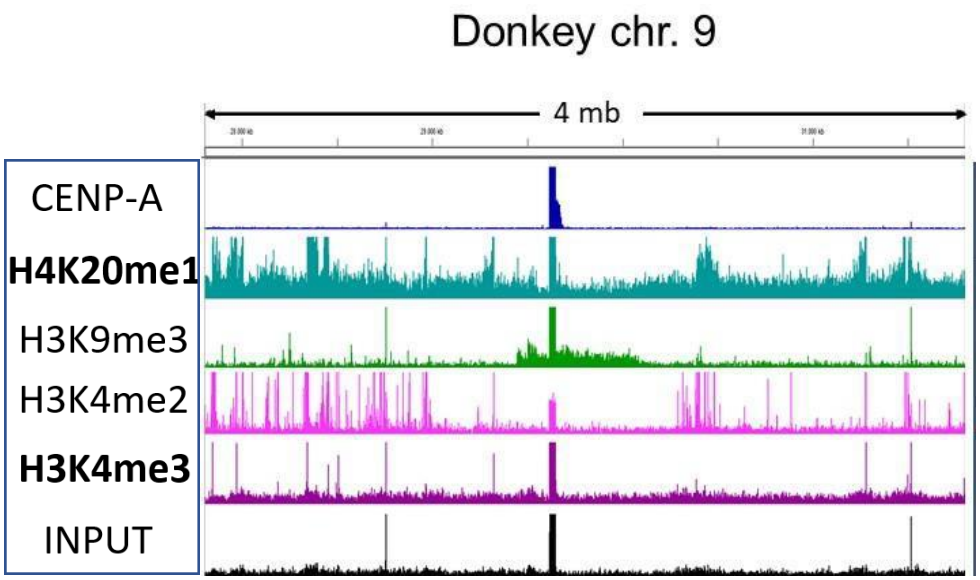


Figure R3-5: Epigenetic profile of the centromeric locus of donkey chromosome 8 on EquCabAsiA (panel A) and its orthologous non centromeric region on horse chromosome 20 on Equcab2.0 (panel B). *Colored tracks correspond to the different sequencing datasets obtained with the antibodies listed on the left. For all the datasets, unnormalized read count is plotted.*

Figure R3-5 shows the results of the analysis at the centromere of donkey chromosome 8 (panel A) and its orthologous region on horse chromosome 20 (panel B).
In the donkey centromeric region, the input shows a complex peak, consequence of a poorly assembled reference sequence and of the presence of duplicated sequences as previously discussed in Part 1 and in previous work (Nergadze et al. 2018) . This organization is not present in the orthologous horse region. This situation influenced the data analysis, so an epigenetic profile of this region is hardly trackable.
In the horse orthologous non centromeric region H4K20me1 signal seems to be absent, although some minor peaks are present in the region. H3K4me3 marker seems to concord with H3K4me2.
Overall, due to mapping artefacts in the donkey, little information can be obtained from this centromere.

The centromere of donkey chromosome 9

A



B

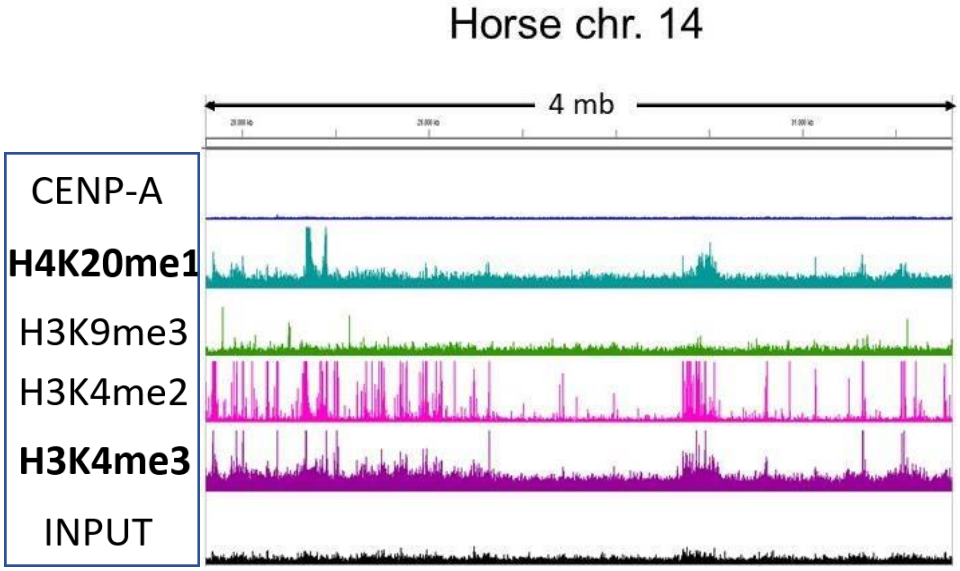
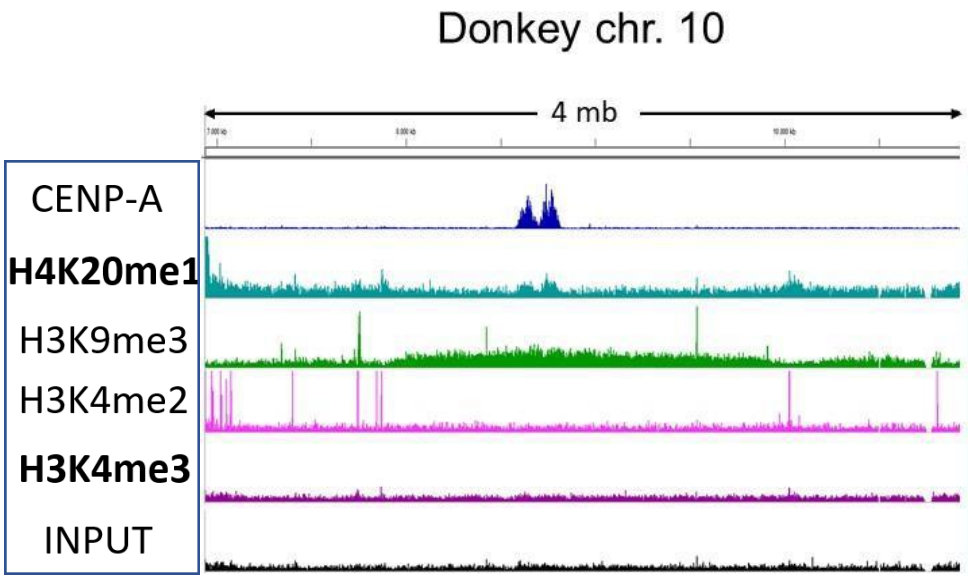


Figure R3-6: Epigenetic profile of the centromeric locus of donkey chromosome 9 on EquCabAsiA (panel A) and its orthologous non centromeric region on horse chromosome 14 on Equcab2.0 (panel B). *Colored tracks correspond to the different sequencing datasets obtained with the antibodies listed on the left. For all the datasets, unnormalized read count is plotted.*

Figure R3-6 shows the results of the analysis at the centromere of donkey chromosome 9 (panel A) and its orthologous region on horse chromosome 14 (panel B).
As previously reported in Part 1, a sequence duplication (represented by the narrow peak in all datasets) is present in this donkey region and is absent in the horse (panel B). This situation influenced the data analysis, so an epigenetic profile of this region is hardly trackable; H4K20me1 seems to have a lower enrichment in the region surrounding the centromere exactly corresponding to the heterochromatic domain. H3K4me3 signal is absent in the centromeric locus but also in the surrounding area H3K4me2. In the donkey sample, we detected a ~700 kb wide heterochromatic region.
In the horse we observe lack of signal of both H4K20me1 and H3K4me3 markers. Moreover, we observed lack of heterochromatin signal in the horse.

The centromere of donkey chromosome 10.

A



B

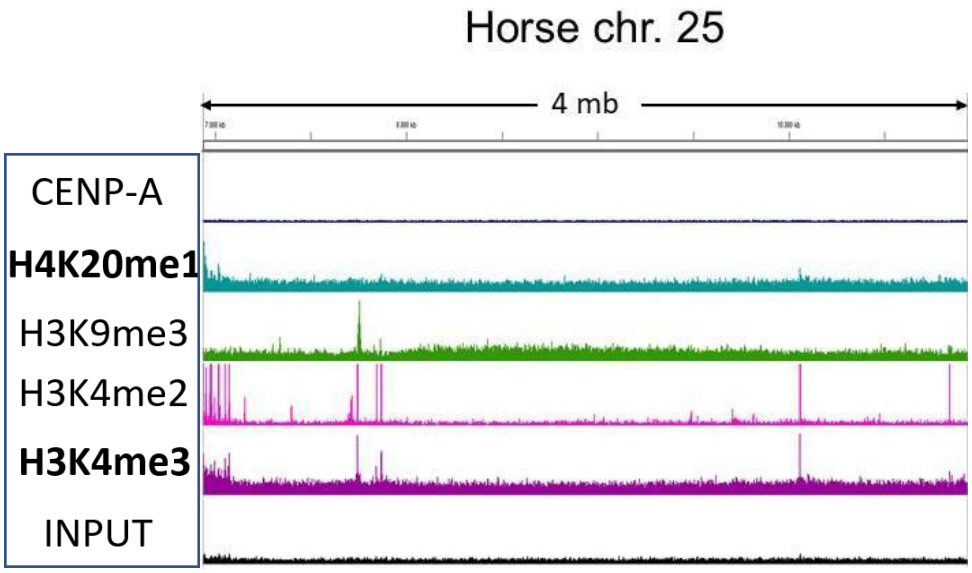
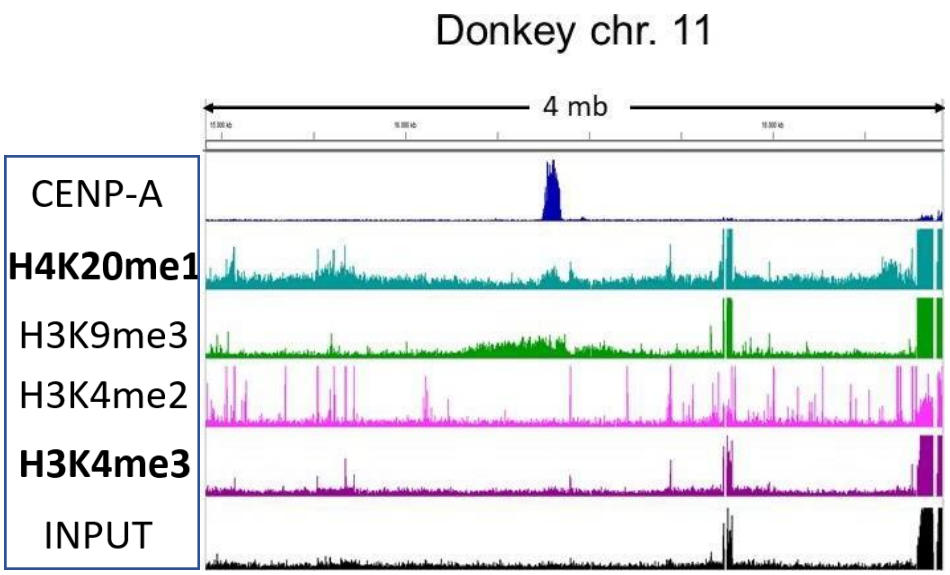


Figure R3-7: Epigenetic profile of the centromeric locus of donkey chromosome 10 on EquCabAsiA (panel A) and its orthologous non centromeric region on horse chromosome 25 on Equcab2.0 (panel B). *Colored tracks correspond to the different sequencing datasets obtained with the antibodies listed on the left. For all the datasets, unnormalized read count is plotted.*

Figure R3-7 shows the results of the analysis at the centromere of donkey chromosome 10 (panel A) and its orthologous region on horse chromosome 25 (panel B).
A ChIP-seq signal (~ 200 kb) of the H4K20me1 marker is present in correspondence of the CENP-A binding domain. Moreover, a slight signal depression of the marker is observable surrounding the centromeric locus and coinciding with the heterochromatic domain.
H3K4me3 signal is absent from the centromeric locus confirming the transcriptional silent status of this centromeric region as shown previously in our laboratory (Riccardo Gamba PhD thesis 2017). A ~2 Mb wide heterochromatic region surrounding the CENP-A binding domain is present.
In the corresponding orthologous non centromeric region of the horse sample, a slight H4K9me3 enrichment seems to be present, although it is difficult to distinguish it from the input. H3K4me3 signal is absent also in the horse but H4K20me1 signal is not present either.

The centromere of donkey chromosome 11

A



B

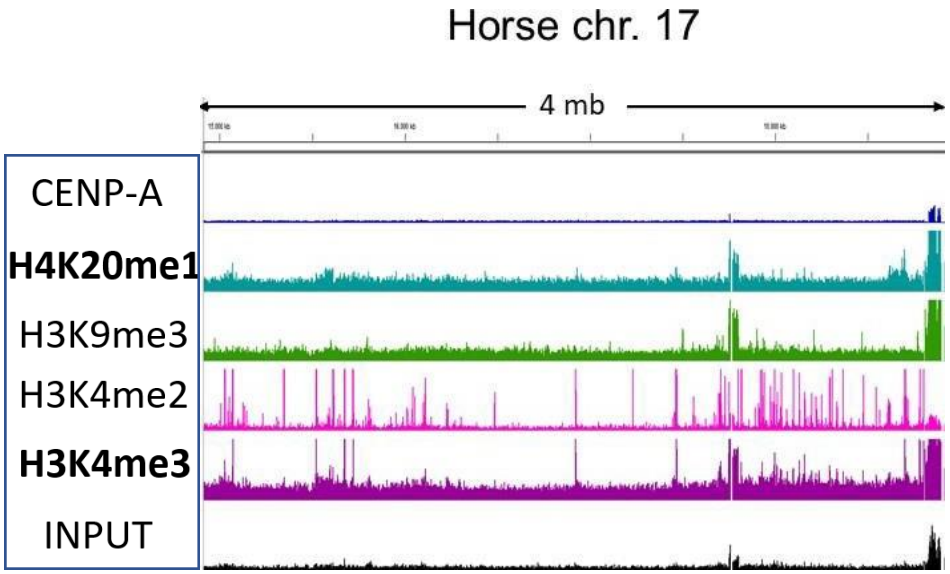
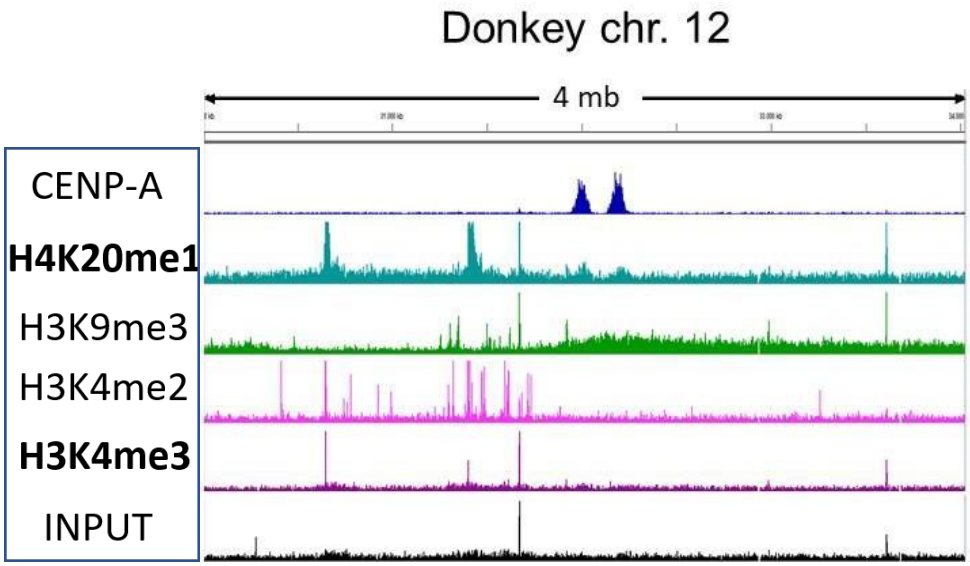


Figure R3-8: Epigenetic profile of the centromeric locus of donkey chromosome 11 on EquCabAsiA (panel A) and its orthologous non centromeric region on horse chromosome 17 on Equcab2.0 (panel B). *Colored tracks correspond to the different sequencing datasets obtained with the antibodies listed on the left. For all the datasets, unnormalized read count is plotted.*

Figure R3-8 shows the results of the analysis at the centromere of donkey chromosome 11 (panel A) and its orthologous region on horse chromosome 17 (panel B).
A ChIP-seq signal (~ 100 kb) of the H4K20me1 marker is present in correspondence of the CENP-A binding domain. Moreover, a signal depression of the marker is visible surrounding the centromeric locus and coinciding with the heterochromatic domain. A ~800 kb wide heterochromatic region surrounds the CENP-A binding domain; it is briefly interrupted outside the right border of the CENP-A peak by the site of transcription of a gene (UFM1) as shown previously in our laboratory (Riccardo Gamba PhD thesis 2017. This gene is transcribed both in the horse and in the donkey.
H3K4me3 signal is absent from the centromeric locus confirming the transcriptional silent status of this centromere. In the corresponding horse orthologous non centromeric region, both markers seem not to be enriched (apart from the H3K3me3 peak corresponding to the UFM1 gene).

The centromere of donkey chromosome 12

A



B

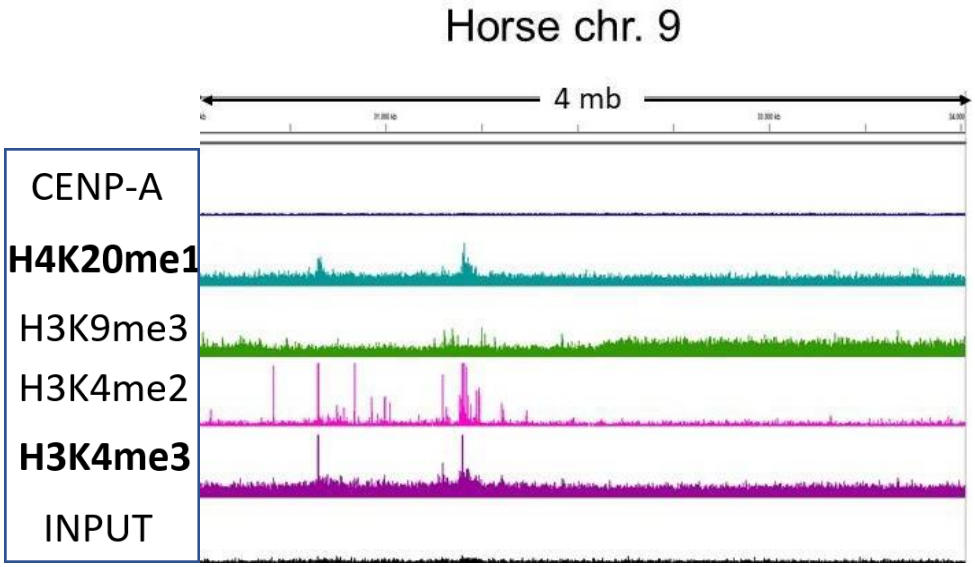
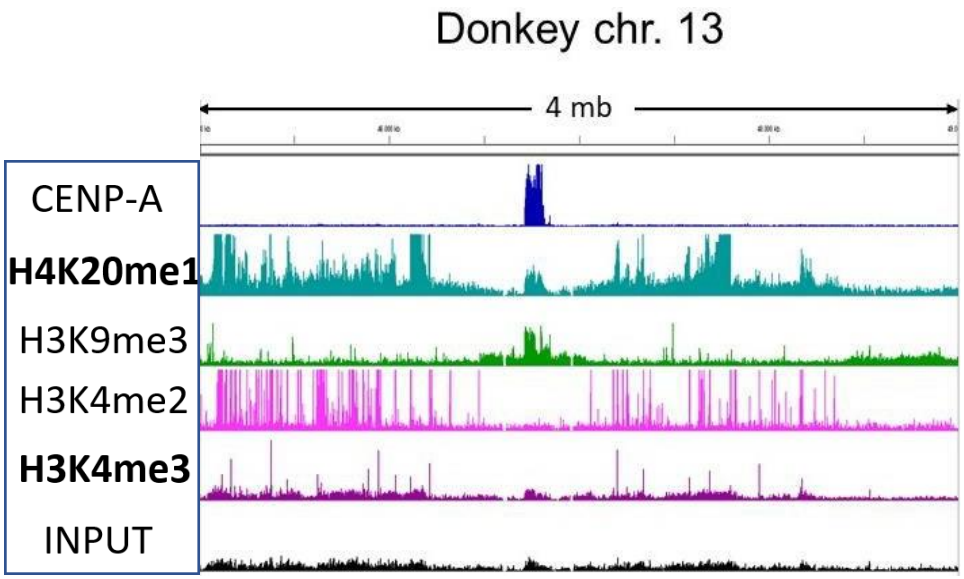


Figure R3-9: Epigenetic profile of the centromeric locus of donkey chromosome 12 on EquCabAsiA (panel A) and its orthologous non centromeric region on horse chromosome 9 on Equcab2.0 (panel B). *Colored tracks correspond to the different sequencing datasets obtained with the antibodies listed on the left. For all the datasets, unnormalized read count is plotted.*

Figure R3-9 shows the results of the analysis at the centromere of donkey chromosome 12 (panel A) and its orthologous region on horse chromosome 9 (panel B).
In the donkey sample, we observed a 2.5 Mb wide heterochromatic region that surrounds the CENP-A binding domains. In the same region of the horse sample, a slight enrichment seems to be present, although it is difficult to distinguish it from the input.
A ChIP-seq signal of the H4K20me1 marker is present in correspondence of the CENP-A binding domain; two H4K20me1 peaks (~ 100 kb each) are present and perfectly matching the two centromeric epialleles. H4K20me1 signal is absent in the horse sample.
In both species, H3K4me3 is absent from this region confirming the lack of transcriptional activity reported also by the analysis of H3K4me2 as previously shown (Riccardo Gamba PhD thesis 2017).

The centromere of donkey chromosome 13

A



B

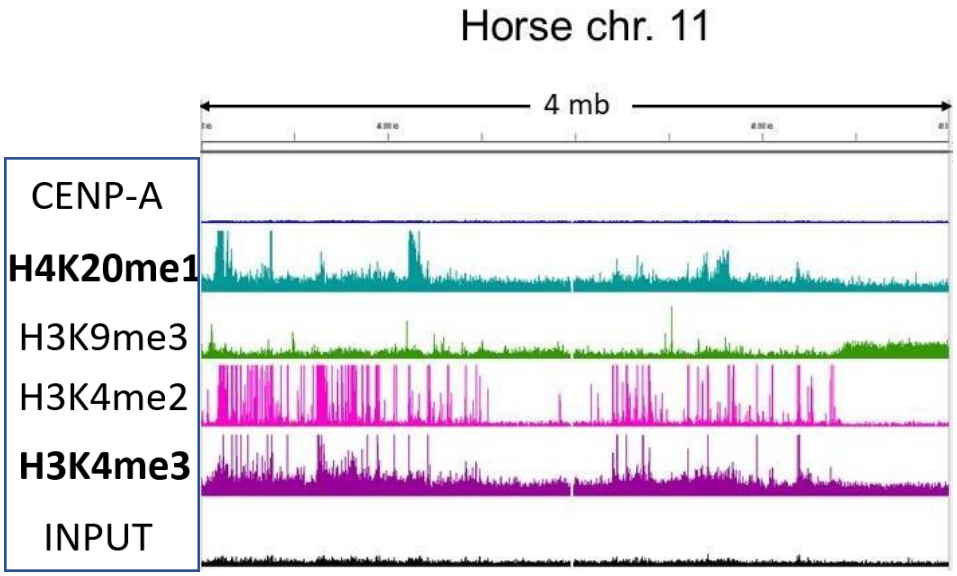
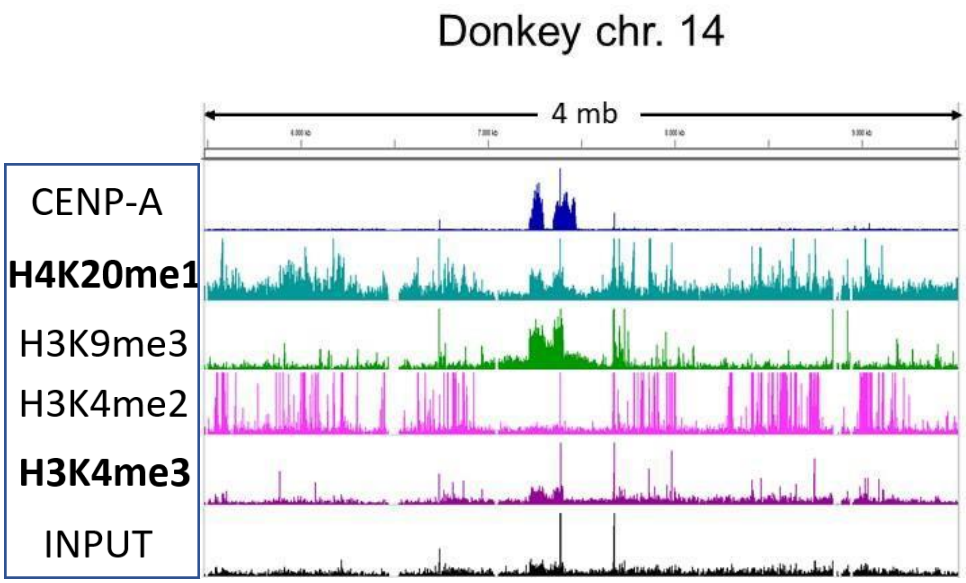


Figure R3-10: Epigenetic profile of the centromeric locus of donkey chromosome 13 on EquCabAsiA (panel A) and its orthologous non centromeric region on horse chromosome 11 on Equcab2.0 (panel B). *Colored tracks correspond to the different sequencing datasets obtained with the antibodies listed on the left. For all the datasets, unnormalized read count is plotted.*

Figure R3-10 shows the centromeric locus of donkey chromosome 13 (panel A) and its orthologous region on horse chromosome 11 (panel B).
A ~500 kb wide heterochromatic region surrounds the CENP-A binding domain.
A ChIP-seq signal (~ 110 kb) of the H4K20me1 marker is present in correspondence of the CENP-A peak. Moreover, a signal depression of the marker is visible surrounding the centromeric locus and coinciding with the heterochromatic domain.
No H3K4me3 signals are present in this region. H3K4me2 is absent as well, both in the horse and in the donkey orthologous sequence.
In the horse orthologous non centromeric region H4K20me1 is not enriched.

The centromere of donkey chromosome 14

A



B

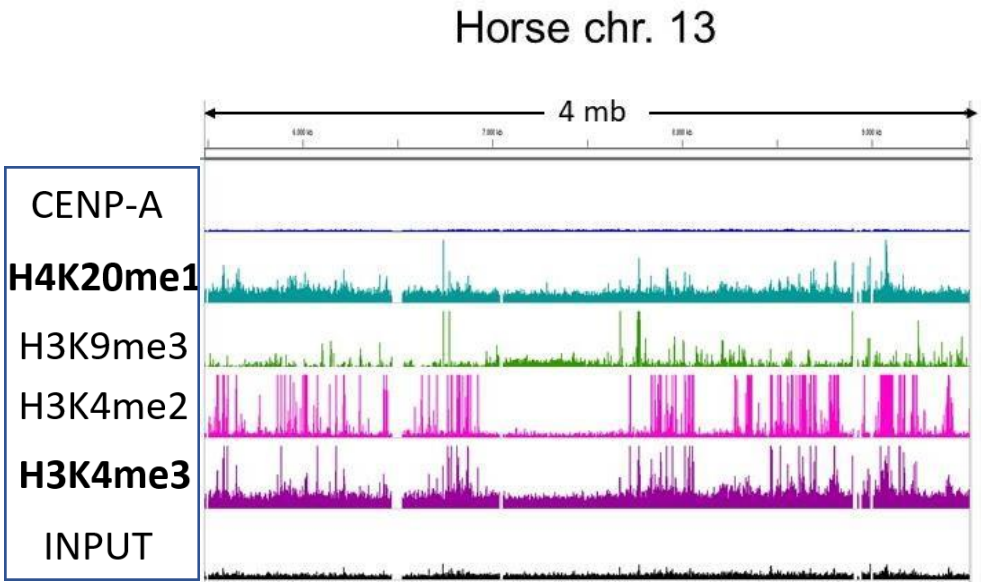
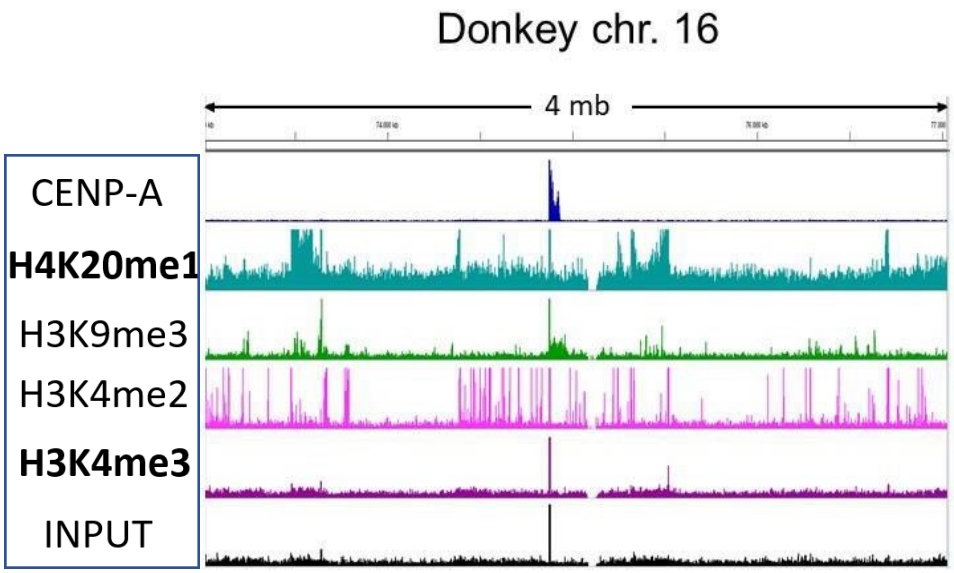


Figure R3-11: Epigenetic profile of the centromeric locus of donkey chromosome 14 on EquCabAsiA (panel A) and its orthologous non centromeric region on horse chromosome 13 on Equcab2.0 (panel B). *Colored tracks correspond to the different sequencing datasets obtained with the antibodies listed on the left. For all the datasets, unnormalized read count is plotted.*

Figure R3-11 shows the results of the analysis at the centromere of donkey chromosome 14 (panel A) and its orthologous region on horse chromosome 13 (panel B).
The peaks in the donkey input dataset suggest the presence of some sequence duplication. Despite the bioinformatic bias a heterochromatic region of about 500 kb surrounding the CENP-A peaks was identified.
H4K20me1 profile is hard to track in the donkey, due to the sequence duplication, although some degree of enrichment is present in the positions of the two centromeric epialleles. No enrichments of this marker are visible in the horse orthologous non centromeric region.
H3K4me3 enrichment is hard to establish, but comparing the ChIP sample to the input, it seems not to be enriched.
In the horse orthologous non centromeric region H4K20me1 and H£K3me3 are not enriched.

The centromere of donkey chromosome 16

A



B

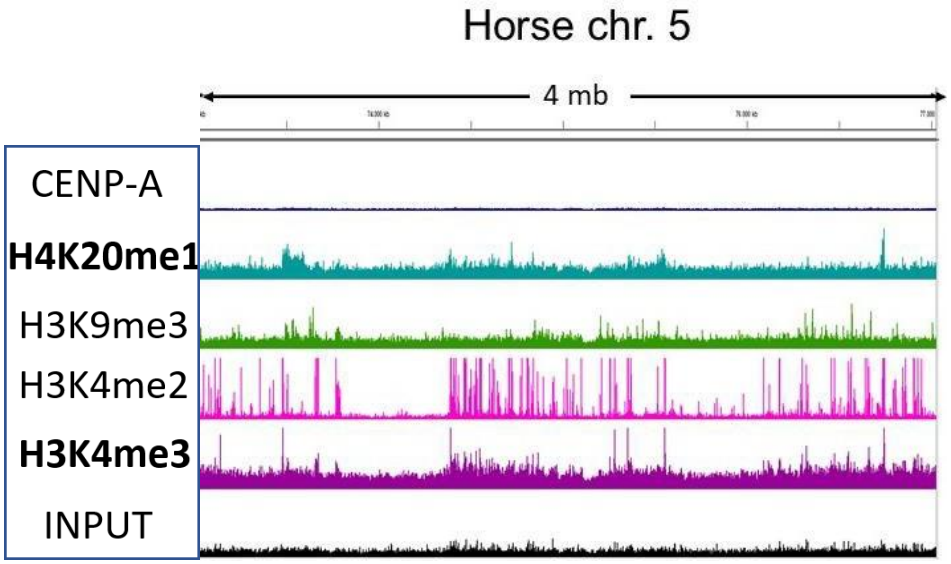
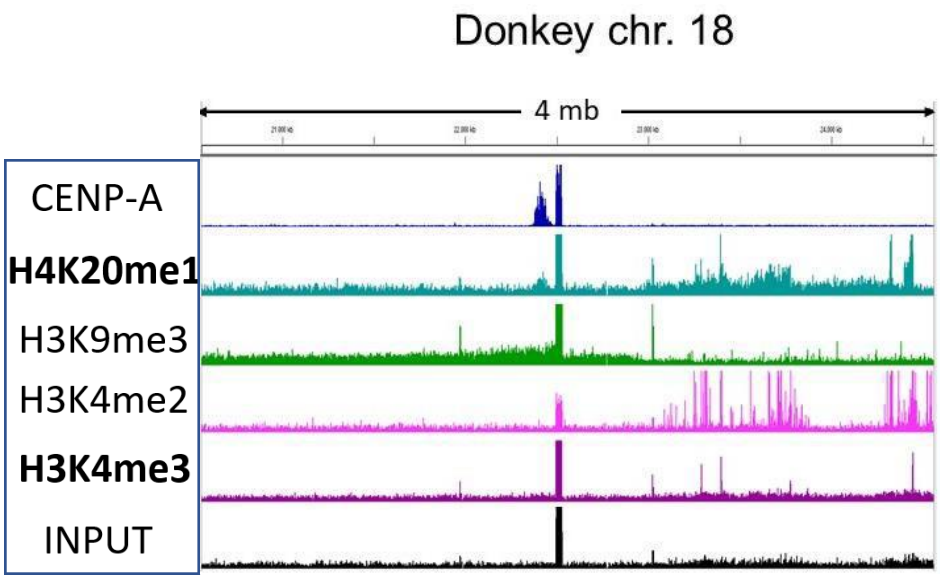


Figure R3-12: Epigenetic profile of the centromeric locus of donkey chromosome 16 on EquCabAsiA (panel A) and its orthologous non centromeric region on horse chromosome 5 on Equcab2.0 (panel B). *Colored tracks correspond to the different sequencing datasets obtained with the antibodies listed on the left. For all the datasets, unnormalized read count is plotted.*

Figure R3-12 shows the results of the analysis at the centromere of donkey chromosome 16 (panel A) and its orthologous region on horse chromosome 5 (panel B).
This centromeric locus partially resides on an amplified genomic sequence as reported in Part 1. This situation impairs the analysis due to a poor sequence assembly. A narrow heterochromatic peak (~100kb) was detected and mostly overlaps the CENP-A binding site; however, the heterochromatic region is probably extended on the array of duplicated sequences.
H4K20me1 seems not to be enriched in this region, but this result may be impaired by the presence of the sequence duplication, which is present only in the donkey.
Except from this narrow heterochromatic domain, the donkey centromere seems to be present in a transcriptionally open environment as shown previously in our laboratory (Riccardo Gamba PhD thesis 2017) and confirmed by H3K4me3. This chromatin status seems to be similar also in the horse orthologous region.

The centromere of donkey chromosome 18

A



B

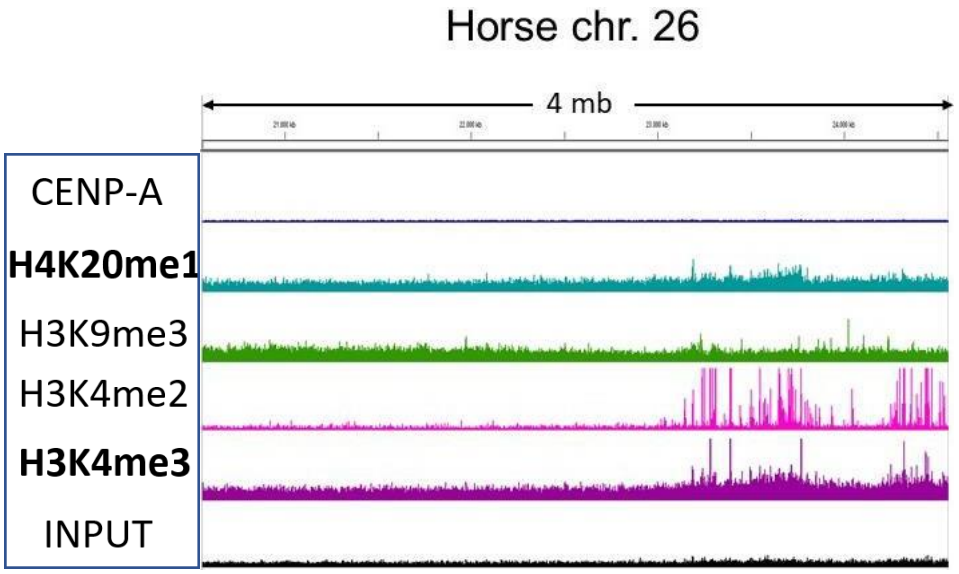
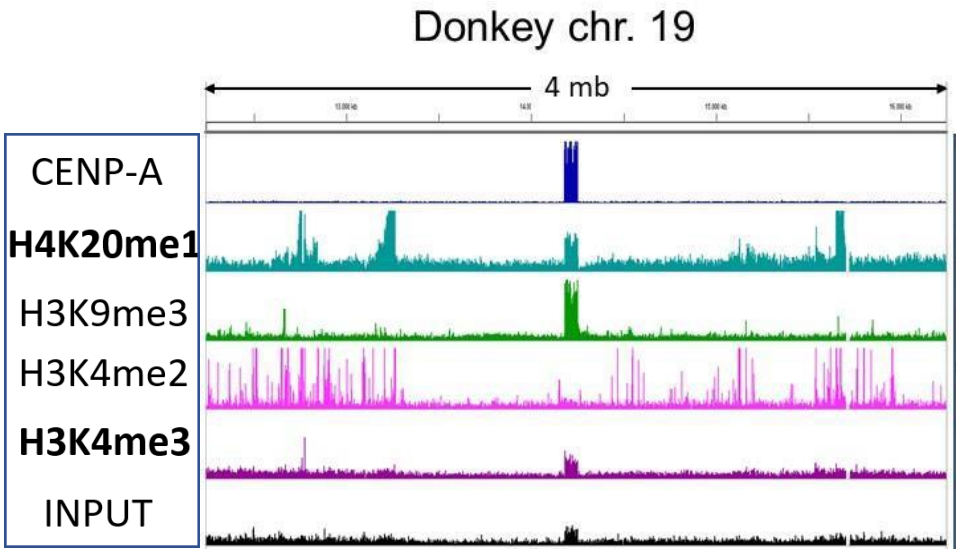


Figure R3-13: Epigenetic profile of the centromeric locus of donkey chromosome 18 on EquCabAsiA (panel A) and its orthologous non centromeric region on horse chromosome 26 on Equcab2.0 (panel B). *Colored tracks correspond to the different sequencing datasets obtained with the antibodies listed on the left. For all the datasets, unnormalized read count is plotted.*

Figure R3-13 shows the results of the analysis at the centromere of donkey chromosome 18 (panel A) and its orthologous region on horse chromosome 26 (panel B).
As previously reported in Part 1 (Nergadze et al. 2018), part of this centromere resides on an amplified sequence in the donkey genome which is represented by the narrow spike peak in the input. We identified a ~6 Mb wide heterochromatic region which embeds the CENP-A peaks at its end proximity.
Two H4K20me1 peaks are present colocalizing with the Gaussian-like CENP-A peak (~ 100 kb) and with the spike peak (~30 kb). This marker is completely absent in the horse orthologous region.
H3K4me3 marker is absent in both species.

The centromere of donkey chromosome 19

A



B

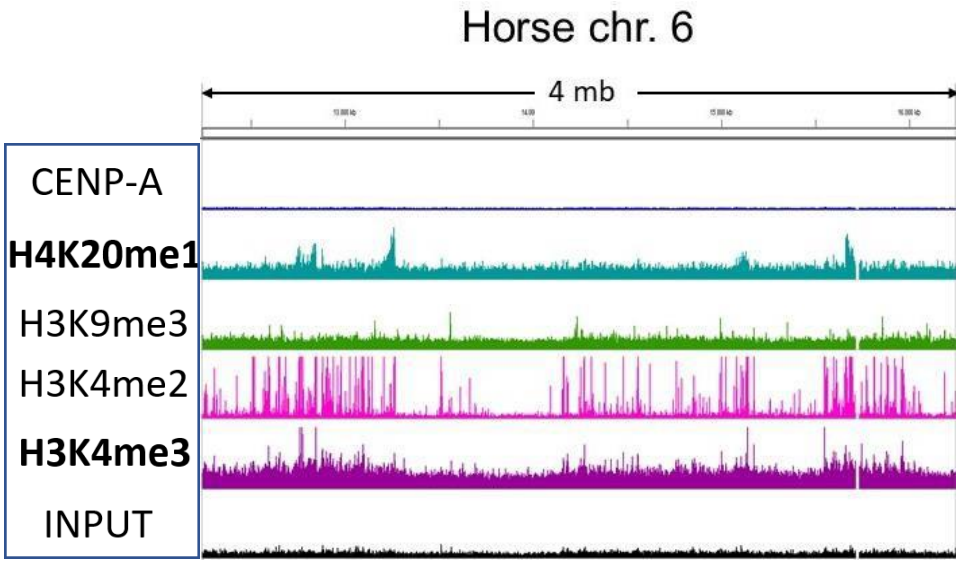
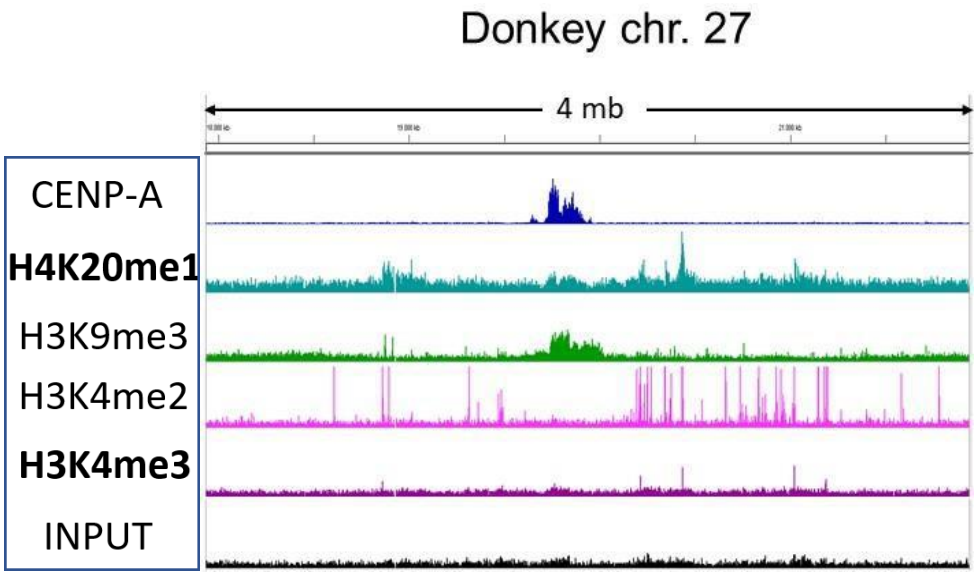


Figure R3-14: Epigenetic profile of the centromeric locus of donkey chromosome 19 on EquCabAsiA (panel A) and its orthologous non centromeric region on horse chromosome 6 on Equcab2.0 (panel B). *Colored tracks correspond to the different sequencing datasets obtained with the antibodies listed on the left. For all the datasets, unnormalized read count is plotted.***.**

Figure R3-14 shows the centromeric locus of donkey chromosome 19 (panel A) and its orthologous region on horse chromosome 6 (panel B).
This centromeric locus partially resides on an amplified genomic sequence as reported in Part 1. This situation impairs the datasets analysis since the lack of a well-defined sequence assembly. Possibly for this reason, we detected a narrow heterochromatic region around the centromeric peak.
H4K20me1 seems to be enriched as well in the amplified sequence; however a precise data analysis could not be executed.
H3K4me3 signal seems to have the same enrichment of the input peak; therefore no enrichment of this marker is present.

The centromere of donkey chromosome 27

A



B

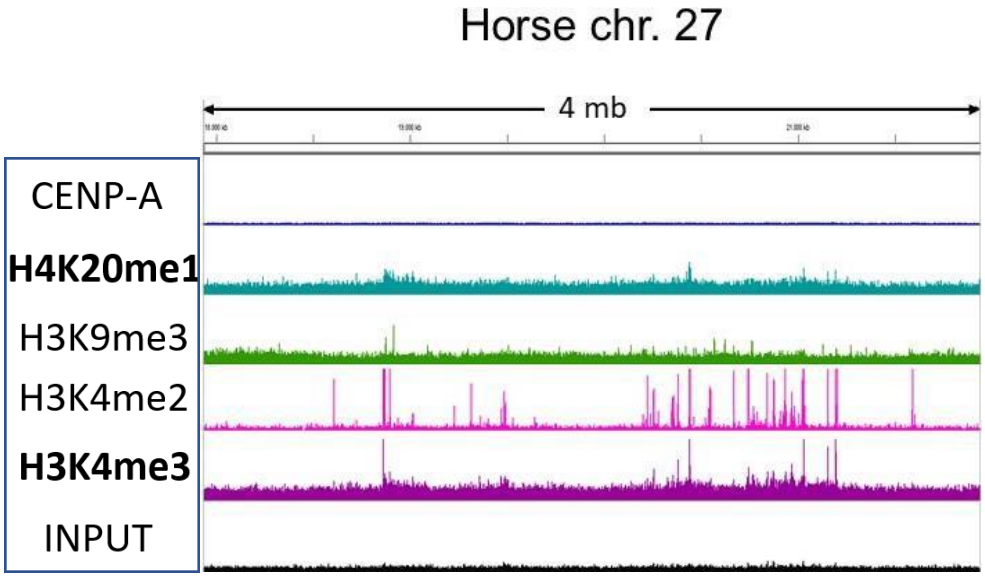
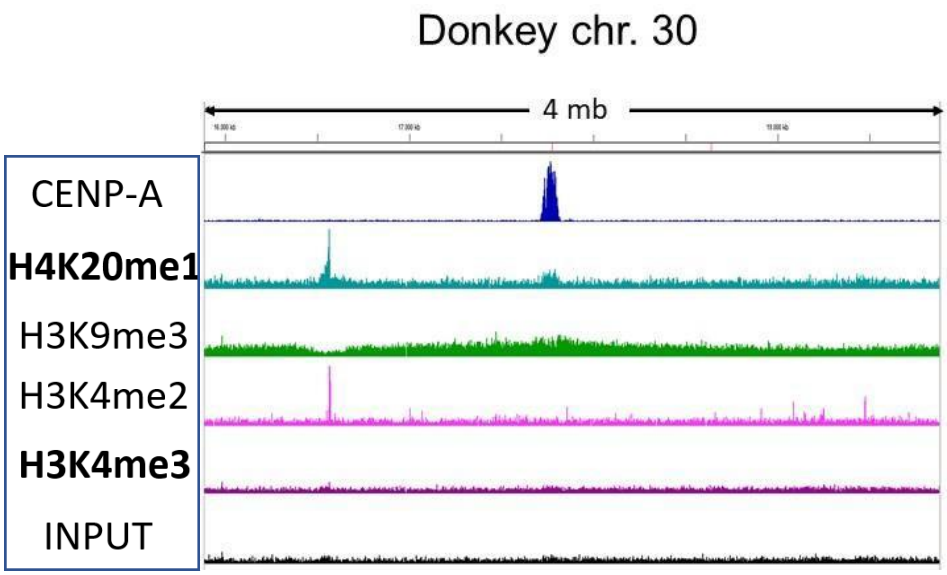


Figure R3-15: Epigenetic profile of the centromeric locus of donkey chromosome 27 on EquCabAsiA (panel A) and its orthologous non centromeric region on horse chromosome 27 on Equcab2.0 (panel B). *Colored tracks correspond to the different sequencing datasets obtained with the antibodies listed on the left. For all the datasets, unnormalized read count is plotted.*

Figure R3-15 shows the results of the analysis at the centromere of donkey chromosome 27 (panel A) and its orthologous region on horse chromosome 27 (panel B).
At this centromere, we detected a short heterochromatic domain of ~200 kb in width which corresponds almost exactly to the centromeric domain. On the horse orthologous region, no heterochromatin was detected.
A slight enrichment seems to be present for H4K20me1 in the regions exactly corresponding to the CENP-A peaks (~180 kb), despite being present also in the surrounding region. In the horse this epigenetic marker is absent from this region.
H3K4me3 signal is completely absent both from this region and from the horse orthologous non centromeric region.

The centromere of donkey chromosome 30

A



B

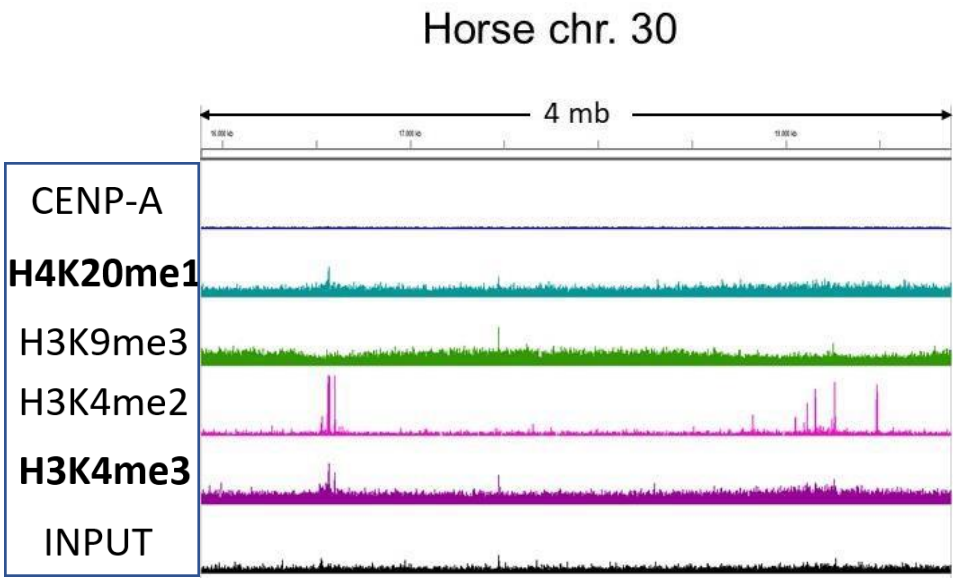
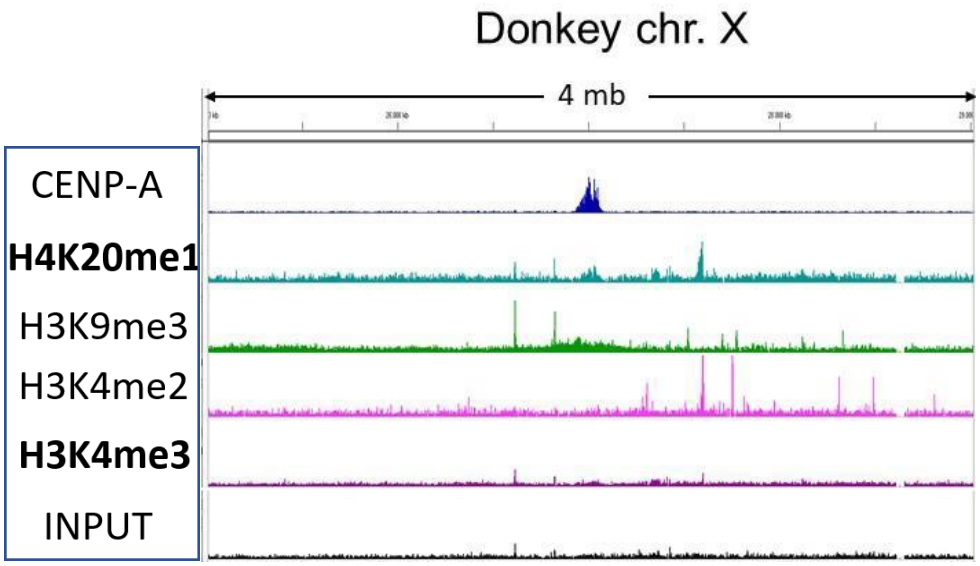


Figure R3-16: Epigenetic profile of the centromeric locus of donkey chromosome 30 on EquCabAsiA (panel A) and its orthologous non centromeric region on horse chromosome 30 on Equcab2.0 (panel B). *Colored tracks correspond to the different sequencing datasets obtained with the antibodies listed on the left. For all the datasets, unnormalized read count is plotted.*

Figure R3-16 shows the results of the analysis at the centromere of donkey chromosome 30 (panel A) and its orthologous region on horse chromosome 30 (panel B).
We detected a ~6 Mb wide heterochromatic region that surrounds the CENP-A binding domain.
An H4K20me1 signal (~100 kb) is present on the corresponding region of the centromeric domain; another H4K20me1 enriched region is also present in the transcription site of a predicted gene (KCTD3) which interrupts the heterochromatic domain on the left.
H3K4me3 is absent from the centromeric locus, which is transcriptionally silent, as well as in the horse orthologous non centromeric region.
In the horse, there is a slight notable enrichment in the H3K9me3 marker, but it hard to distinguish it from the background, but we still could identify the active transcriptional site of the predicted KCTD3 gene.

The centromere of donkey chromosome X

A



B

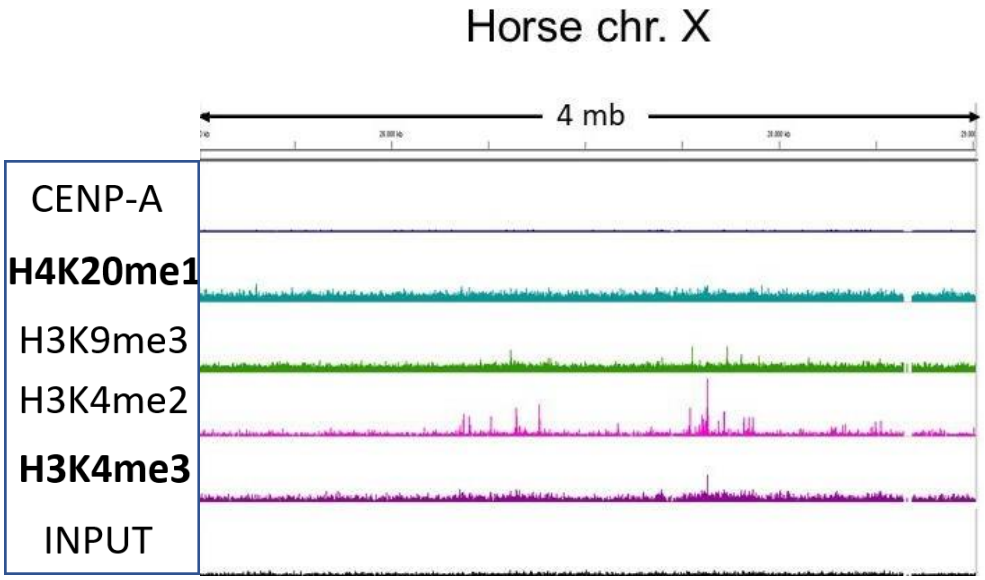


Figure R3-17: Epigenetic profile of the centromeric locus of donkey chromosome X on EquCabAsiA (panel A) and its orthologous non centromeric region on horse chromosome X on Equcab2.0 (panel B). *Colored tracks correspond to the different sequencing datasets obtained with the antibodies listed on the left. For all the datasets, unnormalized read count is plotted.*

Figure R3-17 shows the results of the analysis at the centromere of donkey chromosome X (panel A) and its orthologous region on horse chromosome X (panel B).
At this centromere we roughly identified a putative heterochromatic region of 500 kb around the CENP-A peak. The signal may be lower because only one copy of the X chromosome is present in this cell line, which derives from a male individual.
An H4K20me1 peak (~100 kb) is detectable in correspondence of the CENP-A binding region, which is absent in the horse orthologous region.
H3K4me3 seems to not to be present similarly to the other transcription marker.
In the horse this region lacks the heterochromatic H3K9me3 marker; moreover, neither H3K4me3 nor H4K20me1 were detected.

DISCUSSION

The centromeric and pericentromeric regions of satellite-based centromeres are characterized by the presence of extended arrays of satellite repeats and therefore very difficult to assemble. As far as satellite-free centromeres are concerned, the identification of the centromeric core (CENP-A binding domain) is easier but the exact limits of the pericentromere cannot rely on any particular DNA sequence feature. Through a ChIP-seq analysis of the H3K9me3 histone modification, marker for constitutive heterochromatin, we demonstrated that heterochromatic domains surround all the 16 donkey satellite-less centromeres (Riccardo Gamba PhD thesis 2017) as well as the one of horse chromosome 11. For this reason, satellite-free centromeres offer a good model to unravel the epigenetic landscape of the centromeric and pericentromeric chromatin.

Our data demonstrate that the extension of pericentromeric heterochromatic domains is highly variable: from 100-200 kb, such as for donkey chromosome 27, up to several megabases, such as 3.5 Mb at donkey chromosome 5 and 6 Mb at donkey chromosome 30. For other centromeres, where sequence amplification is present (such as in donkey chromosome 16), the precise size of the heterochromatic region could not be assessed.

The presence of the heterochromatic domain surrounding the centromere strongly suggests that it is a requirement for the centromeric function. This observation is consistent with previous studies demonstrating that the heterochromatic state facilitates CENP-A binding in both *Drosophila* and humans (Henikoff S et al. 2001).

In some cases, the orthologous non centromeric regions in the donkey or in the horse are also enriched for the heterochromatic marker. For example, the satellite-less centromere of horse chromosome 11 and its orthologous region on donkey chromosome 13 are comprised within a ~2.8 Mb region enriched for H3K9me3. Centromere of donkey chromosome 30 and its orthologous non-centromeric sequence display the same profile. This intriguing observation suggests that these genomic regions were already embedded in heterochromatin before neocentromere formation. However, other centromeric regions do not share the same pattern with their orthologous non centromeric regions. Therefore, it is possible that a heterochromatic environment may favour neocentromere formation. However, in some cases, the centromerization process may induce the establishment of a heterochromatic domain.

In many cases, heterochromatic boundaries seem to be limited by transcribed flanking regions, as for donkey chromosome 12, donkey chromosome 7 and horse chromosome 11.

Interestingly, in donkey chromosome 11 and donkey chromosome 30, the heterochromatic domain is interrupted by the transcription of predicted genes, *KCTD3* and *UFM1*, respectively. These genes map near the CENP-A binding

domain. The evidence of transcription at these sites, in both species, was provided by RNA-seq, H3K4me2, H3K36me2 and RNApol II ChIP-seqs as previously described in our laboratory (Riccardo Gamba PhD thesis 2017).
The H3K4me2 marker is absent in the majority of the neocentromeres in both species, suggesting that this feature is not required. These results are in agreement with the previous observation that satellite-less human clinical neocentromeres (Alonso A et al. 2010) and chicken evolutionary new centromeres (Hori T et al. 2014; Bailey AO et al. 2015) lack markers of "open" and transcriptionally active chromatin, such as H3K36me2 and H3K4me3. However, several authors proved that these markers are interspersed with CENP-A nucleosomes, at satellite containing centromeres of humans and *Drosophila* (Sullivan BA and Karpen GH 2004; Bailey AO et al. 2015). Moreover, at centromeres of donkey chromosomes 9, 16 and 19, it seems that they are closely surrounded by open chromatin. We previously demonstrated that, at these centromeres, the CENP-A binding fragment is amplified in the donkey (Nergadze SG et al. 2018), therefore, the size of the actual genomic sequence enriched for the various markers is larger than the one we can estimate by mapping the reads on the corresponding horse single copy sequence.
In this thesis I mainly analyzed two other markers (H3K4me3 and H4K20me1) whose analysis was not presented in Riccardo Gamba PhD thesis (2017).
H3K4me3 is an histone modification that has been typically associated with transcriptionally active chromatin in vertebrates and, in particular, to transcription start sites of active genes (Ruthenburg AJ et al. 2007). The analysis revealed that satellite-less centromeres are not enriched for this marker, as expected, since its enrichment is associated to active genes. On the contrary, regions which were proved to be transcribed outside the centromeric domain are enriched for this marker. Another interesting aspect is that also the orthologous non-centromeric regions are not enriched for H3K4me3, confirming the same trend of the H3K4me2 and H3K36me2. Moreover, its general enrichment pattern profile is more similar to those obtained with the RNApol II ChIP-seq (Riccardo Gamba PhD thesis 2017), according to the fact that these genomic regions are actively transcribed. Interestingly, in donkey chromosome 11 a transcription site within the heterochromatic region, next to the CENP-A peak is present. The H3K4me3 marker is enriched in correspondence of this site.
We can conclude that H3K4me3, similarly to the other markers for open and transcribed chromatin, is not necessary for the centromeric function.
In vertebrates, the H4K20me1 histone modification was previously showed to be present in the centromeric nucleosomes that contain the CENP-A histone variant (Hori T et al. 2014; Bailey AO et al. 2015). Studies on the chicken neocentromeres (Hori T et al. 2014) showed that this histone modification is fundamental for the kinetochore assembly and centromeric function. They found that H4K20me1 marker, although being present genome-wide, was particularly enriched at the

neocentromeric regions, colocalizing with the CENP-A nucleosomes.
In our model we observed a different distribution of this marker. We observed that this histone modification exactly co-localizes with the CENP-A binding domains in most of the satellite-less centromeres, as in the case of donkey chromosome 10, even if this is not the genomic region with the highest enrichment. On the other hand, the orthologous non centromeric regions lack this marker. Interestingly, the region corresponding to the H3K9me3 heterochromatin environment, around the CENP-A binding domain, shows an overall "de-enrichment" (described as signal depression) of H4K20me1 level compared to the surrounding region.
We advance the hypothesis that, although the centromeric core strictly requires the presence of H4K20me1, lysine 20 of histone 4 has to be demethylated within the surrounding pericentromeric domain. This modification may contribute to the stabilization of the pericentromeric boundaries.
H4K20me1 modification is also present at genome-wide level and, in accordance with the work of Beck DB et al. (2012). There is a correspondence between gene transcription and H4K20me1 enrichment, as visible for example, in donkey chromosome 12 and 14.
In conclusion, in agreement with previous data (Hori T et al. 2014) we demonstrate that the H4K20me1 histone modification is coupled with the centromeric function. In addition, thanks to our model system, we can propose that this marker is absent before the establishment of the centromere.

PART 4 ECTOPIC CENP-A BINDING SITES

RESULTS

We know that centromere protein A (CENP-A) is one of the most important determinant of centromere function (Palmer DK et al. 1991; Mendiburo MJ et al. 2011). CENP-A is a histone H3 variant occupying the centromeric locus, but it is also present on non-centromeric loci throughout the genome. Very little or none is known about the possible role of this protein on loci other than the centromere (Bodor DL et al. 2014). However it may play a role on the epigenetic control of some molecular pattern, as for gene expression.
In this work, we isolated and analyzed CENP-A binding sites mapping outside the centromeres, called ectopic or secondary CENP-A binding sites. As we identified satellite-less centromeres through ChIP-seq using an antibody against CENP-A on horse chromatin we also identified signals present in different genomic location other than the neocentromere on horse chromosome 11. It was already known that CENP-A binds the centromeric DNA, but also that CENP-A nucleosomes are present all across the genome (Bodor et al. 2014). The centromeric function determines the abundance of the histone H3 variant distributed on the DNA sequence unit (Bodor DL et al. 2014). Centromeric domains are highly enriched for CENP-A containing nucleosomes, although normal H3 histone variants are still present on the centromeric domain (Blower MD et al. 2002; Sullivan BA and Karpen GH 2004; Sullivan LL et al. 2011). Furthermore CENP-A nucleosomes, tend to localize to transcription factor genome sites in human cancer genome, possibly involved in gene expression and regulation (Athwal RK et al. 2015) and they tend to nucleate around heterochromatin spots (Gonzalez M et al. 2014).
Presence of CENP-A throughout the genome was verified with this experimental setup: we performed a peak calling with MACS software (Zhang Y et al. 2008) using reads from the ChIP and from the input to identify enriched non-centromeric CENP-A binding sites (as described in Material and Methods).

Secondary CENP-A peaks mapping and identification

ChIP-seq reads obtained from skin primary fibroblasts from four different horses (HSF-C, Sparky, CrowdPleaser, Locketaway) and one donkey (Asino Nuovo) using an anti-CENP-A antibody, were mapped on the horse reference genome EquCab2.0. We performed peak calling with MACS software (Zhang Y et al. 2008) using reads from the ChIP and from the input to identify enriched non-centromeric CENP-A binding sites. The majority of mammalian centromeric regions are not assembled at chromosome level because of the high amount of satellite DNA sequences in the regions. Instead those sequences are unmapped and

placed into an extrachromosomal contig called Unplaced Chromosome (chrUN). Because of a satellite-less centromere in the horse we excluded from the analysis CENP-A peaks enriched in the centromeric region of horse CHR11 as well as chrUN for the horse dataset, while 16 centromeric regions and the chrUN where excluded from the donkey dataset [Table R4-1].

For the HSF-C horse (thereafter referred as "HorseC"), two experimental replicas were done. 43.4 million and 70.2 million of paired-end reads were obtained respectively from the first and from the second replicas. After peak-calling and subsequent trimming of primary CENP-A sites, we obtained 352 and 771 secondary CENP-A peaks from the first and second replica, respectively.

From Sparky, CrowdPleaser and Locketaway (three different horse individuals) fibroblast cell lines 42.9 M, 34.3 M and 39.5 M of paired-end reads respectively were obtained. Reads from these three datasets were pooled together since their ChIP efficiency was low. Read mapping, peak-calling and centromeric region trimming produced 146 ectopic CENP-A peaks in this read mixed dataset (referred to as "pooled horses dataset").

From the ChIP-seq experiment performed on Asino Nuovo fibroblasts (referred to as "donkey"), used for the previous work (Part 1, Part 2, Part 3 and Nergadze et al. 2018), 28.9 M of paired-end reads were obtained allowing us to identify 988 CENP-A secondary binding sites.

SPECIES	CHROMOSOME (EquCab2.0)	REGION REMOVED
HORSE	ECA11	27,643,400-28,050,000
DONKEY	EAS4 (ECA28)	12,869,491-13,056,914
	EAS5 (ECA19)	4,926,684-5,182,243
	EAS7 (ECA8)	41,958,140-42,151,138
	EAS8 (ECA20)	26,385,695-26,527,108
	EAS9 (ECA14)	29,651,201-29,706,482
	EAS10 (ECA25)	8,591,820-8,847,027
	EAS11 (ECA17)	16,758,813-16,871,575
	EAS12 (ECA9)	31,949,336-32,311,449
	EAS13 (ECA11)	46,658,369-46,919,553
	EAS14 (ECA13)	7,243,859-7,524,156
	EAS16 (ECA5)	74,884,905-74,962,292
	EAS18 (ECA26)	22,371,639-22,525,907
	EAS19 (ECA6)	14,191,812-14,254,865
	EAS27 (ECA27)	19,637,493-19,968,661
	EAS30 (ECA30)	17,713,606-17,826,113
	EASX (ECAX)	26,952,446-27,097,266

Table R4-1: Chromosomal regions removed from the HorseC replicas, pooled horses and donkey CENP-A datasets.

To further investigate on the CENP-A role in genome wide positioning on different mammal species we decided to use also ChIP-seq datasets retrieved from the Sequence Read Archive (SRA) obtained from mouse and human cells.
Reads obtained (Iwata-Otsubo A et al. 2017) from a ChIP-seq experiment, using an anti-CENP-A antibody on mouse liver chromatin, were mapped on the mouse reference genome (mm10). The same pipeline analysis, applied for the datasets above, was used for this dataset. We were able to identify 309 ectopic CENP-A peaks in the mouse reference genome after deleting the peaks mapped on the Unplaced Chromosome.
We searched for CENP-A ChIP-seq datasets also on human noncancerous cell lines. A data set was found and analyzed (Hayden KE et al. 2013) however

revealed no ectopic CENP-A sites. We then analyzed a CENP-A ChIP-seq dataset obtained from HeLa human cancer cell line (Lacoste N et al. 2014). These authors performed two experimental replicas obtaining 85.9 M and 112.4 M of paired-end reads, respectively. We proceeded into the data analysis as above. After read mapping on the human reference genome (hg38) and peak calling, satellite-based centromeric regions (assembled in the hg38 reference genome) and chrUN were removed from the analysis [Table R4-2]. We identified 3615 and 4250 ectopic CENP-A peaks from the two replicas. The number of secondary CENP-A binding sites in all datasets is reported in Table R4-3.

SPECIES	CHROMOSOME (hg38)	REGION REMOVED
HUMAN	CHR1	121700000-143308083
	CHR2	90402511-96000000
	CHR3	87800000-94000000
	CHR4	48200000-51800000
	CHR5	46100000-51400000
	CHR6	58500000-62600000
	CHR7	58100000-62506779
	CHR8	43200000-47200000
	CHR9	42200000-60518558
	CHR10	38000000-41600000
	CHR11	50871348-55800000
	CHR12	33200000-37800000
	CHR13	16500000-18900000
	CHR14	16100000-18200000
	CHR15	17500000-20500000
	CHR16	35300000-46280682
	CHR17	22700000-27400000
	CHR18	15400000-21500000
	CHR19	24200000-28100000
	CHR20	25700000-30400000
	CHR21	10814560-13000000
	CHR22	13700000-17400000
	CHRX	58100000-63800000

Table R4-2: Chromosomal regions removed from the HeLa CENP-A datasets.

SPECIES	ANIMAL	CELL LINE	ANTIBODY	REPLICA	N. OF CENP-A PEAKS	N. READS
HORSE	HSF-C	fibroblast	anti-CENP-A	1	352	43378388
				2	771	78207302
	Sparky*	fibroblast	anti-CENP-A			
	Crowd*	fibroblast	anti-CENP-A	1	146	125065618
	Locket*	fibroblast	anti-CENP-A			
DONKEY	AsinoNuovo	fibroblast	anti-CENP-A	1	988	28937922
MOUSE	Mouse (Iwata-Otsubo et al. 2017)	liver	anti-CENP-A	1	309	13900000
HUMAN	Human (Lacoste et al. 2014)	HeLa	anti-CENP-A	1	3615	85900000
				2	4250	112400000

Table R4-3: Description of the raw data used to perform the work described here. * Datasets pooled together.

Figure R4-1 shows an example of ectopic CENP-A peaks on the horse sample compared to the centromeric peak on horse chromosome 11. Ectopic CENP-A peaks tend to have a Gaussian-like shape read distribution, however read count is much lower when compared to centromeric regions.

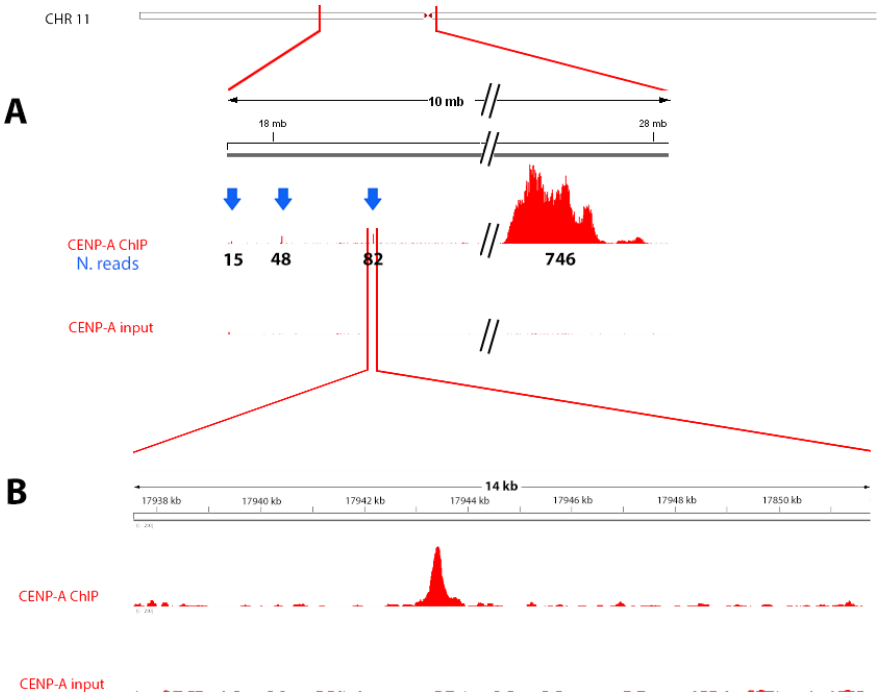


Figure R4-1: Representation of a typical secondary CENP-A peak compared to a centromeric peak: *A) 10 Mb region of the horse chromosome 11 containing three secondary CENP-A peaks (pointed by blue arrows) and the peak corresponding to the satellite-less centromere (from 27,643,400 bp to 28,050,000 bp). The number under each peak is the number of reads that aligned on that region. B) Gaussian-like shape of one secondary peak is visualized.*

Localization of secondary CENP-A binding sites

Here three examples of chromosomal representation are reported to which all the datasets were mapped on. Four chromosome examples showing the localization of ectopic CENP-A of Horse replica 1, Horse replica 2, pooled horses dataset and donkey (since they were all mapped on the horse reference genome Equcab2) [Figure R4-2]. Five chromosome examples showing the localization of ectopic CENP-A for the mouse dataset [Figure R4-3]. Three chromosome examples showing the localization of ectopic CENP-A of HeLa replica 1 and HeLa replica 2 [Figure R4-4].

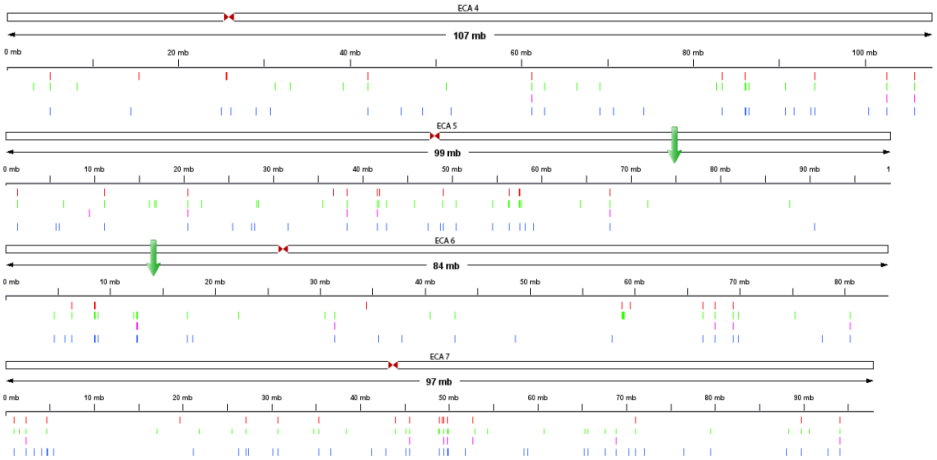


Figure R4-2: Ectopic CENP-A mapping on horse chromosomes 4, 5, 6 and 7. *Secondary CENP-A peaks are represented as coloured lines: red for HorseC replica 1, green for HorseC replica 2, pink for the pooled horses dataset and blue for donkey dataset. Horse centromeres are shown as head-to-head red arrows, donkey neocentromeres as green arrows.*

Figure R4-2 shows the distribution of ectopic CENP-A on horse chromosome 4-5-6-7. CENP-A is scattered across all the chromosomes in all the four datasets analyzed. To evaluate the conservation of the peaks among different replica and from different datasets we took advantage of the BED tools software (Quinlan AR and Hall IM 2010).

Table R3-4 reports the number of shared and conserved CENP-A peaks in each dataset. The number of conserved peaks mapping in genic loci is indicated within parenthesis. The dataset comparison between HorseC replica 1 (352 peaks) and HorseC replica 2 (750 peaks) showed that 266 peaks are shared. Among the shared peaks, 103 mapped in sequences related to genes. In the pooled horse dataset out of 146 peaks, 100 are shared between the two replicas of the HorseC datasets. 48 of these peaks are within genic regions. Donkey and HorseC replica 1 shared 257 peaks (105 within genes), while donkey and HorseC replica 2 shared 429 peaks

(182 within known genes). The pooled horses dataset shared 122 peaks with donkey (60 within known genes). Among the two HorseC replicas, the pooled horses dataset and the donkey, 95 out of 100 peaks were found in common (47 within known genes). This comparative analysis demonstrates that a great proportion of secondary CENP-A peaks, as highlighted by Figure R4-2, is shared not only between replicas but also between *Equus* species.

Number of peaks		Number of conserved secondary peaks (within known genes)
horseC replica 1	352	266 (103)
horseC replica 2	771	
horseC replica 1+2	266	100 (48)
pooled horses dataset	146	
donkey	988	257 (105)
horseC replica 1	352	
donkey	988	429 (182)
horseC replica 2	771	
donkey	988	122 (60)
pooled horses dataset	146	
donkey	988	95 (47)
horseC replica1+2	100	
pooled horses dataset		

Table R4-4: Comparison of the CENP-A peaks found in HorseC replicas, pooled horses dataset and donkey.

Figure R4-3 shows the ectopic CENP-A sites of the mouse dataset mapped on mouse chromosome 1-2-3-4-5. Similarly to the horse and donkey results, the 309 secondary CENP-A binding sites of this mouse dataset are scattered along all chromosomes.

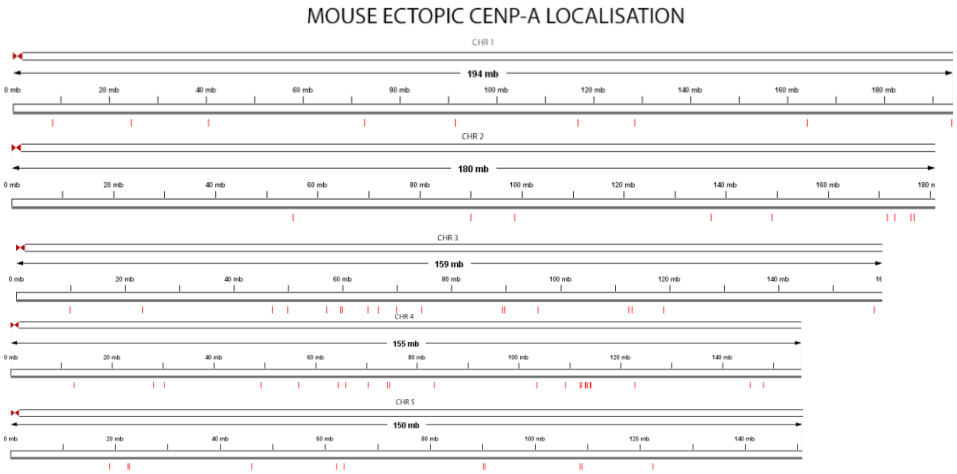


Figure R4-3: Ectopic CENP-A mapping on the mouse chromosome 1, 2, 3, 4 and 5 (mm10). *Secondary CENP-A peaks are represented as red lines. The red head-to-head arrows indicate the position of the mouse centromeres.*

Figure R4-4 shows the secondary CENP-A peaks of HeLa cells mapped on human chromosome 1-2-3. A high number of ectopic CENP-A peaks is located in the pericentromeric regions of most chromosomes as shown in Figure R4-4. Authors of this work (Lacoste N et al. 2014) reported that CENP-A density decreases with centromeric distance, possibly due to CENP-A spreading out of the centromeric core. However, we also detected genome wide ectopic CENP-A. Based on comparison of peaks between the two replicas we established that 2036 out of 3615 secondary CENP-A peaks of replica one are shared with replica 2.

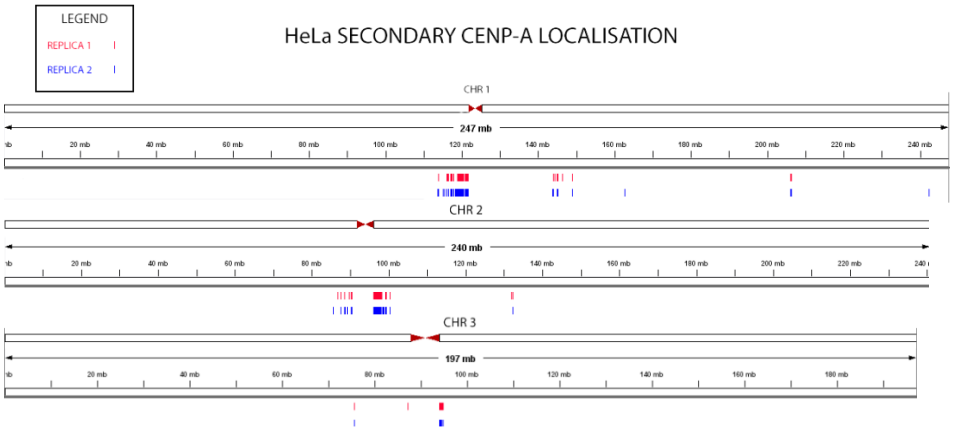


Figure R4-4. Ectopic CENP-A localisation on the human chromosome 1, 2 and 3 (hg38). *Secondary CENP-A peaks are represented as coloured lines red lines for HeLa replica 1, blue lines for HeLa replica 2. The red head-to-head arrows indicate the position of the human centromeres.*

Characterization of loci bound by CENP-A

Ectopic secondary CENP-A was reported to preferentially map on promoter-TSS regions of active genes in cell lines overexpressing the CENP-A protein (Athwal RK et al. 2015). So, we investigated whether ectopic CENP-A also preferentially binds some functional genomic region in our datasets using cell lines expressing normal levels of CENP-A. We used the HOMER software (Heinz S et al. 2010), a ChIP-seq annotation tool, to run a comparison analysis of ectopic CENP-A peaks with respect to the nearby genes and/or regulatory elements within the tested genome. HOMER provides an annotation for each peak as output, indicating their location with respect to exons, introns, promoter-TSS regions (defined as the region that starts from -1000 bp from the TSS site and ends to +100 bp), Transcription Terminating Site regions (TTS, defined as the region that starts from -100 bp and ends to +1000 bp around the TTS). Homer also identifies intergenic peaks and the distance to the closest gene.

Table R4-5 reports the peak summary analysis. For each species analyzed, numbers and frequencies of secondary CENP-A peaks found in each genomic region is reported. HorseC replica 1 and replica 2 data were pooled together since they share the great majority of the peaks.

Pooled horse dataset comprises three horses (Sparky, CrowdPleaser and Locketaway), while HorseC datasets where analyzed separately.

	NUMBER OF CENP-A PEAKS (%)					
	PROMOTER-TSS	EXON	INTRON	TTS	INTERGENIC	TOTAL
HORSEC	60 (7.1)	24 (2.8)	245 (29.0)	11 (1.3)	504 (59.7)	844
POOLED HORSES DATASET	21 (14.4)	4 (2.7)	38 (26.0)	2 (1.4)	81 (55.5)	146
DONKEY	63 (6.4)	28 (2.8)	228 (23.1)	16 (1.6)	653 (66.1)	988
MOUSE	14 (4.5)	3 (1.0)	84 (27.2)	3 (1.0)	205 (66.3)	309
HeLa REPLICA 1	301 (8.3)	261 (7.2)	727 (20.1)	68 (1.9)	2258 (62.5)	3615
HeLa REPLICA 2	322 (7.6)	320 (7.5)	956 (22.5)	76 (1.8)	2576 (60.6)	4250

Table R4-5: Analysis of CENP-A peaks with respect to nearby protein-coding genes. *The number of peaks is organized based on the genomic regions assigned by HOMER. Within parenthesis, peak frequency (%) is reported, which is calculated as the number of peaks in that region divided by the total number of peaks found in that dataset.*

The expected frequency of peaks in each type of genomic region was calculated as the fraction of the genome occupied by such sequences. Figure R4-5 shows the frequencies of ectopic CENP-A peaks in the HorseC related to each genomic region compared to the expected values.

Since the HOMER annotation requires a gene dataset we used the Ensembl Genes dataset for the analysis done on the datasets mapped on the horse genome. We did not use the NCBI RefSeq horse gene dataset mainly for two reasons:

1. Because the horse RefSeq annotation contains very few entries
2. Because Ensembl annotation is more suitable when conducting a more exploratory research (Zhao S and Zhang B 2015)

When choosing an annotation database, it's important to keep in mind that there is no perfect database. Some could miss gene annotations, others may overestimate the number of genes or transcripts. Based upon RNA-Seq data analysis (Zhao S and Zhang B 2015), if robust gene expression estimation must be done (e.g. clinical data), databases such as RefGene are preferred, since they are less complex genome annotations. In our case (a more pilot and exploratory driven study) we could miss some information if databases such as RefGene were used, especially because the low level of ectopic CENP-A. For this reason we chosen the Ensembl annotation database. It has more entries, so we can get more information.
Only one isoform per gene was maintained in the gene dataset (the isoform with the highest number of exons). Ectopic CENP-A peaks in the HorseC bind the promoter-TSS region way more frequently than the expected (7.1% vs 0.9%, p-value = $1.2x10^{-81}$), almost 7 times more. Also, promoter regions showed higher frequencies in CENP-A binding (2.8% vs 1.4%, p-value = $2.8x10^{-04}$). Frequency of intergenic peaks is significantly lower than the expected value (59.7% vs 68.2%, p-value = $3.5x10^{-07}$), probably due to the downscaling factor gave by promoter-TSS and exons regions [Figure R4-5]. The most likely interpretation of this finding is that CENP-A may play a role in gene regulation since it preferentially binds promoter regions.

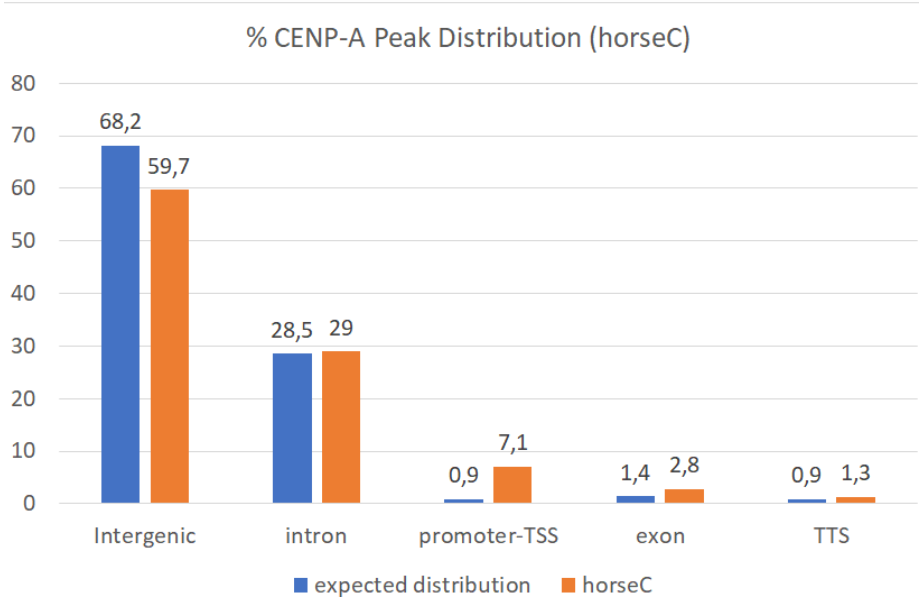


Figure R4-5: Distribution of secondary CENP-A in HorseC. *HorseC secondary peaks analyzed by HOMER (Heinz S et al. 2010). * and **, statistically significant differences referred to the expected distribution with a p-value < 10^{-3} and < 10^{-81}, respectively.*

The same analysis performed on the pooled horses dataset (14.4% vs 0.9%, p-value = $1.1x10^{-66}$) an on the donkey (6.4% vs 0.9%, p-value = $5.2x10^{-91}$) showed same enrichment profile when comparing ectopic CENP-A to promoter-TSS regions (data not shown).

For the mouse analysis, the GENECODE VM9 (Ensembl 84) dataset was used. Only one isoform per gene was maintained in the gene dataset (the isoform with the highest number of exons). As observed in the *Equus* species, in mouse liver cells the frequency of ectopic CENP-A peaks in the promoter-TSS region is higher than expected (4.5% vs 0.9%, p-value = $1.4x10^{-11}$). Other frequencies were similar to the expected ones (data not shown).

For the HeLa CENP-A datasets, the NCBI RefSeq Genes dataset was used since in the human species this annotation dataset is well assembled and revised. Only one isoform per gene was maintained in the gene dataset (the isoform with the highest number of exons). In both replicas of HeLa cells, the frequency for the promoter-TSS region is more than 10-times higher than the expected distribution. The great number of peaks obtained with these samples raises the statistical significance of the observed difference (replica 1 and 2 respectively 8.3% and 7.6% vs 0.7% of the expected distribution, p-values <<< 10^{-134}); enrichment also in the exon regions was observed. Intergenic peak distribution was similar to the expected values.

However peak distribution for the intronic regions is lower compared to the expected distribution, maybe due to a consequence of the high frequencies observed in the promoter-TSS and exon categories. (Data not shown).
We then performed a positional analysis of the ectopic CENP-A peaks respect to the promoter-TSS region to test whether CENP-A containing nucleosomes preferentially reside in a specific DNA position. Again, through the HOMER software tool, we performed a positional analysis "Distance from TSS". Figure R4-6 shows the plot of the distance in bp of each peak from the closest TSS. In all samples the graphs show a greater concentration of peaks in the -1000 +1000 bp around TSSs compared to more distant regions.

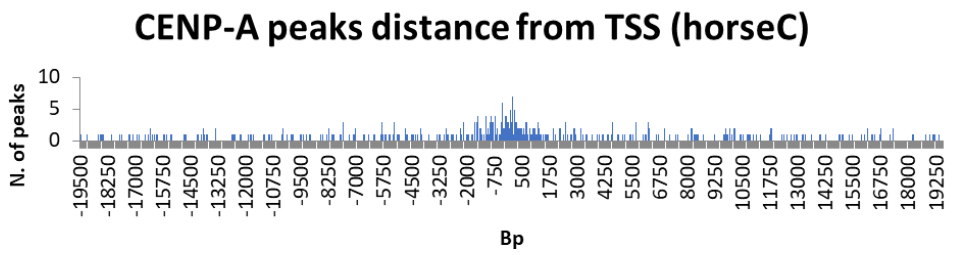


Figure R4-6. CENP-A peaks distribution around TSS in the HorseC. *This plot represents the distribution of the CENP-A peak respect to the closer TSS in the HorseC. On the X-axis we have the distance measured in base pairs, while on the Y-axis we have the number of peaks sharing the same distance from TSS.*

There is clearly a greater concentration of ectopic CENP-A peaks -1000 +1000 bp around the TSSs. Same results were obtained when the same analysis was done on the pooled horse dataset, on donkey dataset and on both HeLa cells replicas. A clear result was hard to establish in the mouse probably due to the low number of total ectopic CENP-A peaks, or to the annotation database used (Data not shown).

Gene expression analysis

We focused then on the task of revealing any type of connection between ectopic CENP-A and specific gene expression. Using the ATLAS database (https://www.proteinatlas.org/) we obtained the expression profiles from 37 human tissues (Uhlen M et al. 2015) and 13 mouse tissues (Söllner JF et al. 2017) and calculated the expected frequencies of genes belonging to the following classes: 1) tissue-specific genes, 2) genes expressed in all tissues, 3) genes with a mixed pattern of expression and 4) genes whose expression was not described in the database. We then analyzed the expression pattern of the genes associated with CENP-A sites, calculated their frequencies and compared them with the expected frequency values. When more than one CENP-A binding site was associated to a

single gene, this gene was counted only once. For the horse, donkey and human analysis we used the human ATLAS proteome database (due to lacking same data for the equids and since high protein sequence homology between species), while for the mouse analysis we used the mouse ATLAS proteome database. Genes were also classified according to the genomic position of the CENP-A peaks (i.e. promoter-TSS, exon, intron, TSS, and "more than 1 peak per gene" genes that have more than 1 peak along their sequences).

Gene expression analysis in the HorseC and donkey

The results of the gene expression analysis in the HorseC and donkey are reported in Figure R4-7. A chart for each genomic region was created. In each chart the frequencies of each expression profile are shown, calculated as the number of genes with that particular expression profile divided by the total number of genes found in the same genomic region. As an example, the first chart on the upper left corner represents the frequencies of each expression profile calculated on HorseC (orange) and donkey (grey) genes that have at least 1 ectopic CENP-A peak in their promoter-TSS region.

We only reported cerebral cortex and testis tissue-specific genes since very few genes, bound by ectopic CENP-A, were expressed specifically in other tissues. Therefore, we classified the gene expression as follows: cerebral cortex, testis, other tissues, expressed in all, not detected and mixed. In all the considered genomic categories, except for the TTS region, we have found that genes expressed exclusively in the human cerebral cortex are significantly more represented than the expected distribution. While the expected frequency of the cerebral cortex's genes is 4.4%, in the promoter-TSS category for both species we found frequencies of 28.3% (HorseC, p-value = $8.8x10^{-17}$) and 22.4% (donkey, p-value = $2.6x10^{-09}$); in the exon category for both species we found frequencies of 29.2% (HorseC, p-value = $1.8x10^{-04}$) and 50.0% (donkey, p-value = $1.8x10^{-20}$); in the intron category, for both species we found frequencies of 31.9% (HorseC, p-value = $4.4x10^{-68}$) and 25.0% (donkey, p-value = $1.3x10^{-38}$); in the "more than 1 peak per gene" category for both species we found frequencies of 34.6% (HorseC, p-value = $2.0x10^{-07}$) and 25.9% (donkey, p-value = $2.0x10^{-04}$).

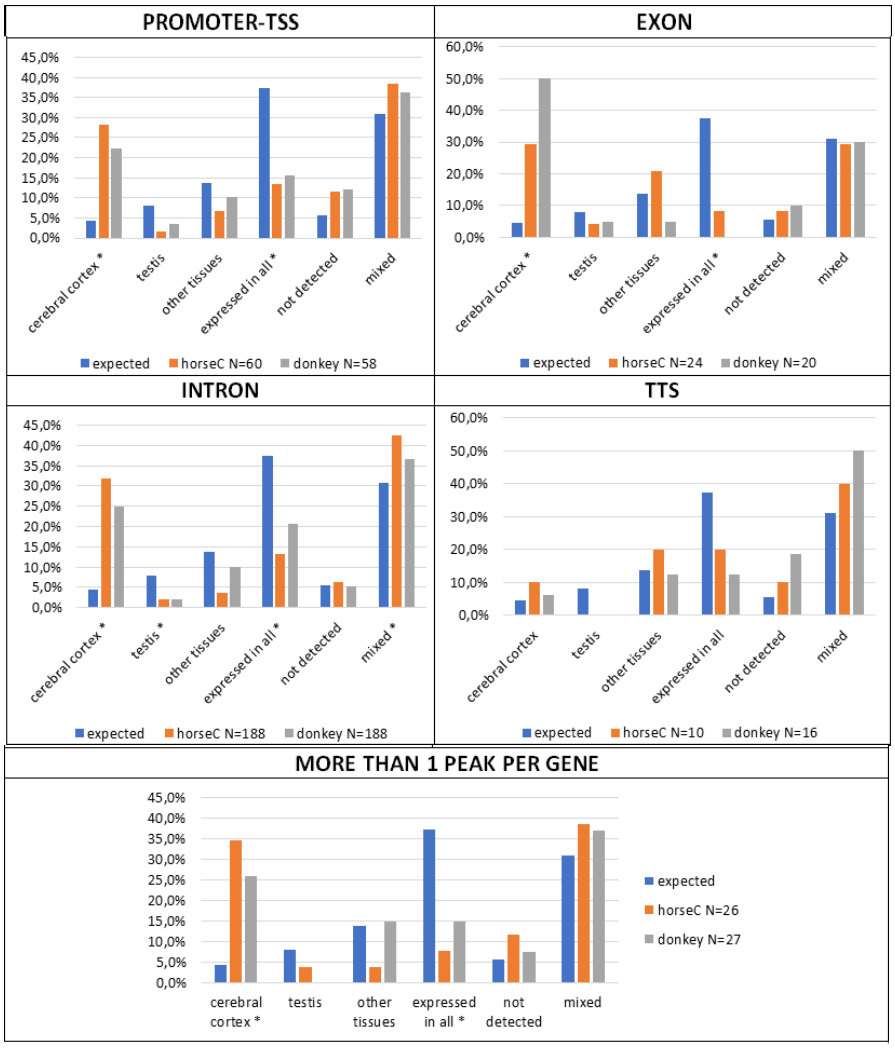


Figure R4-7: Gene expression analysis for the secondary CENP-A peaks present in the HorseC and donkey. *In these charts, the frequencies of expression profile of genes associated to each CENP-A category are shown (horse C, orange bars; donkey, grey bars). The "N" present in the legend represents the number of genes present in that functional genomic region. The expression profiles marked with * are significantly different from the expected values (blue bars).*

We performed the same gene expression analysis (data not shown) in the pooled horse dataset after combining all the genes from the different functional genomic regions (promoter-TSS, exon, intron and TSS regions) due to the low number of ectopic CENP-A peaks found. Genes expressed exclusively in the human cerebral cortex were found to be more represented than the expected (38.8% vs 4.4%, p-value = $1.4x10^{-26}$), while genes expressed in all tissues were less represented (6.3% vs 37.3%). We performed the same gene expression analysis (data not shown) in

the mouse dataset after combining all the genes from the different functional genomic regions (promoter-TSS, exon, intron and TSS regions) due to the low number of ectopic CENP-A peaks found. Genes expressed exclusively in the mouse brain were more represented than expected (17.2% vs 8.1%, p-value = $2.9x10^{-03}$), while no differences in other expression profiles were detected.
All these results indicate that, at least in these species, the ectopic CENP-A seems to be loaded preferentially at loci associated with genes which are exclusively expressed in the brain.
When we tested, with the same gene expression analysis, datasets from HeLa replica 1 and replica 2, we did not detect any association of ectopic CENP-A and brain-related genes. Since HeLa cells are immortalized tumoral cells, maybe ectopic CENP-A is not physiological regulated as it is in other cell lines. We performed a correlation analysis between ectopic CENP-A associated genes and proto-oncogenes and tumor-suppressor genes, however we obtained no significant results (data not shown).

Motif analysis

We proceeded on the identification of putative sequence motifs present within regions of 500 bp spanning the center of secondary CENP-A peaks. The *de novo* motif discovery tool of the HOMER software was used to perform this analysis. When we performed this analysis on the HorseC, donkey and pooled horse datasets, we identified a specific sequence motif which is common in all functional genomic regions bound by ectopic CENP-A. The identified motif is the REST binding motif, also called neuron-restrictive silencer element NRSE [Figure R4-8A]. The significance of this enrichment is reported in Table R4-6. Figure R4-8 shows the REST logo motif and the frequencies of secondary CENP-A binding sequences containing this motif compared to frequencies of random background sequences in the HorseC, in the donkey and in the pooled horse dataset. We used as statistically significant threshold a p-value < $1.0x10^{-50}$, as suggest by the HOMER manual. In the HorseC [Figure R4-8] we have a 266-fold enrichment of this motif within the promoter-TSS (p-value = $1.0x10^{-67}$), 400-fold in the intron (p-value = $1.0x10^{-473}$) and 356-fold in the intergenic (p-value = $1.0x10^{-67}$) regions. As shown in Figure R3-8C-D this motif was found to be also enriched in the pooled horse dataset (no significant enrichment for exon and TTS categories) and on the donkey dataset. The REST binding motif identified by HOMER software shared almost identical consensus homology the one previously described (Bessus A et al. 1997), as well as with those found in the TRANSFAC online motif database (http://gene-

regulation.com/cgi-bin/pub/databases/transfac/getTF.cgi?AC=T01974#2). Mouse and HeLa datasets however, did not exhibit enriched motifs with a statistically significant p-value.

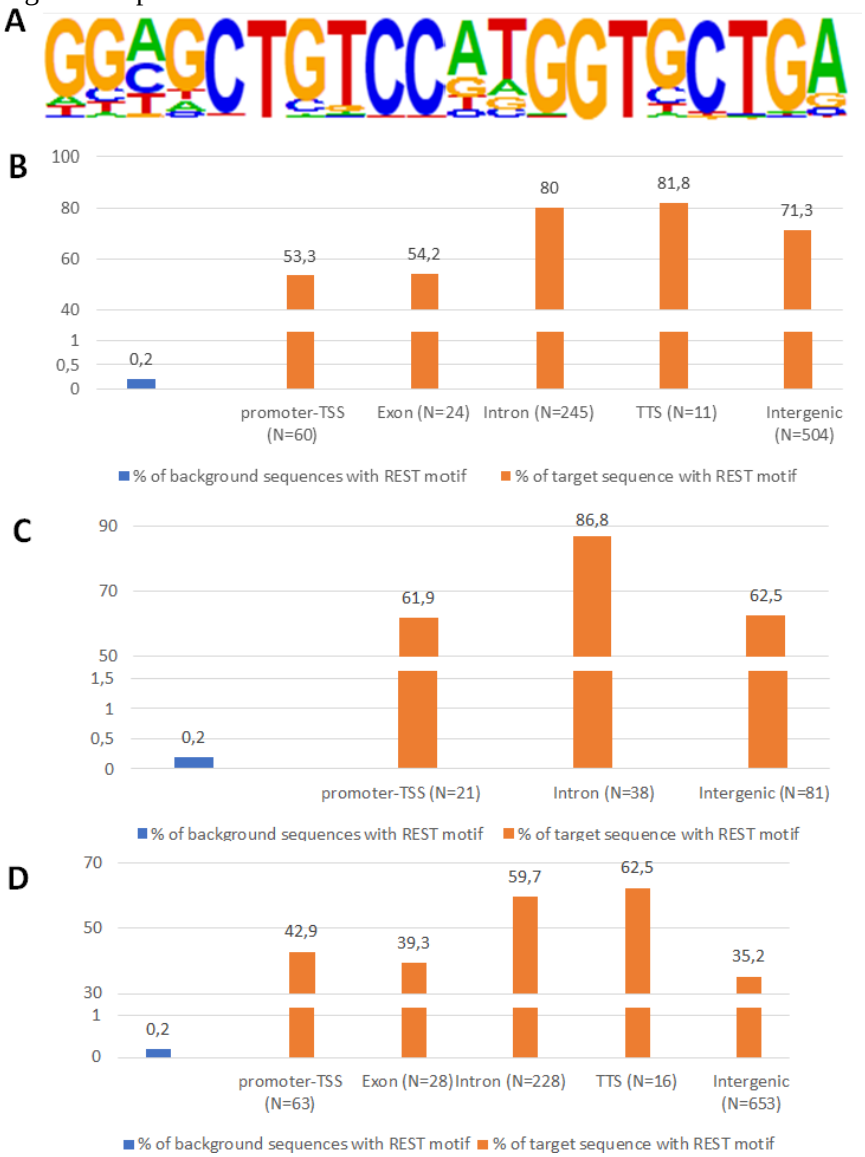


Figure R4-8: REST motif enrichment in the HorseC, pooled horses dataset and donkey secondary CENP-A binding sequences. *A) the logo is the motif bounded by REST protein. The dimension of each letter is proportional to the probability of finding that nucleotide in that position. B) For each genomic category considered, the frequencies of HorseC secondary CENP-A binding sequences containing the REST binding motif (orange bars) are compared to the frequency of random sequences that contain the REST binding motif (blue bar). C) The same comparison as in B, for the pooled horses dataset. D) The same comparison as in B for the donkey.*

HORSEC REST MOTIF ANALYSIS					
	p-value	N. target sequences with REST motif	% of target sequence with REST motif	N. background sequence with REST motif	% of background sequences with REST motif
promoter-TSS (N=60)	$1x10^{-67}$	32	53.3	102	0.2
Exon (N=24)	$1x10^{-27}$	13	54.2	114	0.2
Intron (N=245)	$1x10^{-473}$	196	80	102	0.2
TTS (N=11)	$1x10^{-21}$	9	81.8	107	0.2
Intergenic (N=504)	$1x10^{-863}$	357	71.3	80	0.2
POOLED HORSES DATASET MOTIF ANALYSIS					
promoter-TSS (N=21)	$1x10^{-28}$	32	61.9	102	0.3
Intron (N=38)	$1x10^{-84}$	196	86.8	102	0.2
Intergenic (N=81)	$1x10^{-118}$	357	62.5	80	0.2
DONKEY MOTIF ANALYSIS					
promoter-TSS (N=63)	$1x10^{-52}$	27	42.9	101	0.2
Exon (N=28)	$1x10^{-23}$	11	39.3	81	0.2
Intron (N=228)	$1x10^{-313}$	136	59.7	81	0.2
TTS (N=16)	$1x10^{-23}$	10	62.5	89	0.2
Intergenic (N=653)	$1x10^{-455}$	229	35.2	78	0.2

Table R4-6: REST motif analysis in the HorseC, pooled horses dataset and donkey secondary CENP-A binding sequences. *This table reports the interval of confidence for the REST motif enrichment (p-value), the number of secondary CENP-A binding sites containing the motif, the % of ectopic CENP-A binding sequences that contain the motif, the number of background sequences containing the motif and the % of background sequences that contains the motif for each genomic category considered.*

DISCUSSION

CENP-A is the main centromeric variant of the H3 histone protein, thus being the main epigenetic marker for the centromeric function (Drinnenberg IA et al. 2014; Fachinetti et al. 2013; Perpelescu M and Fukagawa T 2011). Since its importance during the cell cycle for proper chromosome segregation, most of the scientific studies addressed their effort on the characterization of this protein at the centromeric core. Very few studies focused on the identification and functional characterization of CENP-A in non-centromeric regions. Recent work suggested new secondary roles for this protein beside its centromeric function (Athwal RK et al. 2015; Lacoste N et al. 2014). For this reason we investigated the possible role of ectopic CENP-A in horse, donkey, mouse and human cell lines, expressing an endogenous level of this protein. A high number of non-centromeric relatively short (up to few hundreds of nucleotides) CENP-A peaks scattered on the genome was observed [Figure R4-1]. We identified very large number of ectopic CENP- A peaks, 352 in the "HorseC" replica 1, 771 in HorseC replica 2, 146 in the pooled horse dataset and 988 in donkey dataset. We also analyzed ChIP-seq reads obtained from mouse liver chromatin using an anti-CENP-A and we identified 309 ectopic CENP-A peaks. When we analyzed an anti-CENP-A ChiP seq dataset generated from HeLa human cancer cell, we identified 3615 and 4250 ectopic CENP-A sites respectively in two HeLa cells replicas.

Positional analysis revealed that secondary ectopic CENP-A peaks are scattered all over the genome in all the analyzed datasets, binding other genomic regions beside the centromeric one. So, we further confirmed that CENP-A is present not only in the centromeric regions but also at genome-wide level, in normal cell lines (plus HeLa cell line) expressing normal level of this protein. We also detected that a great majority of secondary CENP-A peaks is shared not only between replicas but also between *Equus* species (since their datasets were all mapped on the horse reference genome), suggesting an active not random process of CENP-A loading at specific loci. Ectopic CENP-A peaks found in the mouse dataset showed the same genome-wide distribution. In the HeLa cell datasets, the great majority of the ectopic CENP-A peaks localize near the centromeres as previously observed (Lacoste N et al. 2014). However, we were still able to find ectopic CENP-A peaks localized throughout the genome and we observed that HeLa replicas shared a high number of peaks. It will be interesting to compare secondary CENP-A peaks found in the horse, donkey and HeLa datasets through homology studies.

We thus investigated the positional localization of ectopic CENP-A peaks with respect to the nearby genes and/or regulatory elements since, as previously reported, CENP-A nucleosomes were found in the locations of promoter-TSS regions of active genes in cell lines overexpressing CENP-A protein (Athwal RK et

al. 2015; Lacoste N et al. 2014). Strikingly we observed that ectopic CENP-A binds promoter-TSS regions at a higher frequency compared to the expected value in all datasets analyzed from the different species. However, in the mouse, no clear enrichment around the TSS was found, possibly due to the low number of reads in this dataset, to the annotation database used, or to the fact that there is no enrichment at TSS in this species.

CENP-A containing nucleosomes are therefore distributed in the genomes of different species not randomly fashioned but with a precise localization: gene related sequences into the genome. Since CENP-A containing nucleosomes are conformationally different from H3 containing nucleosomes (Tachiwana H et al. 2011), their presence in these functional genomic regions may suggest that they could have a different role as compared to normal nucleosomes, for example being differently recognized by specific binding proteins. Moreover, we proved that ectopic CENP-A positioning is correlated to tissue-specific genes. Surprisingly, in the horses, donkey and mouse we found a statistically significant enrichment of genes expressed exclusively in the cerebral cortex in all the functional genomic regions bound by CENP-A. This is of important interest since these results were obtained from horse and donkey fibroblasts. These results clearly indicate that ectopic CENP-A is loaded in gene promoter regions which are not expressed in this cell type, at least in the species studied in this work. This is the opposite of what observed by Lacoste N et al. (2014), where they observed an enrichment of CENP-A peaks in actively transcribed gene regions. In the HeLa cell line, results indicate that ectopic CENP-A seem to behave differently from the other species here analyzed. We did not find an enrichment of brain related genes. However, we found a statistically significant under-enrichment for genes expressed exclusively in the human testis in both replicas. Nor an association between CENP-A and cancer- associated genes was clearly identified. However, we cannot exclude different results in non-cancer tissues.

Finally, we reported that all the functional genomic regions bound by ectopic CENP-A in the equid species are highly enriched in the NRSE sequence (neuron-restrictive silencer element). The NRSE motif is recognized by the REST protein, a Kruppel-type zinc finger protein that represses neuronal gene transcription in non-neuronal cells by recruiting an histone-deacetylase and EHMT2 methyltransferase, hence acting as a chromatin modifier (Mulligan P et al. 2008). It is important to point out that our ChIP-seq experiments with anti-CENP-A-antibody were performed on horse and donkey fibroblast cell lines, in which of course brain-related genes should be repressed. Since we have found a great association between secondary CENP-A binding sites and genes expressed exclusively in the cerebral cortex, ectopic CENP-A seems to be associated with genes repressed by the REST protein. These results suggest that ectopic CENP-A containing nucleosomes are preferentially recruited at brain-related genes, probably, contributing at their repression in non-neuronal cells, at least in the *Equus* species. A possible

explanation is that, nucleosomes which have the centromeric H3 histone variant, act as null nucleosomes, therefore being hardly recognized by RNA polymerase.
All together, these data show a correlation between CENP-A positioning on gene-related regulatory elements. Moreover, two distinct analysis (gene expression and motif discovery) directly suggested that CENP-A is associated to brain-related genes. Taking into account positional evidences of ectopic CENP-A at promoters of these genes, we can propose that CENP-A containing nucleosomes can act as a further layer of gene repression when important genes must be strictly repressed in some cell type.

PART 5 FUNCTIONAL ANNOTATIONS OF HORSE CENTROMERES

RESULTS

Following the examples of the ENCODE consortium, an internationally coordinated Functional Annotation of Animal Genomes (FAANG) project was established. The aim of the consortium is to produce comprehensive maps of functional elements in the genomes of domesticated animal species based on common standardized protocols and procedures.

In 2017 we joined the FAANG consortium, as members of the horse genome community. A large set of markers will be evaluated by the consortium, including DNA methylation, several histone modifications and RNA-seq. Our role in the project is to analyze the epigenetic environment of the centromeric region in different tissues and individuals. To this purpose, we started to characterize the ECA11 centromere of the two FAANG female horses by ChIP-seq in fibroblast cell lines.

The chromatin extracted from fibroblast cell lines of the two horses (FAANG Horse 1 and FAANG Horse 2) was immunoprecipitated with an anti-CENP-A antibody and the purified DNA was paired-end sequenced through an Illumina HiSeq2500 platform, at IGA Technologies Service. For each sample, a fraction of non-precipitated chromatin, input, was saved as control, purified and sequenced in parallel with ChIP DNA. Using the IGV software, we analyzed the reads obtained from the sequencing on EquCab3 reference genome to visualize the CENP-A binding domains.

Figures R5-1 and R5-2 (red tracks) show the results of the ChIP-seq experiments on the two mares. As previously described, the distribution of the reads at the centromeric domains, spanning about 120 kb, is irregular probably due to assembly problems of EquCab3 in this region. Moreover, according to previous results from our group (Purgato S et al. 2015; Nergadze SG et al. 2018), a positional variation of the centromeric function on the two different horses can be detected.

We then evaluated the transcriptional profile of these centromeric loci, taking advantage of RNA-seq and microRNA-seq datasets produced by the FAANG collaboration.

We obtained from the FAANG consortium polyA-RNA-seq datasets from 30 tissues and miRNA-seq datasets from 8 tissues. Here, I report our preliminary

analysis of polyA- and miRNA-seq from eight tissues: Adipose Loin, Lamina, Left Ventricle, Liver, *Longissimus dorsi*, Lung, Ovary, and Parietal Cortex.

The results of the RNA-seq analysis are summarized in figure R5-1. For both mares, the 15 Mb region around the centromeric locus of horse chromosome 11 is shown. Panel A shows the CENP-A ChIP-seq tracks, obtained by us in the fibroblast cell lines, and the RNA-seq tracks provided by our FAANG collaborators from Adipose Loin, Lamina, Left Ventricle and Liver in both horses; panel B shows the CENP-A ChIP-seq track and the RNA-seq tracks from *Longissimus dorsi*, Lung, Ovary and Parietal Cortex in both horses.

In all tissues, except parietal cortex, PolyA-RNAs are absent in the 3 Mb region comprising the CENP-A binding domain while a high expression level was detected in the flanking regions as shown in Figure R5-2. These results were consistent with the interpretation that the satellite-less centromere of horse chromosome 11 is a gene desert (Wade CM et al. 2009). However in parietal cortex a large region of about 100 kb seem to be highly transcribed. Comparative analysis of these transcripts with online databases showed a partial overlap with a previously identified horse transcript, whose function is unknown.

The transcript we identified in the parietal cortex tissue is not classified as protein-coding mRNA from a preliminary analysis. Also it does flank the centromeric peak although they do not overlap. The actual centromeric position in the parietal cortex tissue may be in a different location (we used fibroblast cell lines.).

Further investigations will be necessary to verify the origin of these transcripts and to test whether they derive from gene expression regulation or from experimental artifacts.

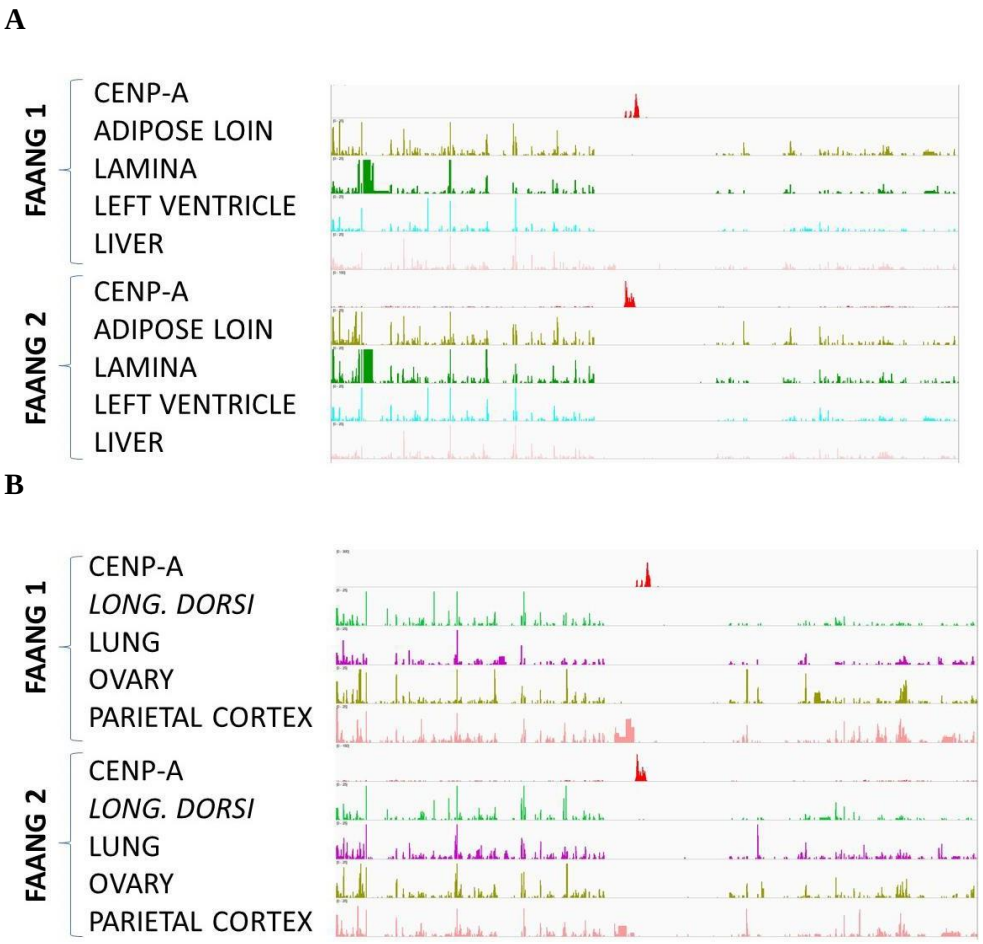


FIGURE R5-1. RNA-seq results of 8 different tissues of FAANG horse 1 and 2 within a 15 Mb region comprising the centromeric locus. *The two panels show the results of read mapping on the horse reference genome (EquCab3) of the RNA-seq datasets obtained from the 8 tissues on both FAANG horses; CENP-A peak resulting from our ChIP-seq experiment identifies the centromere position for both horses. (A) Tracks of CENP-A ChIP-seq, Adipose Loin RNA-seq, Lamina RNA-seq, Left Ventricle RNA-seq and Liver RNA-seq in both horses; (B) tracks of CENP-A ChIP-seq, Longissimus dorsi RNA-seq, Lung RNA-seq, Ovary RNA-seq and Parietal Cortex RNA-seq in both horses.*

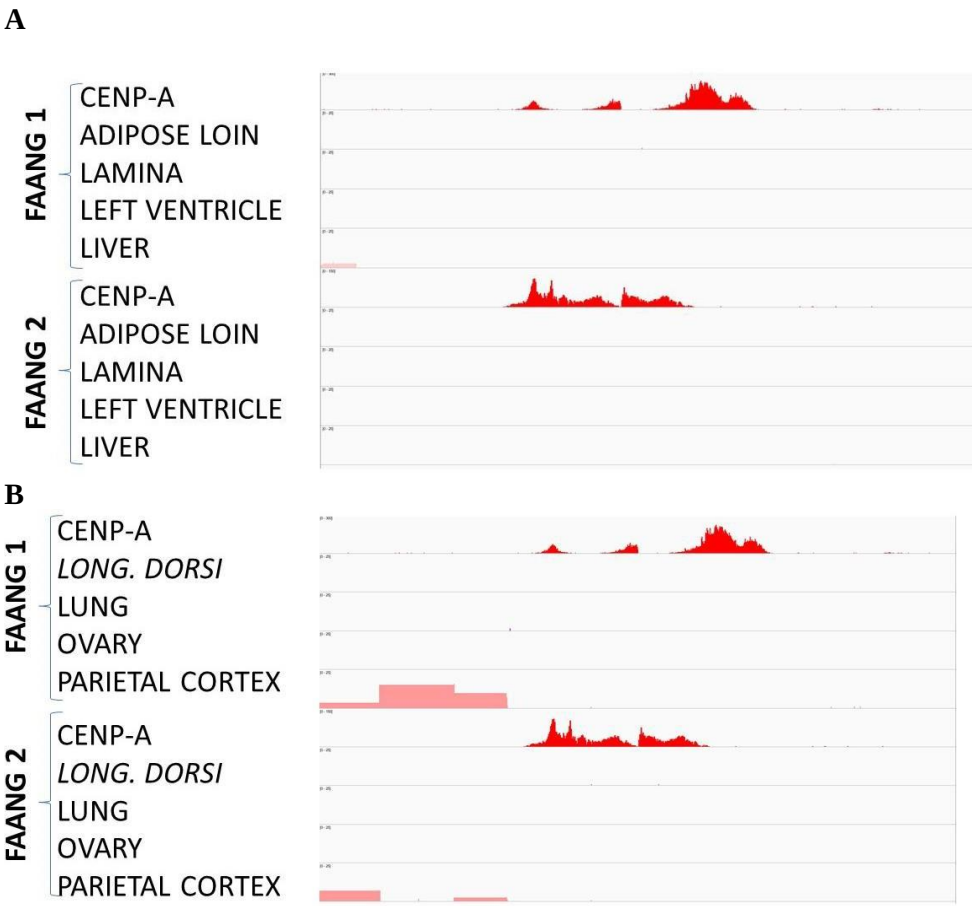


FIGURE R5-2. RNA-seq results of 8 different tissues of FAANG horse 1 and 2 within ~1 Mb genomic region around the centromeric locus. *The two panels show the results of read mapping on the horse reference genome (EquCab3) of the RNA-seq datasets obtained from the 8 tissues on both FAANG horses; CENP-A peak resulting from our ChIP-seq experiment identifies the centromere position for both horses. (A) Tracks of CENP-A ChIP-seq, Adipose Loin RNA-seq, Lamina RNA-seq, Left Ventricle RNA-seq and Liver RNA-seq in both horses; (B) tracks of CENP-A ChIP-seq, Longissimus dorsi RNA-seq, Lung RNA-seq, Ovary RNA-seq and Parietal Cortex RNA-seq in both horses.*

We then evaluated whether micro RNAs were present at the ECA 11 centromeric locus in the same above mentioned tissues. The same 15 Mb region comprising the centromere was examined and results are shown in figure R5-3. In all the tissues, the region is characterized by the presence of microRNA but it is possible to observe a general lower expression and/or absence of transcripts around the 3 Mb pericentromeric locus.

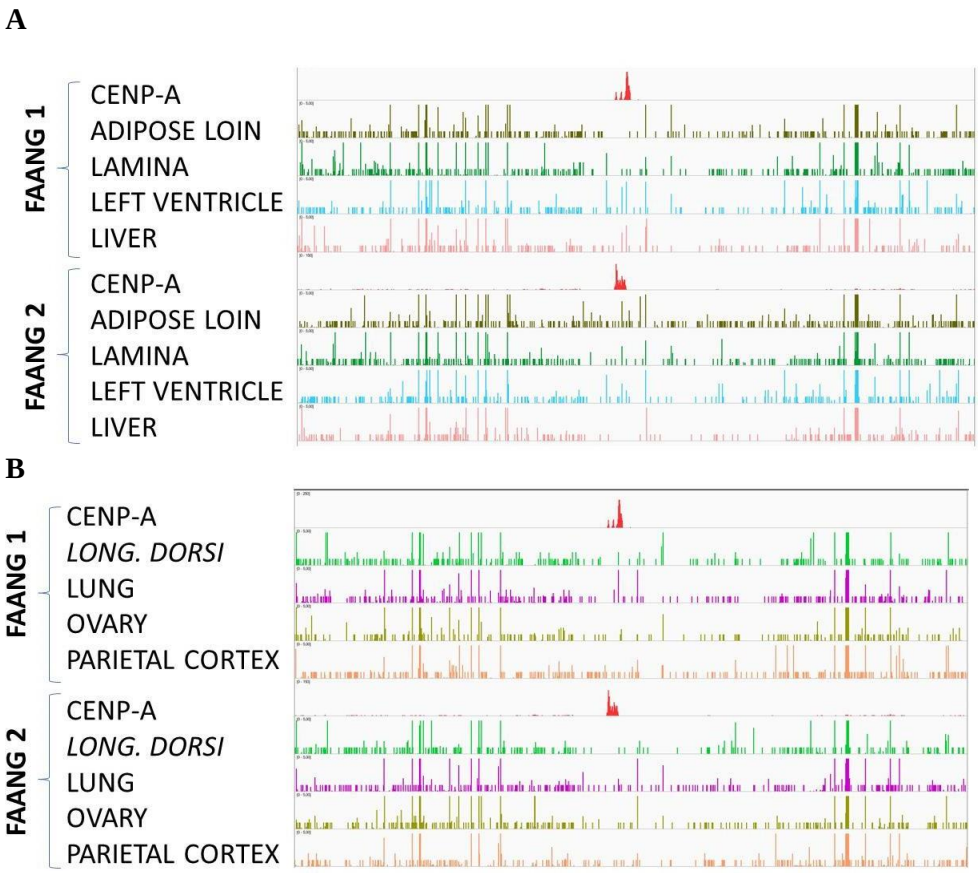
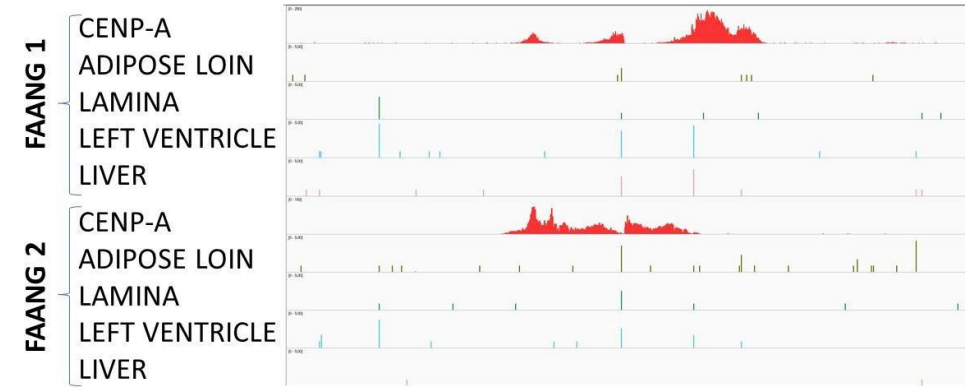


FIGURE R5-3. miRNA-seq results of 8 different tissues of FAANG horse 1 and 2 within a 15 Mb genomic region around the centromeric loci. *The two panels show the results of read mapping on the horse reference genome (EquCab3) of the miRNA-seq datasets obtained from the 8 tissues on both FAANG horses; CENP-A peak resulting from our ChIP-seq experiment identifies the centromere position for both horses. (A) Tracks of CENP-A ChIP-seq, Adipose Loin miRNA-seq, Lamina miRNA-seq, Left Ventricle miRNA-seq and Liver miRNA-seq in both horses; (B) tracks of CENP-A ChIP-seq, Longissimus dorsi miRNA-seq, Lung miRNA-seq, Ovary miRNA-seq and Parietal Cortex miRNA-seq in both horses.*

We then analyzed specifically the centromeric locus to test the presence of miRNAs in the CENP-A binding domain (Figure R5-4). When zooming into the centromeric locus we were able to identify multiple miRNAs, in all the tissues analyzed. There, miRNAs are all expressed at a very low level. Therefore, further experiments will be required to verify the origin of these transcripts and to test whether they represent real ncRNA species deriving from experimental artifacts. We observed a certain degree of variability. For example, Adipose Loin datasets in the two individuals show a different number of miRNAs (higher in Horse 2).

Unfortunately we cannot tell whether this is a biological variation with a meaningful relevance since the number of reads is different in the two datasets. To solve this problem a specific differential expression analysis must be performed to unravel tissues variability and interindividual miRNAs variability, which at the end may be influenced by the positional instability of the centromeric function. With this pilot analysis, we were able to assess that, although protein-coding genes are absent from this region as previously reported (Wade CM et al. 2009), satellite-less centromere in horse chromosome 11 is transcribed, at least for some miRNA transcripts.

A



B

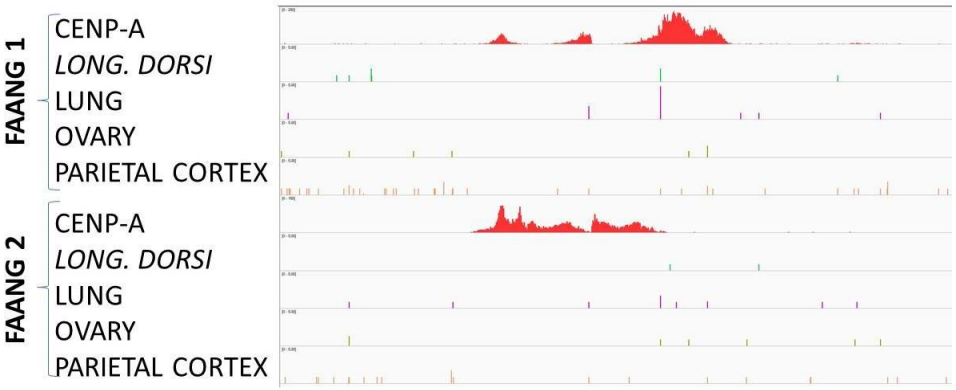


FIGURE R5-4. miRNA-seq results of 8 different tissues of FAANG horse 1 and 2 within a ~1 Mb genomic region around the centromeric loci. *The two panels show the results of read mapping on the horse reference genome (EquCab3) of the miRNA-seq datasets obtained from the 8 tissues on both FAANG horses; CENP-A peak resulting from our ChIP-seq experiment identifies the centromere position for both horses. (A) Tracks of CENP-A ChIP-seq, Adipose Loin miRNA-seq, Lamina miRNA-seq, Left Ventricle miRNA-seq and Liver miRNA-seq in both horses; (B) tracks of CENP-A ChIP-seq, Longissimus dorsi miRNA-seq, Lung miRNA-seq, Ovary miRNA-seq and Parietal Cortex miRNA-seq in both horses.*

DISCUSSION

Transcription at centromeric level was first reported for mouse satellite-based centromeres (Cohen AK et al. 1973) and since then several reports confirmed that transcription events and, sometimes, the transcript itself play a role in centromeric function and formation (Chueh AC et al. 2009; Grenfell AW et al. 2016). The presence of repetitive DNA normally associated with centromeres has so far hindered a detailed molecular dissection of the RNA transcripts, making the centromere an enigmatic locus. On the other hand, evolutionarily new centromeres and human neocentromeres were found to arise in gene desert regions (Cardone MF et al. 2006; Wade CM et al. 2009)

Thanks to our model system, we can evaluate the transcriptional profile of centromeres which are lacking satellite DNA and, unlike the clinical neocentromeres, are fixed in a species.

Since we joined the FAANG (Functional Annotation of ANimal Genomes) project we were able to start the analysis of NGS datasets produced by the horse community. NGS data for RNA, epigenetic markers and other genomic data is currently under production and we are participating by characterizing the centromere functional locus of chromosome 11 in different tissues of the two FAANG female horses. In particular, a preliminary analysis of the transcriptional state of these loci was reported in the previous section and compared to the localization of CENP-A binding domains carried out in fibroblast cell lines from the FAANG horses.

Results show that the ECA 11 satellite-less centromeres of both horses share the same peculiarity of other equids centromeres, for example the typical satellite-less centromere length spanning about 100 kb. Moreover, peak shape is irregular. This is probably due to a duplicon present in this region (small central peak). The position of the CENP-A binding domain is different in the two horses confirming that centromere sliding is occurring at satellite-less centromeres.

We then evaluated the transcriptional profile of the ECA 11 centromeric locus, taking advantage of RNA-seq and miRNA-seq datasets produced within the FAANG effort. Preliminary results regarding a 15 Mb genomic region around the centromere in eight different tissues from both horses demonstrated that the centromeric (~ 100 kb in length) and the pericentromeric (~ 3 Mb in length) loci are gene-deserts. No protein-coding genes were identified in the tissues as expected, confirming previously published data. However, looking in detail a 1 Mb region around the CENP-A peak we were able to identify a 50 kb transcribed region upstream of the CENP-A binding domain in the Parietal Cortex RNA-seq datasets of both FAANG horses. A comparison of these transcripts with those present in the online databases showed a partial overlap with an already identified transcripts, whose function is unknown.

This proximal transcriptional activity seems not to impair the centromeric function.

A consideration must be pointed out: the ChIP-seq experiment carried out on both horses to unravel the centromeric location was performed using fibroblast cell lines but we do not know whether the CENP-A binding domains are in the same position in different tissues. The answer to this question is actually one of our major tasks in the FAANG project. To this purpose we will perform centromere identification in different tissues to correlate the centromeric location with the epigenetic markers tested by the other members of the consortium.

We examined the presence of microRNAs in the ECA 11 centromeric region in the same tissues previously used for the RNA-seq. We identified the presence of several miRNAs in this genomic region. All the processed datasets showed transcription within the centromeric area, despite a lower transcription level respect to the surrounding region.

When we zoomed in at ~ 1 Mb around the centromere we observed a great number of miRNA transcripts all over the centromeric locus. Each tissue analyzed in both horses showed the presence of ncRNAs in correspondence of the centromeric peak. We can observe some degree of variation of the transcriptional expression level among tissues and between individuals. However, to completely identify possible differences in transcription, an accurate differential expression analysis is needed to identify tissue-specific and individual specific transcripts which may depends on the positional variation of the centromeric locus. It would be interesting to test whether centromere sliding can affect in some way miRNAs transcription.

Presence of transcription at the centromeric core has always been a fascinating task to study but a hard challenge to approach on. Since the presence of centromeric repetitive DNA in the majority of mammalian species, studies on transcription were addressed on the sole presence of satellite-based transcripts. Moving along on the study of centromere lacking satellite DNA, clinical and evolutionarily neocentromeres, results addressed the point that gene desert regions are preferred for the centromeric seeding. However short transcripts per se and the transcription machinery have been proposed to play a relevant role in centromeric formation and function. Indeed, with our work we highlighted the presence of miRNA transcripts within the centromeric core. Although our ChIP-seq data with transcription markers showed that the centromeric region is transcriptionally silent, short regions of open chromatin may be present in some cell cycle phase allowing transcription of small RNA molecules.

Thanks to the availability of a tissue biobank from FAANG horses (Burns EN et al. 2018), we plan to identify the position of CENP-A binding domains of ECA11 in

several tissues from different embryonic origin to investigate whether centromere sliding could occur also during development, and to test whether its positional variability may influence, or be influenced by micro RNA transcripts.

CONCLUDING REMARKS

PART 1 BIRTH EVOLUTION AND TRANSMISSION OF SATELLITE-LESS MAMMALIAN CENTROMERIC DOMAINS

In this part and in the attached paper (Nergadze SG et al 2018) we investigated the satellite-free centromeres of *Equus asinus* by using ChIP-seq with anti-CENP-A antibodies. We identified an extraordinarily high number of centromeres lacking satellite DNA in the donkey: 16 out of 31. All of them lay in LINE and AT rich regions. A subset of these centromeres is associated with DNA amplification. The location of CENP-A binding domains can vary in different individuals giving rise to epialleles. The analysis of epiallele transmission in equid hybrids showed that centromeric domains are inherited as Mendelian traits but their position can slide in one generation.

PART 2 COMPARATIVE ANALYSIS OF CENP-A BINDING DOMAINS IN 7 EQUID SPECIES

In this part we demonstrated that an incredible number of satellite-less centromeres is present in the 7 equid species here analyzed through ChIP-seq, a total of 82 satellite-less centromeres: 1 in *Equus caballus* (Nergadze et al. 2018), 10 in *Equus zebra hartmannae*, 11 in *Equus grevyi*, 14 in *Equus burchelli*, 16 in *Equus asinus* (Nergadze et al. 2018), 15 *in Equus kiang*, 15 in *Equus hemionus onager*. Within many of these neocentromeres we detected some features already identified in the donkey satellite-less centromeres (Part 1 and Nergadze et al. 2018). Many neocentromeres displayed sequence rearrangements compared to the horse reference sequence, other neocentromeres exhibited sequence amplification, while in others we identified two distinct epialleles. We were able to correlate two different phenomena, centromere repositioning, described as event of centromere movement during species evolution, with centromere sliding, which is a phenomenon of centromeric repositioning at smaller scale, which may occur from one generation to another. Then to establish a phylogenetic history of those chromosomes carrying repositioned centromeres we inferred three models of equid evolutionary tree using neocentromeres detected through ChIP-seq as markers. Taking into account the criterion of maximum parsimony, the classical phylogenetic tree model was the best model, since it comprised the lowest number of CR events compared to the other two trees. Results also suggested that, at least in some cases, “centromerization” hotspots may be present and used as seeding sites for neocentromeric formation when independent centromere repositioning events may have occurred in different equid lineages. We also demonstrated that

chromosomal rearrangements may have occurred during the equid evolution impairing the neocentromere detection.

PART 3 GENOME-WIDE EPIGENETIC ANALYSIS OF SATELLITE-LESS CENTROMERES

The epigenetic analysis of the satellite-less centromeres revealed that they are not enriched for H3K4me3 marker, since its enrichment is associated with actively transcribed regions. Moreover, the orthologous non-centromeric regions are not enriched for H3K4me3 as well. We can propose that H3K4me3 is not necessary for the centromeric function, as is the case with other markers for open and transcribed chromatin. Moreover, we observed that H4K20me1 histone modification exactly co-localizes with the CENP-A binding domains in most of the satellite-less centromeres. On the other hand, the orthologous non centromeric regions lack this marker. Interestingly, the pericentromeric region, corresponding to the H3K9me3 heterochromatin environment around the CENP-A binding domain, shows an overall "de-enrichment" of H4K20me1 level compared to the surrounding region. In conclusion we demonstrate that the H4K20me1 histone modification is coupled with the centromeric function and we can propose that this marker is absent before the establishment of the centromere.

PART 4 ECTOPIC CENP-A BINDING SITES

In this work, we isolated and analyzed CENP-A binding sites mapping outside the centromeres, called ectopic CENP-A binding sites. We detected that an incredible high number of ectopic CENP-A binding sites are present also in non-centromeric regions and scattered across all the chromosomes in all the species analyzed (horse, donkey, mouse and HeLa cells). We observed that ectopic CENP-A binds gene-promoter-TSS regions at a higher frequency compared to the expected value in all datasets analyzed except for the mouse dataset. Moreover we detected that promoter-TSS regions bound by CENP-A are enriched for brain-related genes. Finally we found that sequences bound by ectopic CENP-A are enriched for the REST motif, sequence bound by a neuronal gene repressor.

PART 5 FUNCTIONAL ANNOTATIONS OF HORSE CENTROMERES

We analyzed two horses from the FAANG project. The position of the CENP-A binding domain is different in the two FAANG horses confirming that centromere sliding is occurring at satellite-less centromeres of these individuals. A preliminary

analysis of the transcriptional state of these loci through RNA-seq indicated that no protein-coding genes were identified in eight different tissues from both horses in the centromeric (~ 100 kb in length) and the pericentromeric (~ 3 Mb in length) loci, which are gene-deserts, although a 50 kb transcribed region upstream of the CENP-A binding domain in the Parietal Cortex RNA-seq datasets of both FAANG horses, which needs more analysis to test its relevance. When miRNA datasets from the same eight tissue were analyzed, we identified several ncRNAs in this genomic region. These datasets showed transcription activity within the centromeric area, despite a lower transcription level respect to the surrounding region. Indeed, this work suggests that miRNA transcripts within the centromeric core are transcribed.

REFERENCES

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

Alexandrov IA, Medvedev LI, Mashkova TD, Kisselev LL, Romanova LY, Yurov YB. Definition of a new alpha satellite suprachromosomal family characterized by monomeric organization. Nucleic Acids Res. 1993;21:2209-2215.

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

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

Alonso A, Hasson D, Cheung F, Warburton PE. A paucity of heterochromatin at functional human neocentromeres. Epigenetics Chromatin. 2010;3:1.

Amor DJ, Choo KHA. Neocentromeres: Role in Human Disease, Evolution, and Centromere Study. Am J Hum Genet. 2002;71:695–714.

Amor DJ, Bentley K, Ryan J, Perry J, Wong L, Slater H, Choo KH. Human centromere repositioning "in progress". Proc Natl Acad Sci U S A. 2004;101:6542-6547.

An O, Dall'Olio GM, Mourikis TP, Ciccarelli FD. NCG 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings. Nucleic Acids Res. 2016;44:992–999.

Athwal RK, Walkiewicz MP, Baek S, Fu S, Bui M, Camps J, Ried T, Sung MH, Dalal Y. CENP-A nucleosomes localize to transcription factor hotspots and subtelomeric sites in human cancer cells. Epigenetics Chromatin. 2015;8:2.

Bailey AO, Panchenko T, Shabanowitz J, Lehman SM, Bai DL, Hunt DF, Black BE, Foltz DR. Identification of the posttranslational modifications present in centromeric chromatin. Mol Cell Proteomics. 2015;15:918-931.

Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, Madrigal P, Taslim C, Zhang J. Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. PLoS Comput Biol. 2013;9:e1003326.

Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. High-resolution profiling of histone methylations in the human genome. Cell. 2007;129:823–837.

Beck DB, Oda H, Shen SS, Reinberg D. PR-Set7 and H4K20me1: at the crossroads of genome integrity, cell cycle, chromosome condensation, and transcription. Genes Dev. 2012;26:325–337.

Bensasson D. Evidence for a high mutation rate at rapidly evolving yeast centromeres. BMC Evol Biol. 2011; 11:211

Bessis A, Champtiaux N, Chatelin L, Changeux JP. The neuron-restrictive silencer element: a dual nhancer/silencer crucial for patterned expression of a nicotinic receptor gene in the brain. Proc Natl Acad Sci U S A. 1997;94:5906–5911.

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, Panchenko T, Cleveland DW, Black BE, Shah JV, Jansen LE. The quantitative architecture of centromeric chromatin. eLife. 2014;3:e02137.

Burns EN, Bordbari MH, Mienaltowski MJ, Affolter VK, Barro MV, Gianino F, Gianino G, Giulotto E, Kalbfleisch TS, Katzman SA, Lassaline M, Leeb T, Mack M, Müller EJ, MacLeod JN et al. Generation of an equine biobank to be used for Functional Annotation of Animal Genomes project. Anim Genet. 2018; doi: 10.1111/age.12717.

Buscaino A, Allshire R, Pidoux A. Building centromeres: home sweet home or a nomadic existence? Curr Opin Genet Dev. 2010;20:118-126.

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

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

Carbone L, Nergadze SG, Magnani E, Misceo D, Francesca Cardone M, Roberto R, Bertoni L, Attolini C, Piras MF, De Jong P, Raudsepp T, ChowDhary BP, Guerin G, Archidiacono N, Rocchi 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, Carbone L, De Jong P, Stanyon R, D'Addabbo P, Archidiacono N, She X Eichler E, Warburton P, Rocchi M. Independent centromere formation in a capricious, gene-free domain of chromosome 13q21 in Old World monkeys and pigs. Genome Biol. 2006;7:R91..

Carone DM, Longo MS, Ferreri GC, Hall L, Harris M, Shook N, Bulazel KV, Carone BR, Obergfell C, O'Neill MJ, O'Neill RJ. A new class of retroviral and satellite encoded small RNAs emanates from mammalian centromeres. Chromosoma. 2009;118:113-125.

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

Carone DM, Zhang C, Hall LE, Obergfell C, Carone BR, O'Neill MJ, O'Neill RJ. Hypermorphic expression of centromeric retroelement-encoded small RNAs impairs CENP-A loading. Chromosome Res. 2013;21:49-62.

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 27;9:35.

Chan FL, Marshall OJ, Saffery R, Kim BW, Earle E, Choo KH, Wong LH. Active transcription and essential role of RNA polymerase II at the centromere during mitosis. Proc Natl Acad Sci U S A. 201;109:1979-84.

Choi ES, Strålfors A, Catania S, Castillo AG, Svensson JP, Pidoux AL, Ekwall K, Allshire RC. 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;10:182-188.

Chueh AC, Northrop EL, Brettingham-Moore KH, Choo KHA, Wong LH. LINE Retrotransposon RNA Is an Essential Structural and Functional Epigenetic Component of a Core Neocentromeric Chromatin. PLOS Genet. 2009;5:1000354.

Cohen AK, Huh TY, Helleiner CW. Transcription of satellite DNA in mouse L-cells. Can J Biochem. 1973;51:529–532.

Cooke CA, Bernat RL, Earnshaw WC. CENP-B: a major human centromere protein located beneath the kinetochore. J Cell Biol. 199;110:1475-1488.

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

Drinnenberg IA, DeYoung D, Henikoff S, Malik HS. Recurrent loss of CenH3 is associated with independent transitions to holocentricity in insects. eLife. 2014;3:e03676.

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

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

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

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

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.

Foltz DR, Jansen LET, Bailey AO, Yates JR, Bassett EA, Wood S, Black BE, Cleveland DW. Centromere-specific assembly of CENP-a nucleosomes is mediated by HJURP. Cell. 2009;137:472–484.

Fujita R, Otake K, Arimura Y, Horikoshi N, Miya Y, Shiga T, Osakabe A, Tachiwana H, Ohzeki J, Larionov V, Masumoto H, Kurumizaka H. Stable complex formation of CENP-B with the CENP-A nucleosome. Nucleic Acids Res. 201;434909-4922.

Fukagawa T, Brown WR. Efficient conditional mutation of the vertebrate CENP-C gene. Hum Mol Genet. 1997;6:2301-2308.

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

Garavís M, Escaja N, Gabelica V, Villasante A, González C. Centromeric Alpha-Satellite DNA Adopts Dimeric i-Motif Structures Capped by AT Hoogsteen Base Pairs. Chem Weinh Bergstr Ger. 2015;21:9816–9824.

Gartenberg M. Heterochromatin and the cohesion of sister chromatids. Chromosome Res. 2009;17:229-238

Giulotto E, Raimondi E, Sullivan K. The unique DNA sequences underlying equine centromeres. Prog Mol Subcell Biol 2017;56: 337–354.

Gong Z, Wu Y, Koblízková A, Torres GA, Wang K, Iovene M, Neumann P, Zhang W, Novák P, Buell CR, Macas J, Jiang J. Repeatless and repeat-based centromeres in potato: implications for centromere evolution. Plant Cell. 2012;24:3559–3574.

Gonzalez M, He H, Dong Q, Sun S, Li F. Ectopic centromere nucleation by CENP--a in fission yeast. Genetics. 2014;198:1433–1446.

Grenfell AW, Heald R, Strzelecka M. Mitotic noncoding RNA processing promotes kinetochore and spindle assembly in Xenopus. J Cell Biol. 2016;214:133–141.

Han Y, Zhang Z, Liu C, Liu J, Huang S, Jiang J, Jin W. Centromere repositioning in cucurbit species: Implication of the genomic impact from centromere activation and inactivation. Proc Natl Acad Sci. 2009;106:14937–14941.

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

Hayden KE, Strome ED, Merrett SL, Lee H-R, Rudd MK, Willard HF. Sequences Associated with Centromere Competency in the Human Genome. Mol Cell Biol. 2013;33:763–772.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng J, Murre C, Singh S, Glass C. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell. 2010;38:576–589.

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

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

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

Horvath JE, Gulden CL, Vallente RU, Eichler MY, Ventura M, McPherson JD, Graves TA, Wilson RK, Schwartz S, Rocchi M, Eichler EE. Punctuated duplication seeding events during the evolution of human chromosome 2p11. Genome Res. 200;15:914-927..

Hsieh CL, Lin CL, Liu H, Chang YJ, Shih CJ, Zhong CZ, Lee SC, Tan BC. WDHD1 modulates the post-transcriptional step of the centromeric silencing pathway. Nucleic Acids Res. 2011;39:4048-4062.

Huang J, Zhao Y, Bai D, Shiraigol W, Li B, Yang L, Wu J, Bao W, Ren X, Jin B, Zhao Q, Li A, Bao S, Bao W, et al. Donkey genome and insight into the imprinting of fast karyotype evolution. Sci Rep. 2015;5:14106.

Hudson DF, Fowler KJ, Earle E, Saffery R, Kalitsis P, Trowell H, Hill J, Wreford NG, de Kretser DM, Cancilla MR, Howman E, Hii L, Cutts SM, Irvine DV, Choo KH. Centromere protein B null mice are mitotically and meiotically normal but have lower body and testis weights. J Cell Biol. 199;141:309-319

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

Iwata-Otsubo A, Dawicki-McKenna JM, Akera T, Falk SJ, Chmátal L, Yang K, Sullivan BA, Schultz RM, Lampson MA, Black BE. Expanded Satellite Repeats Amplify a Discrete CENP-A Nucleosome Assembly Site on Chromosomes that Drive in Female Meiosis. Curr Biol. 2017;27:2365-2373

Jónsson H, Schubert M, Seguin-Orlando A, Ginolhac A, Petersen L, Fumagalli M, Albrechtsen A, Petersen B, Korneliussen T, VilstrupJ, Lear T, Myka JL, Lundquist J, Miller DC, Alfarhan A, et al. Speciation with gene flow in equids despite extensive chromosomal plasticity. Proc Natl Acad Sci U S A. 2014;111:18655–18660.

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

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36.

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

Knehr M, Poppe M, Schroeter D, Eickelbaum W, Finze EM, Kiesewetter UL, Enulescu M, Arand M, Paweletz N. Cellular expression of human centromere protein C demonstrates a cyclic behavior with highest abundance in the G1 phase. Proc Natl Acad Sci U S A. 1996;93:10234-10239.

Kobayashi T, Yamada F, Hashimoto T, Abe S, Matsuda Y, Kuroiwa A. Centromere repositioning in the X chromosome of XO/XO mammals, Ryukyu spiny rat. Chromosome Res Int J Mol Supramol Evol Asp Chromosome Biol. 2008;16:587–593.

Kriegs JO, Churakov G, Kiefmann M, Jordan U, Brosius J, Schmitz J. Retroposed elements as archives for the evolutionary history of placental mammals. PLoS Biol. 2006;4:e91.

Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinforma Oxf Engl. 2011;27:1571–1572.

Lacoste N, Woolfe A, Tachiwana H, Garea A, Barth T, Cantaloube S, Kurumizaka H, Imhof A, Almouzni G. Mislocalization of the Centromeric Histone Variant CenH3/CENP-A in Human Cells Depends on the Chaperone DAXX. Mol Cell. 2014;53:631–644.

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

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

Liehr T, Kosyakova N, Weise A, Ziegler M, Raabe-Meyer G. First case of a neocentromere formation in an otherwise normal chromosome 7. Cytogenet Genome Res. 2010;128:189-191.

Lima-De-Faria. Genetics, origin and evolution of kinetochores. Hereditas. 1949; https://doi.org/10.1111/j.1601-5223.1949.tb02883.x

Lo AW, Magliano DJ, Sibson MC, Kalitsis P, Craig JM, Choo KH. A novel chromatin immunoprecipitation and array (CIA) analysis identifies a 460-kb CENP-A-binding neocentromere DNA. Genome Res. 2001;11:448-457.

Lo AW, Craig JM, Saffery R, Kalitsis P, Irvine DV, Earle E, Magliano DJ, Choo KH. 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, Muzny DM, Yang SP, Wang Z, Chinwalla AT, Minx P, Mitreva M, Cook L, Delehaunty KD, Fronick C, Schmidt H, et al. Comparative and demographic analysis of orang-utan genomes. Nature. 2011;469:529–533.

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

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, Totoki Y, Ito TK, Minamino T, Nakamura H, Yamaguchi S, Shibata T, Masutomi K. Involvement of telomerase reverse transcriptase in heterochromatin maintenance. Mol Cell Biol. 2014;34:1576-1593.

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

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

Masumoto H, Ikeno M, Kitagawa K. The centromere of human chromosome Tanpakushitsu Kakusan Koso. 1993 Feb;38:403-411.

McEwen BF, Dong Y, VandenBeldt KJ. Using electron microscopy to understand functional mechanisms of chromosome alignment on the mitotic spindle. Methods Cell Biol. 2007;79:259-293.

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

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D, Garcia JF, DeRisi JL, Smith T, Tobias C 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.

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

Mulligan P, Westbrook TF, Ottinger M, Pavlova N, Chang B, Macia E, Shi YJ, Barretina J, Liu J, Howley PM, Elledge SJ, Shi Y. CDYL Bridges REST and Histone Methyltransferases for Gene Repression and Suppression of Cellular Transformation. Mol Cell. 2008;32:718–726.

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

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

Musilova P, Kubickova S, Vahala J, Rubes J. Subchromosomal karyotype evolution in Equidae. Chromosome Res Int J Mol Supramol Evol Asp Chromosome Biol. 2013;21:175–187.

Nagpal H, Fukagawa T. Kinetochore assembly and function through the cell cycle. Chromosoma. 2016;125:645–659.

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

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

Oakenfull EA, Lim HN, Ryder OA. A survey of equid mitochondrial DNA: Implications for the evolution, genetic diversity and conservation of Equus. Conserv Genet. 2000;1:341–355.

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

Ohno Y, Ogiyama Y, Kubota Y, Kubo T, Ishii K. Acentric chromosome ends are prone to fusion with functional chromosome ends through a homology-directed rearrangement. Nucleic Acids Res. 2016;44:232–244.

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

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

Partridge JF, Borgstrøm B, Allshire RC. Distinct protein interaction domains and protein spreading in a complex centromere. Genes Dev. 2000;14:783-791.

Perpelescu M, Fukagawa T. The ABCs of CENPs. Chromosoma. 2011 Oct 13;120(5):425–46.

Peters AH, Kubicek S, Mechtler K, O'Sullivan RJ, Derijck AA, Perez-Burgos L, Kohlmaier A, Opravil S, Tachibana M, Shinkai Y, Martens JH, Jenuwein T. Partitioning and plasticity of repressive histone methylation states in mammalian chromatin. Mol Cell. 200;12:1577-1589.

Piras FM, Nergadze SG, Magnani E, Bertoni L, Attolini C, Khoriauli L, Raimondi E, Giulotto E. Uncoupling of Satellite DNA and Centromeric Function in the Genus Equus. PLoS Genet. 2010;6:e1000845.

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

Plohl M, Luchetti A, 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;123:313-25

Pluta AF, Saitoh N, Goldberg I, Earnshaw WC. Identification of a subdomain of CENP-B that is necessary and sufficient for localization to the human centromere. J Cell Biol. 1992;116:1081-1093.

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

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

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinforma Oxf Engl. 2010;26:841–842.

Rhoades MM, Vilkomerson H. On the Anaphase Movement of Chromosomes. Proc Natl Acad Sci U S A. 1942;28:433-436.

Rice JC, Briggs SD, Ueberheide B, Barber CM, Shabanowitz J, Hunt DF, Shinkai Y, Allis CD. Histone methyltransferases direct different degrees of methylation to define distinct chromatin domains. Mol Cell. 200;12:1591-1598.

Richard F, Messaoudi C, Lombard M, Dutrillaux B. Chromosome homologies between man and mountain zebra (Equus zebra hartmannae) and description of a new ancestral synteny involving sequences homologous to human chromosomes 4 and 8. Cytogenet Cell Genet. 2001;93:291-296.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative Genomics Viewer. Nat Biotechnol. 2011;29:24–26.

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

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

Ruthenburg AJ, Allis CD, Wysocka J. Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. Mol Cell. 2007;25:15–30.

Saffery R, Sumer H, Hassan S, Wong LH, Craig JM, Todokoro K, Anderson M, Stafford A, Choo KH. Transcription within a functional human centromere. Mol Cell. 2003;12:509-516.

Shang WH, Hori T, Toyoda A, Kato J, Popendorf K, Sakakibara Y, Fujiyama A, Fukagawa T. Chickens possess centromeres with both extended

tandem repeats and short non-tandem-repetitive sequences. Genome Res. 2010;20:1219–1228.

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

Shang WH, Hori T, Martins NM, Toyoda A, Misu S, Monma N, Hiratani I, Maeshima K, Ikeo K, Fujiyama A, Kimura H, Earnshaw WC, Fukagawa T. 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.

Söllner JF, Leparc G, Hildebrandt T, Klein H, Thomas L, Stupka E, Simon E. An RNA-Seq atlas of gene expression in mouse and rat normal tissues. Sci Data. 2017;4:170185.

Sugimoto K, Yata H, Muro Y, Himeno M. Human centromere protein C (CENP-C) is a DNA-binding protein which possesses a novel DNA-binding motif. J Biochem. 199;116:877-881.

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

Sullivan BA, Schwartz S. Identification of centromeric antigens in dicentric Robertsonian translocations: CENP-C and CENP-E are necessary components of functional centromeres. Hum Mol Genet. 1995;4:2189-2197.

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 LL, Boivin CD, Mravinac B, Song IY, Sullivan BA. Genomic size of CENP-A domain is proportional to total alpha satellite array size at human centromeres and expands in cancer cells. Chromosome Res Int J Mol Supramol Evol Asp Chromosome Biol. 2011;19:457–470.

Steiner CC, Mitelberg A, Tursi R, Ryder OA. Molecular phylogeny of extant equids and effects of ancestral polymorphism in resolving species-level phylogenies. Mol Phylogenet Evol. 2012; 65:573–581.

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

Stellfox ME, Bailey AO, Foltz DR. Putting CENP-A in its place. Cell Mol Life Sci. 201;70:387-406.

Tachiwana H, Kagawa W, Shiga T, Osakabe A, Miya Y, Saito K, Hayashi-Takanaka Y, Oda T, Sato M, Park SY, Kimura H, Kurumizaka H. Crystal structure of the human centromeric nucleosome containing CENP-A. Nature. 2011;476:232–235.

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

Torras-Llort M, Moreno-Moreno O, Azorín F. Focus on the centre: the role of chromatin on the regulation of centromere identity and function. EMBO J. 2009;28:2337-2348.

Tyler-Smith C, Gimelli G, Giglio S, Floridia G, Pandya A, Terzoli G, Warburton PE, Earnshaw WC, Zuffardi O. Transmission of a fully functional human neocentromere through three generations. Am J Hum Genet. 1999;64:1440-1444.

Uhlen M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA et al. Tissue-based map of the human proteome. Science. 2015;347:1260419–1260419.

Usdin K, Chevret P, Catzeflis FM, Verona R, Furano AV. L1 (LINE-1) retrotransposable elements provide a "fossil" record of the phylogenetic history of murid rodents. Mol Biol Evol. 1995;12:73-82.

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.

Vakoc CR, Sachdeva MM, Wang H, Blobel GA. Profile of histone lysine methylation across transcribed mammalian chromatin. Mol Cell Biol. 2006;26:9185–9195.

Van Hooser AA, Ouspenski II, Gregson HC, Starr DA, Yen TJ, Goldberg ML, Yokomori K, Earnshaw WC, Sullivan KF, Brinkley BR. Specification

of kinetochore-forming chromatin by the histone H3 variant CENP-A. J Cell Sci. 2001;114:3529-3542.

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

Ventura M, Weigl S, Carbone L, Cardone MF, Misceo D, Teti M, D'Addabbo P, Wandall A, Björck E, de Jong PJ, She X, Eichler EE, Archidiacono N, Rocchi M. Recurrent Sites for New Centromere Seeding. Genome Res. 2004;14:1696–1703.

Ventura M, Antonacci F, Cardone MF, Stanyon R, D'Addabbo P, Cellamare A, Sprague LJ, Eichler EE, Archidiacono N, Rocchi M. Evolutionary formation of new centromeres in macaque. Science. 2007;316:243–246.

Verdaasdonk JS, Bloom K 2011. Centromeres: unique chromatin structures that drive chromosome segregation. Nat Rev Mol Cell Biol. 201;12:320-332.

Verdel A, Jia S, Gerber S, Sugiyama T, Gygi S, Grewal SI, Moazed D. RNAi-mediated targeting of heterochromatin by the RITS complex. Science. 2004;303:672-676.

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. 1993;52:1153–1163.

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

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

Warburton PE, Cooke CA, Bourassa S, Vafa O, Sullivan BA, Stetten G, Gimelli G, Warburton D, Tyler-Smith C, Sullivan KF, Poirier GG, Earnshaw WC. 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, Naritomi K, Tohma T, Nagai T, Hasegawa T, Ohashi H, Govaerts LC, Eussen BH, Van

Hemel JO, Lozzio C, Schwartz 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.

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

Wong LH, Brettingham-Moore KH, Chan L, Quach JM, Anderson MA, Northrop EL, Hannan R, Saffery R, Shaw ML, Williams E, Choo KH. Centromere RNA is a key component for the assembly of nucleoproteins at the nucleolus and centromere. Genome Res. 2007;17:1146-1160.

Yang CH, Tomkiel J, Saitoh H, Johnson DH, Earnshaw WC. Identification of overlapping DNA-binding and centromere-targeting domains in the human kinetochore protein CENP-C. Mol Cell Biol. 1996;16:3576-3586.

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. 2003;102:235-243.

Yang F, Fu B, O'Brien PCM, 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 Int J Mol Supramol Evol Asp Chromosome Biol. 2004;12:65–76.

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

Zeitlin SG, Baker NM, Chapados BR, Soutoglou E, Wang JY, Berns MW, Cleveland DW. Double-strand DNA breaks recruit the centromeric histone CENP-A. Proc Natl Acad Sci U S A. 2009;106:15762–15767.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137.

Zhao S, Zhang B. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics. 2015;16:97.

LIST OF ORIGINAL MANUSCRIPTS

Peer review publication

Nergadze SG*, Piras FM*, Gamba R*, **Corbo M***, Cerutti F, McCarter JGW, Cappelletti E, Gozzo F, Harman RM, Antczak DF, Miller D, Scharfe M, Pavesi G, Raimondi E, Sullivan KF, Giulotto E. Birth, evolution, and transmission of satellite-free mammalian centromeric domains. Genome Research 2018 Jun;28(6):789-799. doi: 10.1101/gr.231159.117.

*These authors contributed equally to this work.

Meeting abstracts

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

Roberti A, **Corbo M,** Bensi M Piras FM, Giulotto E, Raimondi E *Epigenetic Marks at Satellite-Free and Satellite-Based Centromeres.* XV Congress of the Italian Federation of Life Sciences - Rome, Italy, 18-21 September 2018

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

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

Corbo M, Roberti A, Piras FM, Cappelletti E, Bensi M, Nergadze SG, Raimondi E, Giulotto E. *The epigenetic landscape of mammalian centromeres*.2017. 2nd Joint Annual Symposium of the Departments of Biology Biotechnology, Molecular Medicine and CNR-Institute of Molecular Genetics – Pavia, Italy, 20-22 June 2018

Roberti A, Nergadze SG, Bensi M, Gamba R, **Corbo M**, Piras FM, Cappelletti E, Giulotto E, Raimondi E. *Epigenetic modifications at satellite-less evolutionarily*

new centromeres.2017. Congress of the Italian Geneticists Association AGI – Cortona, Italy, 7-9 September 2017

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

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

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

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

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

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

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

Research

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

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

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

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

[Supplemental material is available for this article.]

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

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

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

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

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

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

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

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

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

Results

Satellite-free CENPA binding domains in *Equus asinus*

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

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

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

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

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

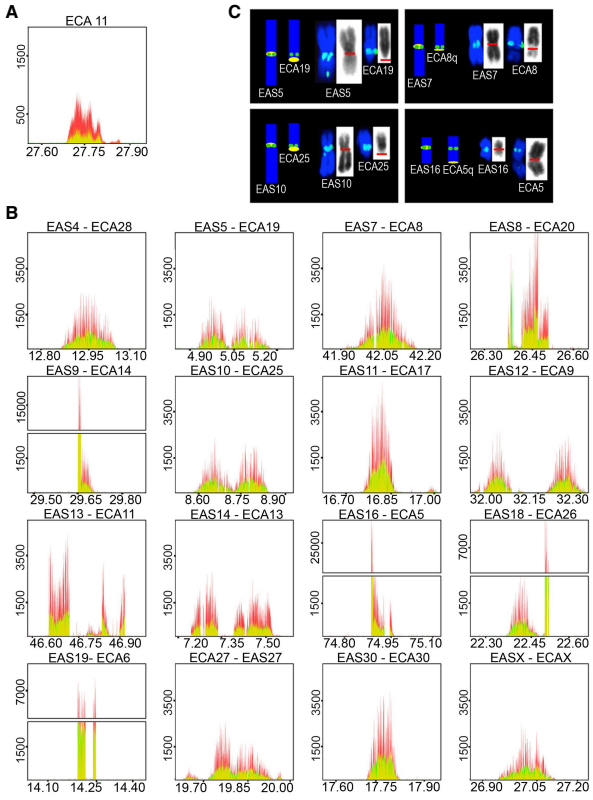


Figure 1. Identification of satellite-free centromeres in *Equus asinus*. ChIP-seq reads from primary fibroblasts of HorseS (*A*) and DonkeyA (*B*) were mapped on the EquCab2.0 horse reference genome. Immunoprecipitation was performed with an antibody against human CENPA (red) or with a CREST serum (green).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

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

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

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

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

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

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

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

Tandem repetitions associated with some satellite-free centromeres

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

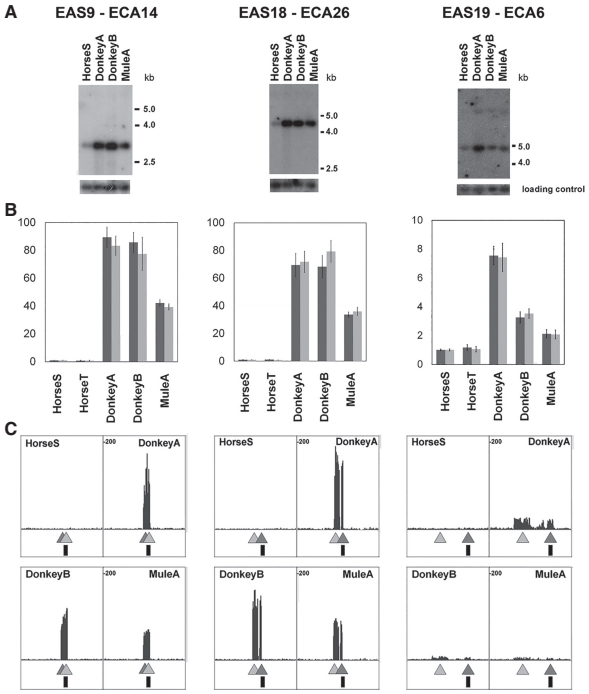


Figure 2. DNA sequence amplification at the centromeres of *E. asinus* Chromosomes 9, 18, and 19. The number of the *E. asinus* chromosome (EAS) and of its ortholog in *E. caballus* (ECA) is reported on *top*. (*A*) Southern blot analysis of genomic DNA from one horse, two donkeys, and a mule (MuleA,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

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

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

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

DNA sequence analysis of the satellite-free centromeric domains

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

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

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

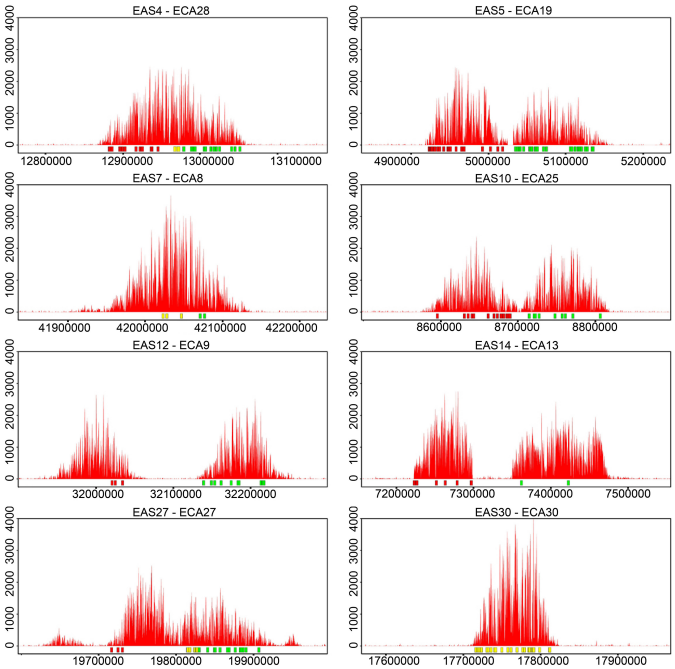


Figure 3. Identification of epialleles through SNV analysis. The positions of single nucleotide variants (SNVs), located within each centromeric domain, are represented as colored rectangles under each ChIP-seq profile. Reads were mapped on the chimeric EquCabAsiA reference genome. The *y*-axis reports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

Centromere sliding occurs in *Equus asinus*

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

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

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

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

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

Germline and somatic transmission of centromeric domains

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

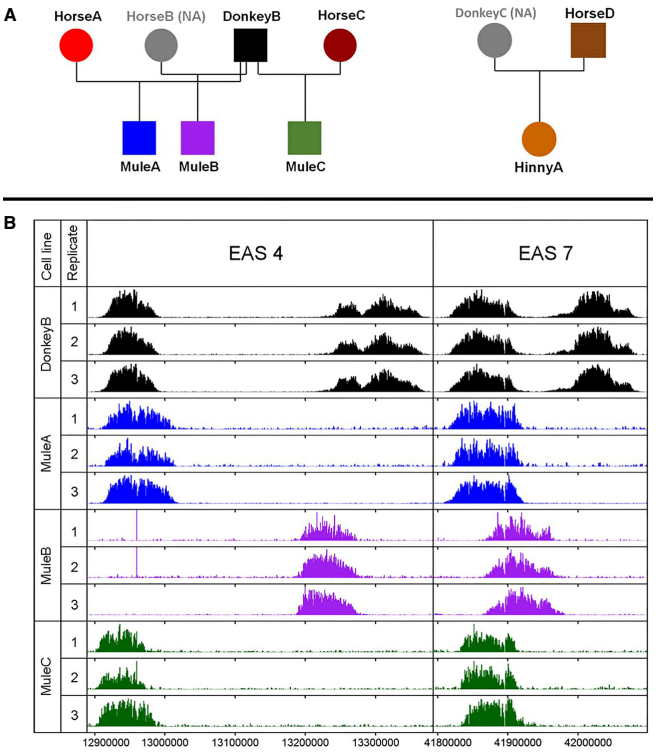


Figure 4. Transmission of satellite-free centromeric domains in hybrids. (*A*) Family trees reporting the genetic relationships among the individuals used in this study. Each color represents an individual, and the same color code is used in *B*. Cell lines from the individuals in gray were not available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

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

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

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

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

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

Discussion

Identification and DNA sequence composition of satellite-free centromeres

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

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

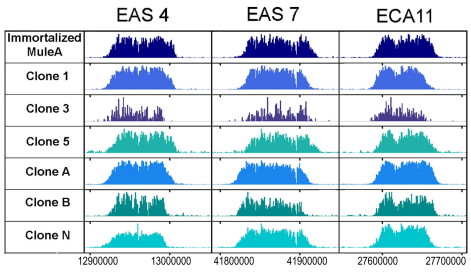


Figure 5. Transmission of satellite-free centromeric domains in clonal cell lines. ChIP-seq analysis of the immortalized cell line obtained from MuleA primary fibroblasts and six clonal derivative cell lines. Three centromeric domains taken as examples are shown (EAS4, EAS7, and ECA11). Results from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

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

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

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

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

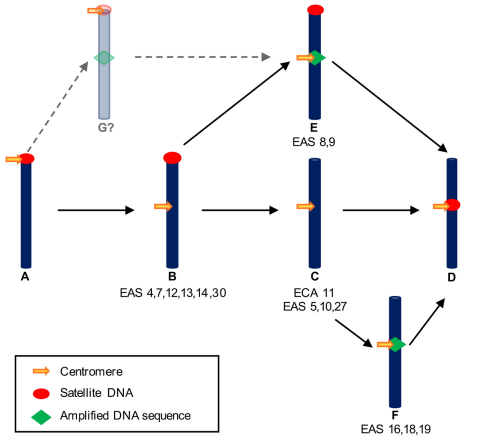


Figure 6. Model for the maturation of a centromere during evolution. Different pathways can be envisaged leading to a fully mature satellite-based repositioned centromere (D) from an ancestral centromere with satellite repeats (A) through satellite-free intermediates (B,C,E,F). The first route (A–D)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

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

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

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

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

Positional variability and transmission of satellite-free centromeric domains

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

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

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

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

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

Conclusions

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

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

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

Methods

Cell lines

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

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

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

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

Chromatin Immunoprecipitation (ChIP)

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

Cytogenetic analysis

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

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

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

Bioinformatic analysis of ChIP-seq data

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

SNV analysis

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

Southern blotting and quantitative PCR (qPCR)

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

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

Data access

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

Acknowledgments

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Geigl EM, Bar-David S, Beja-Pereira A, Cothran EG, Giulotto E, Hrabar H, Oyunsuren T, Pruvost M. 2016. Genetics and paleogenetics of equids. In *Wild equids* (ed. Ransom JI, Kaczensky P), pp. 87–104. Johns Hopkins University Press, Baltimore, MD.

Giulotto E, Saito I, Stark GR. 1986. Structure of DNA formed in the first step of CAD gene amplification. *EMBO J* **5:** 2115–2121.

Giulotto E, Raimondi E, Sullivan K. 2017. The unique DNA sequences underlying equine centromeres. *Prog Mol Subcell Biol* **56:** 337–354.

Gong Z, Wu Y, Koblízková A, Torres GA, Wang K, Iovene M, Neumann P, Zhang W, Novák P, Buell CR, et al. 2012. Repeatless and repeat-based centromeres in potato: implications for centromere evolution. *Plant Cell* **24:** 3559–3574.

Han Y, Zhang Z, Liu C, Liu J, Huang S, Jiang J, Jin W. 2009. Centromere repositioning in cucurbit species: implication of the genomic impact from centromere activation and inactivation. *Proc Natl Acad Sci* **106:** 14937–14941.

Hayden KE, Strome ED, Merrett SL, Lee HR, Rudd MK, Willard HF. 2013. Sequences associated with centromere competency in the human genome. *Mol Cell Biol* **33:** 763–772.

Henikoff S, Furuyama T. 2010. Epigenetic inheritance of centromeres. *Cold Spring Harb Symp Quant Biol* **75:** 51–60.

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

Hori T, Kagawa N, Toyoda A, Fujiyama A, Misu S, Monma N, Makino F, Ikeo K, Fukagawa T. 2017. Constitutive centromere-associated network controls centromere drift in vertebrate cells. *J Cell Biol* **216:** 101–113.

Huang J, Zhao Y, Bai D, Shiraigol W, Li B, Yang L, Wu J, Bao W, Ren X, Jin B, et al. 2015. Donkey genome and insight into the imprinting of fast karyotype evolution. *Sci Rep* **5:** 14106.

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

Kalitsis P, Choo KA. 2012. The evolutionary life cycle of the resilient centromere. *Chromosoma* **121:** 327–340.

Karpen GH, Allshire RC. 1997. The case for epigenetic effects on centromere identity and function. *Trends Genet* **13:** 489–496.

Kobayashi T, Yamada F, Hashimoto T, Abe S, Matsuda Y, Kuroiwa A. 2008. Centromere repositioning in the X chromosome of XO/XO mammals, Ryukyu spiny rat. *Chromosome Res* **16:** 587–593.

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

Leeb T, Vogl C, Zhu B, de Jong PJ, Binns MM, Chowdhary BP, Scharfe M, Jarek M, Nordsiek G, Schrader F, et al. 2006. A human-horse comparative map based on equine BAC end sequences. *Genomics* **87:** 772–776.

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

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

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

Marshall OJ, Chueh AC, Wong LH, Choo KH. 2008. Neocentromeres: new insights into centromere structure, disease development, and karyotype evolution. *Am J Hum Genet* **82:** 261–282.
Martins NM, Bergmann JH, Shono N, Kimura H, Larionov V, Masumoto H, Earnshaw WC. 2016. Epigenetic engineering shows that a human centromere resists silencing mediated by H3K27me3/K9me3. *Mol Biol Cell* **27:** 177–196.
Mayer W, Niveleau A, Walter J, Fundele R, Haaf T. 2000. Embryogenesis: demethylation of the zygotic paternal genome. *Nature* **403:** 501–502.
McKinley KL, Cheeseman IM. 2015. The molecular basis for centromere identity and function. *Nat Rev Mol Cell Biol* **17:** 16–29.
Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D, et al. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol* **14:** R10.
Mendiburo MJ, Padeken J, Fülöp S, Schepers A, Heun P. 2011. *Drosophila* CENH3 is sufficient for centromere formation. *Science* **334:** 686–690.
Mondello C, Smirnova A, Giulotto E. 2010. Gene amplification, radiation sensitivity and DNA double-strand breaks. *Mutat Res* **704:** 29–37.
Montefalcone G, Tempesta S, Rocchi M, Archidiacono N. 1999. Centromere repositioning. *Genome Res* **9:** 1184–1188.
Musilova P, Kubickova S, Vahala J, Rubes J. 2013. Subchromosomal karyotype evolution in Equidae. *Chromosome Res* **21:** 175–187.
Nergadze SG, Belloni E, Piras FM, Khoriauli L, Mazzagatti A, Vella F, Bensi M, Vitelli V, Giulotto E, Raimondi E. 2014. Discovery and comparative analysis of a novel satellite, EC137, in horses and other equids. *Cytogenet Genome Res* **144:** 114–123.
Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, Schubert M, Cappellini E, Petersen B, Moltke I, et al. 2013. Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499:** 74–78.
Palmer DK, O'Day K, Margolis RL. 1990. The centromere specific histone CENP-A is selectively retained in discrete foci in mammalian sperm nuclei. *Chromosoma* **100:** 32–36.
Palmer DK, O'Day K, Trong HL, Charbonneau H, Margolis RL. 1991. Purification of the centromere-specific protein CENP-A and demonstration that it is a distinctive histone. *Proc Natl Acad Sci* **88:** 3734–3738.
Panchenko T, Black BE. 2009. The epigenetic basis for centromere identity. *Prog Mol Subcell Biol* **48:** 1–32.
Phanstiel DH, Boyle AP, Araya CL, Snyder MP. 2014. Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. *Bioinformatics* **30:** 2808–2810.
Piras FM, Nergadze SG, Poletto V, Cerutti F, Ryder OA, Leeb T, Raimondi E, Giulotto E. 2009. Phylogeny of horse chromosome 5q in the genus *Equus* and centromere repositioning. *Cytogenet Genome Res* **126:** 165–172.
Piras FM, Nergadze SG, Magnani E, Bertoni L, Attolini C, Khoriauli L, Raimondi E, Giulotto E. 2010. Uncoupling of satellite DNA and centromeric function in the genus *Equus*. *PLoS Genet* **6:** e1000845.
Plohl M, Meštrović N, Mravinac B. 2014. Centromere identity from the DNA point of view. *Chromosoma* **123:** 313–325.
Probst AV, Almouzni G. 2011. Heterochromatin establishment in the context of genome-wide epigenetic reprogramming. *Trends Genet* **27:** 177–185.
Purgato S, Belloni E, Piras FM, Zoli M, Badiale C, Cerutti F, Mazzagatti A, Perini G, Della Valle G, Nergadze SG, et al. 2015. Centromere sliding on a mammalian chromosome. *Chromosoma* **124:** 277–287.
Raimondi E, Piras FM, Nergadze SG, Di Meo GP, Ruiz-Herrera A, Ponsà M, Ianuzzi L, Giulotto E. 2011. Polymorphic organization of constitutive heterochromatin in *Equus asinus* ($2n = 62$) chromosome 1. *Hereditas* **148:** 110–113.
Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. 2014. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42:** W187–W191.
Raudsepp T, Lear TL, Chowdhary BP. 2002. Comparative mapping in equids: The asine X chromosome is rearranged compared to horse and Hartmann's mountain zebra. *Cytogenet Genome Res* **96:** 206–209.
Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29:** 24–26.
Ross JE, Woodlief KS, Sullivan BA. 2016. Inheritance of the CENP-A chromatin domain is spatially and temporally constrained at human centromeres. *Epigenetics Chromatin* **9:** 20.
Saito I, Groves R, Giulotto E, Rolfe M, Stark GR. 1989. Evolution and stability of chromosomal DNA coamplified with the CAD gene. *Mol Cell Biol* **9:** 2445–2452.
Santos F, Peters AH, Otte AP, Reik W, Dean W. 2005. Dynamic chromatin modifications characterise the first cell cycle in mouse embryos. *Dev Biol* **280:** 225–236.
Schueler MG, Sullivan BA. 2006. Structural and functional dynamics of human centromeric chromatin. *Annu Rev Genomics Hum Genet* **7:** 301–313.
Shang WH, Hori T, Toyoda A, Kato J, Popendorf K, Sakakibara Y, Fujiyama A, Fukagawa T. 2010. Chickens possess centromeres with both extended tandem repeats and short non-tandem-repetitive sequences. *Genome Res* **20:** 1219–1228.
Steiner CC, Mitelberg A, Tursi R, Ryder OA. 2012. Molecular phylogeny of extant equids and effects of ancestral polymorphism in resolving species-level phylogenies. *Mol Phylogenet Evol* **65:** 573–581.
Stoler S, Keith KC, Curnick KE, Fitzgerald-Hayes M. 1995. A mutation in *CSE4*, an essential gene encoding a novel chromatin-associated protein in yeast, causes chromosome nondisjunction and cell cycle arrest at mitosis. *Genes Dev* **9:** 573–586.
Sullivan BA, Karpen GH. 2004. Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. *Nat Struct Mol Biol* **11:** 1076–1083.
Tolomeo D, Capozzi O, Stanyon RR, Archidiacono N, D'Addabbo P, Catacchio CR, Purgato S, Perini G, Schempp W, Huddleston J, et al. 2017. Epigenetic origin of evolutionary novel centromeres. *Sci Rep* **7:** 41980.
van de Werken HJ, Haan JC, Feodorova Y, Bijos D, Weuts A, Theunis K, Holwerda SJ, Meuleman W, Pagie L, Thanisch K, et al. 2017. Small chromosomal regions position themselves autonomously according to their chromatin class. *Genome Res* **27:** 922–933.
Ventura M, Weigl S, Carbone L, Cardone MF, Misceo D, Teti M, D'Addabbo P, Wandall A, Björck E, de Jong PJ, et al. 2004. Recurrent sites for new centromere seeding. *Genome Res* **14:** 1696–1703.
Ventura M, Antonacci F, Cardone MF, Stanyon R, D'Addabbo P, Cellamare A, Sprague LJ, Eichler EE, Archidiacono N, Rocchi M. 2007. Evolutionary formation of new centromeres in macaque. *Science* **316:** 243–246.
Vidale P, Magnani E, Nergadze SG, Santagostino M, Cristofari G, Smirnova A, Mondello C, Giulotto E. 2012. The catalytic and the RNA subunits of human telomerase are required to immortalize equid primary fibroblasts. *Chromosoma* **121:** 475–488.
Voullaire LE, Slater HR, Petrovic V, Choo KH. 1993. A functional marker centromere with no detectable alpha-satellite, satellite III, or CENP-B protein: activation of a latent centromere? *Am J Hum Genet* **52:** 1153–1163.
Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, Lear TL, Adelson DL, Bailey E, Bellone RR, et al. 2009. Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* **326:** 865–867.
Yang F, Fu B, O'Brien PC, Nie W, Ryder OA, Ferguson-Smith MA. 2004. Refined genome-wide comparative map of the domestic horse, donkey and human based on cross-species chromosome painting: insight into the occasional fertility of mules. *Chromosome Res* **12:** 65–76.
Zeitlin SG, Baker NM, Chapados BR, Soutoglou E, Wang JY, Berns MW, Cleveland DW. 2009. Double-strand DNA breaks recruit the centromeric histone CENP-A. *Proc Natl Acad Sci* **106:** 15762–15767.
Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9:** R137.

Received October 11, 2017; accepted in revised form April 13, 2018.

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

Solomon G. Nergadze, Francesca M. Piras, Riccardo Gamba, et al.

Genome Res. published online April 30, 2018
Access the most recent version at doi:10.1101/gr.231159.117