Nucleosome organisation and CENP-C distribution at satellite-free centromeres in *Equus asinus*

A thesis presented to the National University of Ireland, Galway, for the degree of Doctor of Philosophy

by

Joseph Gerald William McCarter, B.Sc
Centre for Chromosome Biology
Department of Biochemistry
National University of Ireland, Galway

October 2016

Chair of Biochemistry: Prof. Noel F. Lowndes
Head of Department: Dr. Michael P. Carty
PhD Academic Supervisor: Prof. Kevin F. Sullivan

Table of Contents

List of figures

List of tables

List of commands

Acknowledgements

I would like to begin by expressing my sincerest thanks to my supervisor Prof. Kevin Sullivan for giving me the opportunity to work in his lab. He provided guidance, encouragement and always had a positive outlook when new ideas were brought to the table.

I would like to thank Prof. Elena Giulotto and all her lab members, past and present for welcoming me to their lab at various stages throughout the project. Thank you to the members of Kevin's lab, who all helped along the way. Thank you to all the members of the CCB, both past and present, whether it be advice in experimental approach, borrowing reagents or general craic, it was always a positive working environment.

I would like to thank my sister Amy - You were always very supportive throughout my PhD and especially in the last year. I really appreciate all the guidance, dinners, lunches and bottles of wine! Now that your on the home straight too, I'll do my very best to return the gesture.

To Teri and Michael – You're some craic! - Thank you for all the support, advice and good nights out. There are plenty more to come!

To the group, Phillip, Edward, Marianne, Brenda, Jamie, Ciaran, Mairéad, Joanne, Siobhan, Siobhan, Siobhan, Emile, Aoife, Declan, Shane, Eimear, Sinead, Sinead, Tracey, Claire, Gareth & David. I'm very lucky to have you all as friends. Thank you all for all the good times over the years and especially over the course of my PhD.

To Frankie "trousers" and Wolf – There are two things I associate with this PhD and you two (1) Friday night pints in the Drift Inn (2) That weekend you came out to Pavia to visit. Enough said!

To my family – It is with no doubt in my mind that this was only possible through your love and support. My parents, John and Breda, my sisters Cathy, Jane and Amy, my brother William, my brothers-in-law Paul and Robert and my beautiful niece and nephew, Holly-Jane and James. Thank you all so much for everything throughout my PhD.

Abbreviations

AN	Asino Nuovo
Avg	average
BAM	Binary sequence alignemnt/map format
BJ	Blackjack
bp	base pairs(s)
C-terminus	carboxy terminus
CATD	CENP-A targeting domain
CCAN	Constitutive centromere associated network of proteins
cen27L	Centromere on chromosome 27 - left allele
cen9R	Centromere on chromosome 9 - right allele
CENP	Centromere protein
ChIP	Chromatin immunoprecipitation
ChIP-seq	Chromatin immunoprecipitation Sequencing
CID	Centromere identifying protein
CREST	calcinosis, Raynaud's phenomenon, esophageal dysmotility, sclerodactyly, and telangiectasia
Cy5	Cyanine 5
DAPI	4' , 6-diamidino-2-phenylindole
DNA	Deoxyribonucleic acid
EAS	E. asinus
ECA	Equus caballus
ENC	Evolutionary new centromere
FFT	Fast Fourier Transform
FISH	Fluorescent in situ hybridisation
FRiP	Fraction of reads in peaks
g	gravity
HFD	Histone fold Domain
HJURP	Holiday junction recognition protein
HMM	Hidden Markov Model
HRP	horseradish peroxidase
IgG	Immunoglobulin G
IGV	Integrative genomics viewer
kb	kilobase pair(s)
kDa	kilodaltons
MACs	Model-based Analysis of ChIP-Seq
Med	Median
Mnase	micrococcal nuclease
Mono1	Mononucleosome 1 dataset
Mono2	Mononucleosome 2 dataset

mRNA	Messenger ribonucleic acid
N-terminus	amino terminus
NP-40	Nonidet-P40
PCR	Polymerase chain reaction
qPCR	Quantitative PCR
SAM	Sequence alignment/Map format
SD	Standard deviation
SDS	Sodium dodecyl sulphate
SNP	Single nucleotide polymorphism
TAE	Tris acetate EDTA
Trinuc	Trinucleosome dataset

Abstract

Centromeres are distinct epigenetic loci present in single copy per chromosome. Defined by the presence of the histone H3 variant CENP-A, centromeres play a vital role in chromosome segregation via the kinetochore and mitotic apparatus. CENP-C plays an important role in CENP-A deposition onto chromatin and acts as a primary anchor between the centromeric chromatin and kinetochore interface. Due to the typical association of centromeres with arrays of repetitive satellite DNA, chromatin profiling using next generation sequencing methods remains quite challenging and the subunit organization of centromeres at a molecular level is unclear. We present our analysis, which identified a system of naturally occurring satellite-free centromeres in *Equus asinus* – the domestic donkey. Using a ChIP-seq approach, donkey skin fibroblasts were processed using antibodies against CENP-A and CENP-C, revealing sixteen centromeres in the donkey genome that are not associated with classical satellite DNA sequences but rather are formed at satellite-free sequence domains. The sixteen centromere domains identified were shown to contain very comparable amounts of CENP-A and CENP-C at each centromere despite the genomic footprint occupied by the domains. In this study, we also present our analysis of the transmission of CENP-A domains from parents to offspring, giving more insight into the plasticity of centromeres. A native ChIP-seq approach has also been carried out to investigate nucleosome distribution across these centromere domains. Collectively, this study provides more insight into centromere structure and nucleosome organisation and strongly supports equids as a model system to study centromere behavior.

Chapter 1 - Introduction

1.1 The centromere

The centromere is the chromosomal locus, present in single copy per chromosome that is responsible for the faithful segregation of genomic material and inheritance from on generation to the next. Through a series of dynamic protein interactions involving core centromere-kinetochore subunits, the centromere establishes the site of kinetochore attachment, which interacts with the spindle microtubules and the mitotic apparatus, to mediate alignment of the chromosomes at the metaphase plate which ultimately segregate and are redistributed to the new daughter cells (Allshire and Karpen, 2008; Cheeseman and Desai, 2008; Cleveland et al., 2003).

The centromere, commonly referred to as the primary constriction, due to the highly condensed appearance of the underlying DNA, is typically associated with arrays of repetitive DNA (Waye and Willard, 1985). It is widely accepted that the repetitive DNA, called alpha-satellite in humans, is neither sufficient nor necessary for centromere function (Earnshaw and Migeon, 1985) but still seems to be the preferred substrate for the locus (Allshire and Karpen, 2008). This is demonstrated by the discovery, albeit rare occurrences of neocentromeres, which are built on non-repetitive DNA. Despite neocentromeres being built on unique sequence DNA, they are still able to retain complete functionality (Marshall et al., 2008a; Voullaire et al., 1993).

Centromere function is defined discretely by a specialised chromatin configuration and context through the presence of the histone H3 variant, CENP-A. The histone fold domain of CENP-A targets it to the centromere (Sullivan et al., 1994) and other structural properties of CENP-A nucleosomes directly relate to its function (Black et al., 2004, 2007a). CENP-A is strictly maintained and needs to be replenished in each cell cycle (Bodor et al., 2014; Jansen et al., 2007; Silva et al., 2012). The CENP-A assembly pathway is one that differs from most canonical nucleosomes as it is uncoupled from DNA replication (Shelby et al., 2000) and subsequent loading of the protein occurs later compared to canonical nucleosomes (Jansen et al., 2007) mediated by the CENP-A specific chaperone HJURP (Dunleavy et al., 2009; Foltz et al., 2009) .

There is a large family of centromere proteins that are constitutively expressed throughout the cell cycle called the constitutive centromere associated network of

proteins (CCAN) and are for the most part essentially involved in the structure and maintenance of the locus (Ando et al., 2002; Foltz et al., 2006; Hori et al., 2008a, 2008b; Okada et al., 2006). CENP-C is a crucial CCAN component and plays numerous roles in the CENP-A assembly pathway (McKinley and Cheeseman, 2014), centromeric chromatin structure and kinetochore assembly (Przewloka et al., 2011).

In higher eukaryotes, many aspects of centromere structure and function have been investigated, however much more progress is required in order to fully understand fundamental properties that coordinate and maintain centromere-kinetochore, structure and function. Some of the key aspects surrounding centromere organisation and regulation are discussed here.

1.2 Centromeres and DNA

It has been well documented that centromeric DNA varies among eukaryotes (Cleveland et al., 2003). "Point" centromeres found in budding yeast, *Saccharomyces cerevisiae,* assemble over a 125bp sequence. Complete centromere function is specified over this 125bp region which includes three centromere DNA elements or CDEs: CDEI, CDEII and CDEIII (Fitzgerald-Hayes et al., 1982). A region in the CDEIII is responsible for targeting CENP-A^{CSE4} to the middle of the AT rich CDEII through the recruitment a set of specific DNA binding proteins (Furuyama and Biggins, 2007; Meluh et al., 1998). Regional centromere domains in *Schizosaccharomyces pombe* are well studied. Functional activity is dictated by a central core which is packaged over a 10 kb domain that is flanked by inverted outer repeats (Castillo et al., 2007; Clarke, 1998). The CENP-A homologue, Cnp1^{CENP-A} binds the central core and the flanking repeat regions of the *S. pombe* centromeres (Takahashi et al., 2000). No common unique identifying sequence is found at all *S. pombe* centromeres that specifies centromere maintenance (Clarke, 1998; Ngan and Clarke, 1997; Partridge et al., 2000). A diffuse centromere in nematodes such as *Caenorhabditis elegans*, some insects, and plants assembles a ''holocentromere'' which extends along the entire length of the chromosome (Hughes-Schrader and Ris, 1941). Recent ChIP-chip studies revealed that the *C. elegans* CENP-A nucleosomes occupy repeat-less regions of 10–12 kb (Gassmann et al., 2012). One proposal is that these are CENP-A ''seeds'' which are dispersed along the chromosome

and by some mechanism associate to direct kinetochore assembly and function (Fukagawa and Earnshaw, 2014). In *Drosophila melanogaster*, the centromere domains are composed of islands of simple satellite repeats. These simple satellites contain no detectable consensus sequence motifs, that are common to all centromeres (Sun et al., 1997) and are occupied by the CENP-A homologue, CENP-A^{CID}. Centromeres in humans are built on alpha satellite DNA. There is typically a 171bp monomer unit that is tandemly repeated up to 5,000 kb (Willard, 1985). Centromeric associated satellite DNAs are rapidly evolving and little conservation is seen between among metazoans (Henikoff et al., 2001). CENP-A nucleosomes are bound to these α-satellite arrays (Vafa and Sullivan, 1997).

While "point" centromeres in budding yeast are established by the presence of a specific DNA sequence, this is not the case in higher eukaryotes. However, it is important to note that CENP-B recognizes a 17bp DNA motif commonly known as the "CENP-B box" (Masumoto et al., 2004) which is present across alpha satellite domains. CENP-B was shown not to be an essential centromere binding protein (Ohzeki et al., 2002; Okada et al., 2007) and is typically not associated at neocentromeres where satellite DNA is not present (Voullaire et al., 1993). Neocentromeres and centromeres share many structural and functional properties with one another, with one notable difference being the underlying DNA sequence. Neocentromeres are typically devoid of higher order repeat (HOR) DNA (Marshall et al., 2008a). The association of unique sequence DNA with neocentromeres verifies the epigenetic nature and definition of centromeric loci. The first neocentromere discovered was observed in a patient with mild developmental delay (Voullaire et al., 1993). A supernumery marker chromosome formed due to rearrangement in chromosome 10 in the patient. The formation of a ring chromosome "rdel(10) with a fully functional centromere and a linear chromosome "mardel(10)" lacking alpha-satellite DNA. The mardel(10) marker chromosome tested negative for alpha satellite DNA but positive for centromere marker – CENP-A (Voullaire et al., 1993). 90 or more cases of neocentromeres have been described in human clinical cases since the discovery of mardel(10) (Marshall et al., 2008a).

Evidence suggests that neocentromeres behave identically to satellite containing

centromeres. They have been shown to bind all known centromere proteins apart from the well characterised CENP-B (Saffery et al., 2000) . A role for CENP-B was suggested based on the fact that it enhanced de novo centromere formation in artificial chromosome assays (Ohzeki et al., 2002; Okada et al., 2007) possibly by modifying the histone H3 chromatin landscape over the centromere region (Okada et al., 2007) which in turn would affect the CENP-A assembly pathway mediated through HJURP (Bergmann et al., 2011). This is in contrary to the idea before which dismisses an essential role for CENP-B by demonstrating the viability of CENP-B null mice (Hudson et al., 1998; Kapoor et al., 1998; Perez-Castro et al., 1998). It is worth noting that recent work has shown depletion of CENP-B in human and mouse cells results in a higher frequency of chromosome missegregation and that functional neocentromeres which lack CENP-B tend to also have a higher frequency of chromosome missegregation (Fachinetti et al., 2015). This study also suggested a role for CENP-B in typical satellite containing centromeres where CENP-B stabilises centromeric CENP-A and CENP-C in a sequence binding dependant manner.

These data suggest that while repetitive DNA tends to be the preferred substrate for CENP-A binding it is not essential for centromere function, therefore centromeres are defined epigenetically through distinct chromatin composition and context which is centered around the presence of CENP-A.

1.3 **CENP-A – defining centromere identity**

Centromere binding proteins were originally discovered using autoantisera isolated from patients with scleroderma (Moroi et al., 1980). It was not until a few years later the name “Centromere protein A” was coined when the first three centromere proteins CENP-A, CENP-B and CENP-C were identified (Earnshaw and Rothfield, 1985). CENP-B was the first of these three centromere proteins to be cloned (Earnshaw et al., 1987), followed by CENP-C (Saitoh et al., 1992). Following digestion by micrococcal nuclease, CENP-A was isolated from mononucleosome preparations (Palmer and Margolis, 1985) and later shown to co-purify with core histone particles (Palmer et al., 1987) suggesting CENP-A was a centromere specific histone. By direct isolation of CENP-A cDNA, it was later confirmed that CENP-A was indeed a histone-H3 like subunit and

contained a histone-fold domain (HFD) that was responsible for centromeric targeting (Sullivan et al., 1994). The histone-H3 like HFD shares a ~62% homology with the HFD of histone H3 and very little identity is seen at the amino terminal domains (Sullivan et al., 1994). The CENP-A targeting domain (CATD) is a 22 amino acid (aa) motif unique to CENP-A that is located within the HFD of CENP-A. The CATD is responsible for targeting CENP-A to the centromere (Black et al., 2007a). Histone H3 does not contain the CATD however when modified to contain the CATD histone H3 is itself sufficient to maintain centromere function (Black et al., 2004, 2007b; Fachinetti et al., 2013). Unlike the canonical histones that usually assemble onto chromatin in S phase, this is not the case with CENP-A as this process is uncoupled from DNA replication in human cells (Shelby et al., 1997, 2000). CENP-A mRNA levels accumulate in late G2 with protein deposition following in early G1 after a decrease in mitotic associated CDK activity (Jansen et al., 2007; Shelby et al., 1997, 2000).

1.4 CENP-A nucleosomes

As previously discussed the histone fold domain of CENP-A shares a 62% identity with that of histone H3. The N terminal domain on the other hand diverges substantially (Sullivan et al., 1994). The octameric human CENP-A nucleosome contains two copies of the core histones, H2A, H2B and H4 were CENP-A replaces H3 (Yoda et al., 2000). The 22 amino acid region of the HFD between the first loop and second α-helix (L1-α2) contain the CATD, which is essential for targeting CENP-A to the centromere (Black et al., 2004, 2007b). Additional regions within CENP-A nucleosomes contribute to centromere function through binding to a subset of the CCAN (Constitutive centromere associated network of proteins) proteins, in particular CENP-C and CENP-N (Carroll et al., 2010). CENP-N binds exclusively to CENP-A nucleosomes arbitrated by the CATD and does not bind canonical Histone H3 nucleosomes (Carroll et al., 2009, 2010; Logsdon et al., 2015). The CENP-A nucleosomes can recruit CENP-C through its N-terminal tail (Logsdon et al., 2015), the CATD (Logsdon et al., 2015; Westhorpe et al., 2015) and the C-terminal tail (Carroll et al., 2010; Guse et al., 2011; Kato et al., 2013). CENP-C in turn stabilises CENP-A (Fachinetti et al., 2015) in part through induced structural changes in CENP-A nucleosomes (Falk et al., 2015).

Approximately 147bp of DNA is wrapped around the core nucleosome (Luger et al., 1997) however recent studies show that CENP-A nucleosomes can protect a minimum length of ~120bp and up to canonical nucleosome DNA protection length of 147bp (Hasson et al., 2013; Tachiwana et al., 2011). Shorter DNA protection of CENP-A nucleosomes is proposed for the following reasons; CENP-A nucleosomes have been shown to induce supercoils on plasmid DNA less efficiently than in Histone H3 nucleosomes therefore DNA has the potential to be partially unwrapped on CENP-A nucleosomes (Tachiwana et al., 2011); Structural differences between CENP-A and histone H3 at the amino terminal regions (Figure 1.1) revealed that DNA ends are more stabilised in histone H3 nucleosomes compared to that of CENP-A nucleosomes and that the length of the amino terminal 5' of the aN helix seems to be responsible for this (Tachiwana et al., 2011).

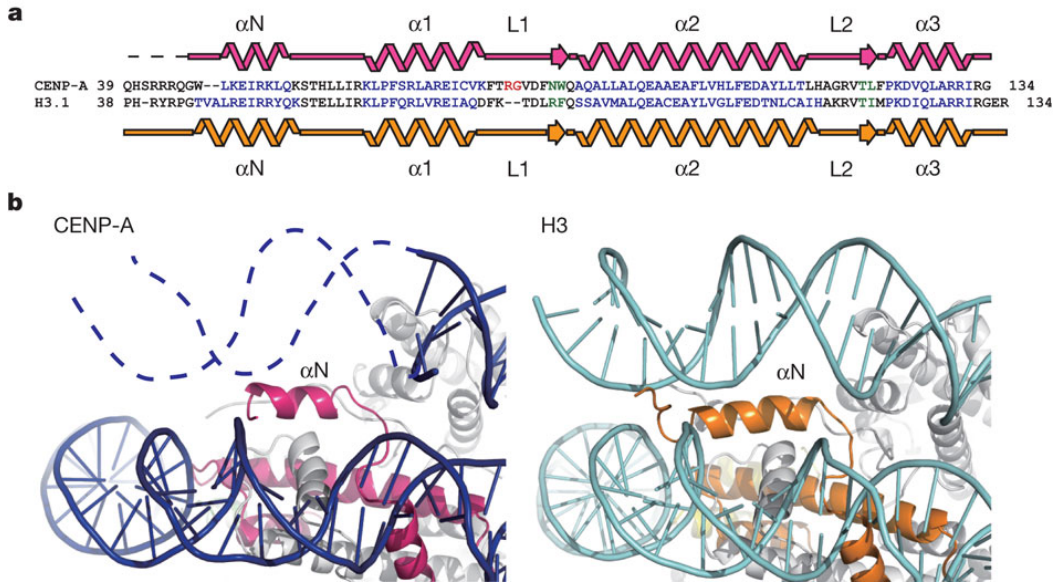


Figure 1.1 Structural comparisons of CENP-A and H3 nucleosomes
Reprinted by permission from Macmillan Publishers Ltd: Nature (Crystal structure of the human centromeric nucleosome containing CENP-A), copyright (2011)

(a) Secondary structure of CENP-A in the nucleosome. The sequences of human CENP-A and H3 are aligned with the secondary structure elements. (b) Close-up views of the αN helices and the DNA edge regions of the CENP-A (left panel) and H3 (right panel) nucleosomes. The dashed line in the left panel shows the DNA region that is not visible in the crystal structure. The CENP-A and H3 molecules are shown in magenta and orange, respectively.

While CENP-A nucleosomes can protect less DNA than canonical nucleosomes (Hasson et al., 2013; Tachiwana et al., 2011), this feature is shown to be enhanced upon CENP-C

binding (Falk et al., 2015). These observations confer functional aspects of the structure of CENP-A nucleosomes that reinforce the epigenetic identity of centromeres.

1.5 CENP-A deposition

In order for centromere identity and function to be maintained, CENP-A must be faithfully replenished at centromeres after each cell cycle (Bodor et al., 2013; Jansen et al., 2007). Failure to replenish CENP-A leads to chromosome segregation errors, which in turn affects genome stability (Bodor et al., 2013; Fachinetti et al., 2013). CENP-A deposition is part of a highly regulated pathway that ensures CENP-A is stably replenished at centromeres in each cell cycle, which subsequently permits accurate kinetochore assembly (McKinley and Cheeseman, 2016). CENP-A deposition pathway can be split into three distinct phases: initiation, deposition, and maintenance (Stellfox et al., 2013).

The bulk of the canonical histone proteins are deposited along DNA during S phase which is mediated by ASF1 and CAF-1 (Mello and Almouzni, 2001). While centromeric DNA is replicated in S phase, CENP-A mRNAs peak in late G2 (Shelby et al., 2000). CENP-A nucleosomes are stably retained through S phase and this requires the CATD (CENP-A targeting domain) (Bodor et al., 2013; Jansen et al., 2007). Immunofluorescence staining on replicated centromeric chromatin fibers showed that CENP-A nucleosomes are distributed equally onto daughter strands (Dunleavy et al., 2011; Ross et al., 2016). Human CENP-A loading occurs in a replication independent manner during G1 and requires exit from mitosis with reduced mitotic CDK activity (Jansen et al., 2007; Silva et al., 2012). The process of CENP-A nucleosome re-distribution over two daughter strands in S phase leaves gaps in the centromeric chromatin where CENP-A nucleosomes previously occupied and as such leaves the CENP-A complement at half-level compared to G1. These gaps are then filled by histone H3.3 nucleosomes which act a place holders as they are exchanged with novel CENP-A nucleosomes in G1 (Dunleavy et al., 2011).

In *S. pombe*, CENP-A^{Cnp1} incorporation requires Mis16 and Mis18 (Fujita et al., 2007; Hayashi et al., 2004). The human orthologs, RbAp46/48 and the three subunit, Mis18 complex (comprised of hMis18α, hMis18β, and M18BP1/KNL-2), also play a role in

CENP-A deposition at centromeres. The Mis18 complex localises to centromeres in late anaphase (Fujita et al., 2007) and is recruited by interacting with CENP-C through the Mis18BP1 subunit of the Mis18 complex (Dambacher et al., 2012; Moree et al., 2011). Initiation of CENP-A incorporation requires phosphorylation of the Mis18 complex by Plk1 and the rapid depletion of Mis18 levels in G1 contemporaneously with CENP-A deposition indicate it plays a role in licensing chromatin for CENP-A loading (Fujita et al., 2007; Hayashi et al., 2004; McKinley and Cheeseman, 2014). Along with the Mis18 complex CENP-A requires a CENP-A specific chaperone called HJURP (Holiday junction recognition protein) (Dunleavy et al., 2009; Foltz et al., 2009). HJURP recognises CENP-A over histone H3 via its N-terminal CENP-A binding domain and the CATD of CENP-A (Bassett et al., 2012; Hu et al., 2011; Shuaib et al., 2010; Zhou et al., 2011). Parallel with the timing of CENP-A deposition, HJURP exists as a dimer and only localises to centromeres in G1 (Dunleavy et al., 2009; Foltz et al., 2009; Zasadzińska et al., 2013). Dimerization of HJURP which is mediated through its C-terminus, is required for de novo CENP-A nucleosome deposition (Zasadzińska et al., 2013).

Once HJURP mediated deposition of CENP-A nucleosomes has taken place, a subsequent "maturation" process happens to maintain stability of centromeres. At this point, the ATP-dependant nucleosome remodelling and spacing factor (RSF) complex (Perpelescu et al., 2009), male germ cell Rac GTPase-activating protein (MgcRacGAP) (Lagana et al., 2010) associate with newly deposited CENP-A and provide stability to CENP-A nucleosomes. CENP-C has also been shown to provide stability to CENP-A nucleosomes by rigidifying nucleosome structure (Falk et al., 2015). Recent studies estimated that human centromeres contain ~400 CENP-A molecules, meaning each centromere would contain approximately 200 CENP-A nucleosomes (Bodor et al., 2014). This study showed that the level CENP-A across all centromeres was present in a 2.5 fold excess. These data are reflective of CENP-A maintenance indicating that centromeres aim to maintain a relatively constant level of CENP-A. While it is evident that CENP-A deposition is tightly regulated, little is known about how CENP-A is organised at centromeres.

1.6 CENP-A nucleosome organisation within centromeric chromatin

CENP-A nucleosomes mark the region of centromeric function and specify the site for kinetochore assembly. A central question in the field aims to elucidate the context in which CENP-A nucleosomes are organised. The importance of this question is ever resilient, as pointed out in numerous studies that have explored how the organisation of the centromeric chromatin fiber in linear space relates to the condensed 3-dimensional architecture which supports proper function (Blower et al., 2002; Marshall et al., 2008b; Ribeiro et al., 2010; Zinkowski et al., 1991).

In situ hybridization experiments on stretched chromatin fibers stained for CENP-A showed that the centromere forming region consisted of CENP-A containing blocks interspersed with histone H3 and their distribution across the linear chromatin fiber shows little uniformity or organization other than that they are interspersed with canonical histone H3 nucleosomes. (Blower et al., 2002; Zinkowski et al., 1991). From these data "repeat-subunit" model was proposed. Here, the centromeric chromatin was proposed to be ordered as an amphipathic coil where CENP-A would be present on one side, establishing the inner kinetochore plate on the external face of the mitotic chromosome.

Three-dimensional deconvolution light microscopy on fly and human metaphase chromosomes showed that CENP-A exhibited a cylindrical-like structure, which extended through the half the width and the entire height and length of the primary constriction region (Blower et al., 2002). Quantitation of CENP-A along with the core histones on stretched chromatin fibers revealed that CENP-A domains contained histones H2A, H2B but not H3 indicating their common use in both nucleosome particles. Notably in flies and humans, dimethylated histone H3 on lysine 4 (H3K4me2) was found to be enriched in H3 subdomains within the centromeric chromatin domain (Sullivan and Karpen, 2004). How these interspersed blocks of CENP-A and histone H3 on centromeric chromatin fold to form an outer surface substructure of CENP-A nucleosomes is currently unknown. However, major attempts to determine this substructure were carried out. Electron microscopy studies on alpha-satellite containing centromeres revealed that CENP-A occupies a much narrow domain than previously observed. Here, CENP-A nucleosomes were shown to occupy a compact domain at the outer surface encompassing two thirds of the primary

constriction length, with the only 6-8% of the total volume of the constriction DNA actually occupied by the protein (Marshall et al., 2008b). An earlier study in neocentromeres concluded a decreased level of overall CENP-A binding at neocentromeres compared to typical centromeres (Irvine et al., 2005). In a more recent study, the neocentromere in the PDNC-4 cell line was shown to bind ~25% less CENP-A compared to the satellite containing centromeres within that cell line (Bodor et al., 2014). Despite neocentromeres binding less CENP-A, Marshall et al, (2008) found similar CENP-A occupancy structure and volume per constriction length when their quantitative metric was applied to the "mardel(10)" neocentromere. These results indicated that despite the differences in the total amount of CENP-A between the two types of centromere, the CENP-A complement is present in a common higher order structure.

This observation is strongly supported by another study, which used super-resolution fluorescence microscopy to examine CENP-A distribution relative to histone modifications and kinetochore proteins in chicken cells (Ribeiro et al., 2010). Destabilising higher order chromatin structure showed clearly the interspersed blocks of CENP-A and histone H3 nucleosomes, a pattern that is now considered conserved (Blower et al., 2002; Sullivan and Karpen, 2004; Zinkowski et al., 1991). By inducing destabilized higher order chromatin structure the study revealed important structural roles for CCAN components. Ribeiro et al. (2010) showed that these CCAN components, in particular CENP-C are required for the structural integrity of mitotic kinetochores. A model was developed, that describes alternating CENP-A and histone H3 blocks folding into a boustrophedron loop topology in which CENP-C (along with other CCAN components) provides structural crosslinks between the loops by binding CENP-A. This model would allow CENP-A and histone H3 nucleosomes to position at the outer face of the centromeric chromatin and permit subsequent binding of the kinetochore machinery.

Collectively these data strongly suggest a higher order centromeric chromatin structure. However, while many advances have been made in investigating nucleosome configuration at the centromere domain, limitations in resolution exist

when immunostaining on stretched chromatin fibers. In order to truly elucidate the molecular architecture of centromeric chromatin, base pair resolution experiments are required.

1.7 The constitutive centromere associated network of proteins

CENP-A plays its vital role in (1) maintaining the centromeric chromatin environment through its consistent replenishment after each cell cycle (2) specifying the site of assembly of the kinetochore. In order to establish the kinetochore, a primary set of centromere-kinetochore proteins assemble which are known as the constitutive centromere associated network of proteins (CCAN) (Cheeseman and Desai, 2008). These proteins are denoted by alphabetic nomenclature after CENP-A –B & -C (CENP-T CENP-W, CENP-S, CENP-X, CENP-L CENP-N, CENP-H, CENP-I, CENP-K, CENP-M, CENP-P, CENP-O, CENP-R, CENP-Q, CENP-U) (Foltz et al., 2006; Izuta et al., 2006; Nishihashi et al., 2002; Okada et al., 2006; Saitoh et al., 1992; Sugata et al., 1999). Some of the key complexes and interactions that result in kinetochore assembly are discussed here.

A number of biochemical approaches identified centromere interacting proteins; CENP-A, -B, -C, -H, -I and Mis12 (Goshima et al., 2003; Liu et al., 2003; Nishihashi et al., 2002; Sugata et al., 2000). These proteins were said to be part of the "CEN-complex" (Obuse et al., 2004). A series of affinity purification experiments (TAPs) were carried out in order to establish the complex of proteins associated with CENP-A and histone H3 nucleosomes across the centromeric chromatin landscape (Foltz et al., 2006). Three proteins, CENP-M, CENP-N and CENP-T were amongst the proteins found to be associated with CENP-A nucleosomes (Foltz et al., 2006). Expressing fluorescently tagged versions of CENP-M, CENP-N and CENP-T as well as affinity purifying using localisation and affinity purification (LAP) tagged fusions confirmed their targeting to centromeres and concluded that these three proteins were associated with the CENP-A NAC along with CENP-C and CENP-H (Foltz et al., 2006; Saitoh et al., 1992; Sugata et al., 2000). Through the same method CENP-U(50) was also found to be associated with this complex (Foltz et al., 2006).

Small interference RNA (siRNA) depletion of a selection of these components, in particular CENP-M, CENP-N and CENP-T resulted in disrupted targeting of the CENP-A

NAC. CENP-M and CENP-N depletion resulted in the loss of each other as well as the loss of CENP-H. The depletion of CENP-T resulted in complete loss of CENP-M (Foltz et al., 2006). CENP-U(50) depletion had no effect on targeting of CENP-A, CENP-B, Hec1, CENP-E, CENP-F, Mis12 or Aurora B to the centromere however its depletion resulted in complete loss of CENP-O and CENP-P at interphase and mitotic centromeres concluding the CENP-U(50) is required for their loading (Foltz et al., 2006). Further siRNA approaches revealed that CENP-M, CENPN and CENP-T were required for mitotic progression as depletion of these components increased the number of cells in mitosis along with failure in chromosomes congression with CENP-T depletion having the most pronounced effect (Foltz et al., 2006).

Contemporaneously, a similar approach was taken by (Okada et al., 2006) in DT40 cells, which focused on inner kinetochore proteins and subsequently identified the CENP-H-I complex. Affinity purifications using FLAG or GFP tags isolated and identified several novel centromere associated proteins; CENP-K, CENP-L, CENP-M, CENP-O, and CENP-P (Okada et al., 2006). LAP tagged affinity purification experiments in human cells with CENP-H, CENP-O and CENP-P then identified CENP-N (as BM039), CENP-Q, CENP-R, and CENP-U. Based on functional and localisation phenotypes these proteins were allocated into groups; (1) The CENP-H class (CENP-H, -I, -K, -L) that when depleted produced a strong mitotic delay with localisation depending on M-class and required for CENP-A incorporation; (2) The CENP-M class (Just CENP-M) that when depleted caused strong mitotic delay with localisation depending on CENP-H, -O classes, which is also required for CENP-A incorporation and finally the (3) CENP-O class (CENP-O, -P, -Q, -U(50)) then when depleted caused a mild non-fatal mitotic delay, with localisation requiring the CENP-H and CENP-M classes which were partially required for CENP-A incorporation (Okada et al., 2006). A series of subsequent studies using *myc* tagged fusions of different CCAN components revealed a direct interaction between CENP-L and CENP-N suggesting these two CCAN components form a complex (Carroll et al., 2009). Protein co-expression experiments with CENP-I (and a subdomain of CENP-I), CENP-H, CENP-K and CENP-M revealed that CENP-M enhances stabilization the quaternary complex indicating that CENP-M is part of a “CENP-HIKM” complex (Basilico et al., 2014).

The CENP-O class (CENP-O, -P, -Q, -U(50), -R) of proteins play an interdependent role for centromere/kinetochore localisation (Amaro et al., 2010; Foltz et al., 2006a; Hori et al., 2008a). Studies in *E.coli* and chicken DT40 cells showed that CENP-O, -P, -Q and –U form a stable complex to which CENP-R associates downstream (Hori et al., 2008a). FRET (Förster resonance energy transfer) measurements of the "CENP-PORQU complex" show that once bound to the centromere/kinetochore interface that the proteins are in close proximity to one another indicating again that these set of proteins are imbedded in the major CCAN complex (Eskat et al., 2012). The CENP-O class/CENP-PORQU complex proteins play a role in spindle damage protection and may help prevent premature chromatid separation via CENP-Q's association with microtubules (Amaro et al., 2010; Hori et al., 2008a).

CENP-S was another protein identified as part of the CENP-A distal (CAD) complex through a series of TAP fusion pull down experiments with CENP-M and CENP-U(5) (Foltz et al., 2006). Characterisation of CENP-S revealed that it had a binding partner, CENP-X and that both proteins contained a histone fold domain (Amano et al., 2009) which presumably functions in centromere targeting of the complex and provides structural support in the CENP-TWSX DNA supercoiling heterotetramer found *in vitro* (Nishino et al., 2012; Takeuchi et al., 2014). More recent studies show that CENP-S and CENP-X assemble in late S phase/early G2 contemporaneously with CENP-T/W providing more evidence regarding the existence of a CENP-TWSX particle *in vivo* (Dornblut et al., 2014).

(Foltz et al., 2006) identified CENP-T through its association with the CENP-A NAC. CENP-T is a HFD containing member of the CCAN that was discovered to form a complex with CENP-W (Hori et al., 2008b). CENP-T/W has also been shown to form a heterotetrameic complex with the CENP-S/X, which may act as a nucleosome like particle in vitro (Nishino et al., 2012). CENP-T provides a way of linking the kinetochore with the centromeric chromatin through its phosphorylation dependant N-terminal interaction with the Ndc80 complex (Nishino et al., 2013; Schleiffer et al., 2012). Tethering experiments in DT40 cells showed that when CENP-T along with CENP-C is tethered to an ectopic site that this is sufficient to assemble a kinetochore in the absence of CENP-A nucleosomes (Gascoigne et al., 2011). The CENP-T/W complex directly associates with H3 nucleosomes through protein DNA contact sites with the requirement of the HFD but does not bind with CENP-A nucleosomes (Hori et al.,

2008b). CENP-T/W localisation to centromeres is downstream of other CCAN components (Basilico et al., 2014; Carroll et al., 2010) and is not impacted by reduced CENP-A levels (Fachinetti et al., 2013). It depends on initial CENP-C recruitment and has been shown to directly interact with CENP-HIKM (Basilico et al., 2014; Klare et al., 2015), although, CENP-T/W doesn't bind CENP-C tightly in the absence of other binding partners (Klare et al., 2015). CENP-T/W is not retained at centromeres throughout the cell cycle (Jansen et al., 2007; Prendergast et al., 2011). The *de novo* loading of CENP-T/W to centromeres happens in S/G2 phase of the cell cycle and is required for centromere and kinetochore maintenance (Prendergast et al., 2011) which more recently was shown to be regulated by the FACT (facilitates chromatin transcription) complex (Prendergast et al., 2016).

While major progress has been made in the field regarding CCAN components and kinetochore assembly, the two primary proteins involved in establishing this network are CENP-A and CENP-C. Understanding how these essential components are organised on the centromeric chromatin fiber will provide a way to elucidate the structural properties of the centromere-kinetochore complex.

1.8 CENP-C

CENP-C is included in the CCAN, which was first identified along with CENP-A and CENP-B as an autoantigen recognised by anti-centromere antibodies (Earnshaw and Rothfield, 1985). CENP-C is a 107kDa protein that is essential for centromere maintenance and kinetochore function (Earnshaw and Rothfield, 1985; Fukagawa et al., 1999; Saitoh et al., 1992). Immuno-electron microscopy showed CENP-C localisation to the inner kinetochore plate which is the interface between the centromeric chromatin and the outer kinetochore plate (Saitoh et al., 1992). As a member of the CCAN, CENP-C is localised at centromeres throughout the cell cycle (Saitoh et al., 1992) but a progressive increasing accumulation of CENP-C is seen from S phase onwards (Knehr et al., 1996). CENP-C levels peak in G1 contemporaneously with CENP-A loading (Jansen et al., 2007; Tomkiel et al., 1994). In addition to this, CENP-C co-purifies with CENP-A nucleosomes (Foltz et al., 2006) which provides evidence for interaction of the two proteins. CENP-C has only been shown to be present at active centromeres and therefore is a functional centromere marker (Voullaire et al., 1993). Depletion of CENP-C causes chromosome missegregation,

mitotic delay and apoptosis (Fukagawa and Brown, 1997; Fukagawa et al., 1999) and while CENP-A is the primary epigenetic marker of centromeres it should be noted that induced ectopic kinetochores by replacing DNA binding regions with CENP-C and CENP-T bypasses requirement for CENP-A nucleosomes (Gascoigne et al., 2011). Together these data enforce the importance of CENP-C in centromere-kinetochore assembly and maintenance.

CENP-C contains a CENP-A recognition region called the central domain which allows CENP-C to bind a discrete locus on CENP-A (Carroll et al., 2010; Kato et al., 2013). The central domain is located between amino acids 426-527 (Kato et al., 2013). Within the central domain region a number of consecutive positively charged residues are present which are likely to bind DNA. CENP-C is also centrally involved in the CENP-A assembly pathway. CENP-A assembly is regulated by Mis18α/β which forms a complex with M18BP1 (Fujita et al., 2007). Along with the HJURP complex (HJURP, Npm1, CENP-A, and H4), Mis18α/β and M18BP1 are recruited to centromeres (Dambacher et al., 2012; Dunleavy et al., 2009; Foltz et al., 2009; Moree et al., 2011). CENP-C binds directly to M18BP1 (Dambacher et al., 2012; McKinley and Cheeseman, 2014; Moree et al., 2011).

CENP-C has been shown to co-localise with CENP-A through ChIP-chip and ChIP-seq approaches (Smith et al., 2011; Wade et al., 2009). Through a series of in-situ hybridization experiments and eventually whole genome sequencing, the centromere on chromosome 11 of the domestic horse was shown to be completely devoid of satellite DNA (Carbone et al., 2006; Wade et al., 2009). ChIP-chip analysis using antibodies with a high titer for CENP-A and CENP-C and a tiling array spanning the centromere region in chromosome 11 in the domestic horse showed co-localised enrichment for centromere DNA for both proteins (Wade et al., 2009). In *Neurospora crassa* the centromere repeats are heterogeneous, unlike the typical homogeneous arrays of satellite DNAs found in most organisms. This is due to the activity of a region-specific premeiotic mutator phenomenon called “repeat-induced point mutation” (RIP). The heterogenous repetitive nature of the centromere regions in *N. crassa* allowed for ChIP-seq studies with CENP-A and CENP-C which also revealed compete co-localisation of the two centromere proteins (Smith et al., 2011). ChIP-seq approaches in *C. elegans* show that holocentric chromatin domains exhibit CENP-C

binding at a selection of CENP-A binding regions indicating preferential kinetochore binding sites in the nematode (Steiner and Henikoff, 2014). Together these data show effectively the evidence of interdependencies between the two proteins and demonstrate the importance of elucidating the molecular organization of centromeres at the chromatin fiber level.

1.9 Neocentromeres – prevalence & relevance

Neocentromere formation as reported in over 90 cases over the years have mostly formed through chromosomal rearrangements processes. The most common method of natural neocentromere formation happens when a chromosomal rearrangement event occurs to produce an acentric fragment which is then rescued by the formation of a new centromere or "neocentromere" (Kalitsis and Choo, 2012; Marshall et al., 2008a; Warburton, 2004). Human neocentromeres tend to form on gene-free regions with no particular sequence preference (Kalitsis and Choo, 2012). However when genome engineering methods allowed for isolation of experimentally induced neocentromeres in DT40 cells (Shang et al., 2013) these newly formed neocentromeres were able to form on both transcriptionally active and inactive sites. Other cases of centromeres forming on active gene regions have been shown in rice (Nagaki et al., 2004). Initially this was surprising due to the compact heterochromatic nature of the centromere region, however other examples of active genes in heterochromatin regions have been shown in fruit flies at low density (Weiler and Wakimoto, 1995) and are expressed despite being associated with heterochromatin proteins (Greil et al., 2003).

Neocentromeres have been generated in a number of ways in several species and have revealed novel insights into centromere structure and function (Fukagawa and Earnshaw, 2014). Notably in *D. melanogaster,* chromosome breakage through γ-irradiation treatment led to the formation of a neocentromere in the pericentromeric heterochromatin region of the X chromosome (Murphy and Karpen, 1995) suggesting centromere activity could spread in *cis*. Another strategy which utilised a Cre-loxP system to evict the endogenous centromere was carried out in fission yeast. This process allowed the formation of neocentromere-containing chromosomes and showed formation of neocentromeres on adjacent telomeric regions (Ishii et al., 2008). In a more recent case, chromosome-engineering again using a cre-loxP-like system in

chicken DT40 cells allowed the isolation numerous neocentromeres (Shang et al., 2013). Through Chromatin Immunoprecipitation followed by next generation sequencing the isolated neocentromeres in chicken DT40 cells were shown to be in the range of 40 ± 6 kb. These neocentromeres like most others did not have any preferential sequence binding. It was also reported overexpression of CENP-A in that system did not lead to any expansion of the CENP-A domains (Shang et al., 2013). This is somewhat in contrast to another study in *D. melanogaster* when its CENP-A ortholog, CENP-A^{CID}, was over expressed, it led to mis-localisation of CID at non-centromeric regions (Heun et al., 2006). The mis-localisation of CID also triggered the assembly of kinetochore proteins to the sites of mis-localisation (Heun et al., 2006). Similar results were shown in a human cell line study where a subset of kinetochore proteins were assembled to induced mis-localised CENP-A regions (Hooser et al., 2001).

Notably, from a clinical perspective the importance of neocentromere studies is emphasised with their link with certain cancers. Probably the best example of the link between neocentromeres and cancer can be seen in atypical lipomas and well differentiated liposarcomas (ALP-WDLPS) which is a class of lipomatous tumours (Marshall et al., 2008a). A large ring chromosome or a supernumerary ring chromosome typically marks this class of tumours. They are usually devoid of satellite DNA and mainly composed of amplified of specific regions of chromosome 12 that are known to contain oncogenes (Pedeutour et al., 1999). These large ring or supernumerary chromosomes contain neocentromeres at the primary constriction (Forus et al., 2001; Sirvent et al., 2000).

1.10 Evolutionary new centromeres (ENCs)

Chromosome evolution occurs frequently in eukaryotes through specific chromosomal rearrangement events, which include insertions, deletions, translocations and duplications. These rearrangement events can occur in a gross fashion whereby several megabases of a chromosomal fragment is involved in the rearrangement event, or at a minor level, which effects only part of a gene, however, at any level, these events are key aspects of speciation (Coghlan et al., 2005). Large changes in chromosomal configuration have been implicated in centromere repositioning. First hypothesised as "centromere translocations", it was suggested that centromeres repositioned via chromosomal translocations (Dutrillaux, 1979). Later, centromere repositioning was proposed to occur through centromeric activation/inactivation events (Clemente et al., 1990). However, centromere repositioning was first examined through a series FISH experiments that were aimed at mapping marker positions in chromosome 9 in primates. This study showed that the centromere repositioned independently from alterations in marker order (Montefalcone et al., 1999). These repositioned centromeres were termed "Evolutionary new centromeres" and unlike human neocentromeres which typically form through chromosomal rearrangements, ENCs have endured millions of years of maturation (Kalitsis and Choo, 2012).

With a few exceptions, the majority of ENCs discovered to date are associated with satellite/satellite-like DNA. In the macaque (*Macaca mulatta*), nine of the autosomal centromeres are evolutionary new and all nine are associated with alphoid DNA (Rocchi et al., 2012; Ventura et al., 2007). There are six ENCs in humans, all of which are associated with alpha-satellite DNA (Capozzi et al., 2009; Rocchi et al., 2012; Ventura et al., 2003). The ENCs in both these cases are thought to have acquired satellite arrays slowly over millions of years. This hypothesis is supported by the discovery of partial satellite-containing centromeres in rice (Nagaki et al., 2004) and satellite-free ENCs in equids (horses, donkeys & zebras) (Carbone et al., 2006; Piras et al., 2010). In rice, the centromere on chromosome 8 has provided valuable insight into the structure and evolution of centromeres. This particular centromere contains a relatively small abundance of satellite repeats (~40 kb) when compared to the entire centromere forming domain in which CENP-A occupies somewhere in the region of 750 kb. The configuration of the centromere on chromosome 8 is thought to be in the

process of accumulating satellite repeats and therefore at a maturing stage in centromere evolution (Nagaki et al., 2004; Yan et al., 2008). Interestingly, a number of species have been discovered to possess centromeres completely devoid of satellite DNA. Parallel to the finding in rice, through ChIP-seq, chickens have been shown to contain two satellite-free centromeres that are enriched in retrotransposons (chromosomes 27 & Z) and one satellite-free centromere on chromosome 5 (Shang et al., 2010). CENP-A ChIP experiments in the orang-utan revealed that the centromere on chromosome 12 in the species was entirely satellite-free (Locke et al., 2011). The same study showed, through FISH, the entire chromosome 12 was satellite-free and it has been postulated that the satellite DNA may have been deleted through an unequal crossover event (Kalitsis and Choo, 2012). In addition, studies in equids have revealed a remarkably high frequency of ENC occurrences throughout the genus (Carbone et al., 2006; Piras et al., 2010; Wade et al., 2009).

The completion of the horse genome revealed for the first time, one ENC in the horse (Wade et al., 2009). This ENC, was the centromere on chromosome 11 and it lacked hybridisation signal when probed with known equid satellites using FISH. Through ChIP-chip analysis using CENP-A and CENP-C antibodies it was shown that this centromere was functional and completely devoid of satellite DNA (Wade et al., 2009) (Figure 1.2).

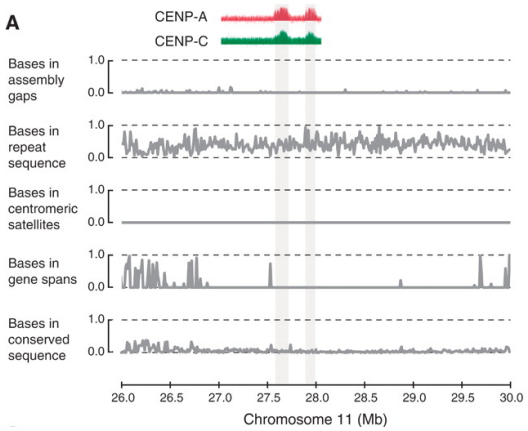


Figure 1.2 Discovery of ENC on horse chromosome 11
Reprinted by permission from The American Association for the Advancement of Science: Science (Genome sequence, comparative analysis, and population genetics of the domestic horse), copyright (2009)

(**A**) Analysis of the centromere of ECA11: 26,000,000 to 30,000,000 bases. ChIP-on-chip analysis with antibodies against CENP-A and CENP-C shows multi-domain (~136 and ~99 kb) binding. There are no satellite tandem repeats, no protein-coding sequences present nearby, and normal levels of noncoding conserved elements.

Equidae (horses, donkeys, zebras) fall under the order of Perissodactyla, an order which consists of two other families, Rhinocerotidae (rhinoceros) and Tapiridae (tapirs) (Wood HE, 1937). Today, the genus equus includes eight different species: the horses (*E. caballus* and *E. przewalskii*); the donkeys (*E. asinus, E. kiang, E. hemionus*) and the zebras (*E. grevyi*, *E. burchelli* and *E. zebra*). The genus Equus shares a common ancestor of ~2-3 MYA (see Figure 1.3) and despite the close evolutionary time span the equids have considerably different karyotype variations (Orlando et al., 2013). Along with chromosome number in the equids (2n=32-66), structural differences can be observed across the species as each contains variable instances of metacentric and submetacentric chromosomes, which presumably originated from chromosomal fusions through evolutionary time (Ryder et al., 1978).

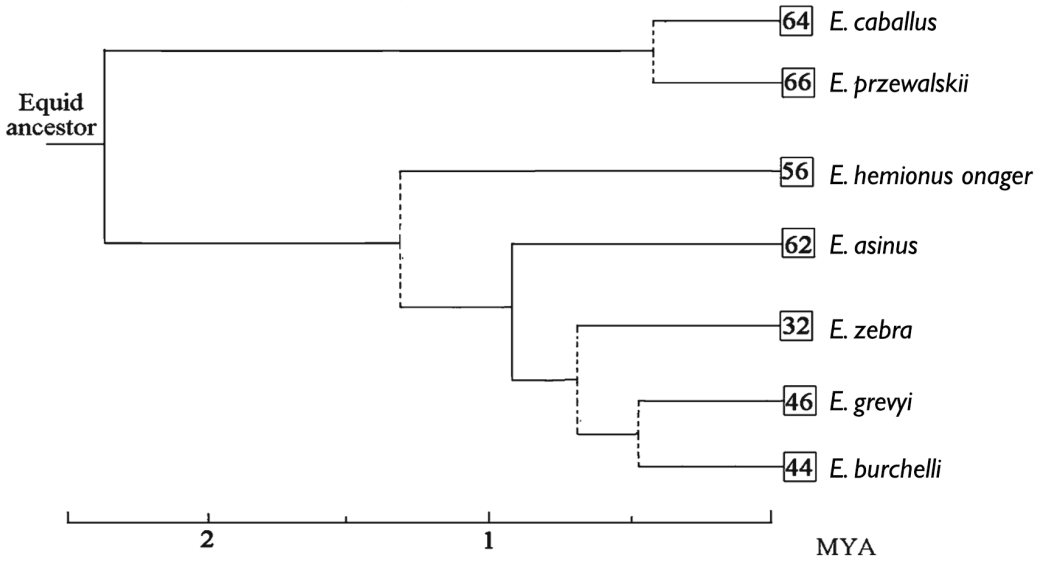


Figure 1.3 Genus Equus karyotype evolution
Adapted from Chromosome Research, "Multidirectional cross-species painting illuminates the history of karyotypic evolution in Perissodactyla", volume 16, Jan 1, 2008, Vladimir A. Trifonov.

Rates of karyotype evolution of equids. Numbers in squares indicate diploid numbers. Scale bar "MYA" (Million years ago) showing evolutionary distance between species.

The most extensively studied species in the genus *Equus* is *E. caballus* or the domestic horse. Cytogenetic and genomics approaches performed on *E. caballus* resulted in the completion of the horse genome and subsequently the discovery of a centromere repositioning event on the chromosome 11 (Piras et al., 2010; Wade et al., 2009).The completion of the horse genome provided an ideal reference for comparative genomics studies in the perissodactyls (Trifonov et al., 2008). Marker order comparison analysis

in other equid species identified a number of centromere repositioning events across the genus. In addition, in an effort to examine the distribution of highly repetitive DNA across the genus *Equus*, a series of FISH experiments using two the known equid satellites along with a total genomic DNA probe were carried out (Piras et al., 2010).This study revealed a remarkable amount of chromosomes across the karyotypes that lacked satellite repeats. One instance was shown in horse, along with eighteen in donkey and several more in two different zebra species (Piras et al., 2010).

The identification of centromere repositioning events across the genus *Equus* demonstrated the surprisingly fast evolution of the karyotypes (Carbone et al., 2006). Collectively, the instances of ENCs identified throughout many different species led to the proposal of a four-state model for centromere formation and evolution (Piras et al., 2010). This model takes into account the observations from many marker order comparisons that showed centromere repositioning events using a phylogenetic based criterion. Shown in Figure 1.4 (Piras et al., 2010) illustrates the steps involved in centromere repositioning. The first step involves the isolation of a centromere from its satellite repeats and moving to a new location on the chromosome to a region devoid of satellites (A-B). This movement would cause the previous centromere-containing satellites to lose centromeric function (C) and finally the repositioned centromere would become fully “mature” after complete accumulation of satellite DNA again (D).

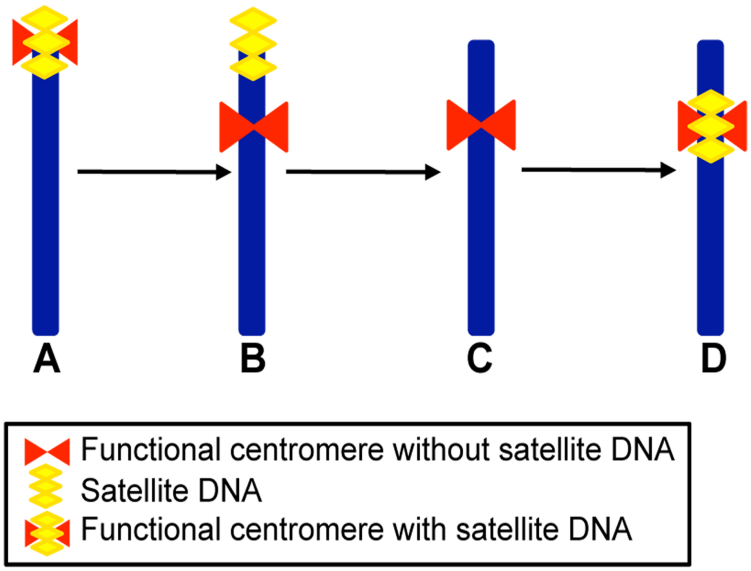


Figure 1.4 Four-state model for evolutionary formation of neocentromeres
Reprinted under the Creative Commons Attribution License: Plos genetics (Uncoupling of Satellite DNA and Centromeric Function in the Genus Equus)(2010)

(A) Ancestral chromosome with terminal centromere (red) containing satellite DNA (yellow) (B) Submetacentric chromosome formed through centromere repositioning in which satellite DNA sequences are maintained (yellow) at the terminal position, while the neocentromere (red) is devoid satellites. (C) Submetacentric chromosome from (B) in which the terminal satellite sequences have been lost. (D) Submetacentric chromosome at full "maturation" defined by presence of satellite DNA (yellow) at the neocentromere.

Research objectives

In order to advance our understanding of key mechanisms involved in centromere regulation its vital we obtain an accurate representation of the molecular architecture of centromeres, at a single base resolution. Equids provide an opportunity to do this through the presence of evolutionary new centromeres.

The specific objectives of this thesis were to:

(1) Use ChIP-seq to identify and characterise satellite-free centromere domains in *E. asinus*. The aim here is to prove the existence of centromere regions devoid of satellite DNA through high throughput DNA sequencing.

(2) Use a native ChIP-seq approach to identify CENP-A nucleosome positions at satellite-free centromere domains. The aim here is to test whether or not preferential CENP-A nucleosome positions are present at the satellite-free centromere domains in *E. asinus*.

(3) Use ChIP-seq to investigate CENP-C distribution at satellite-free centromeres. The aim of this objective is to elucidate the chromatin profile of the CCAN component CENP-C at satellite-free centromeres.

Chapter 2 - Materials & Methods

2.1 Materials – Wet lab

2.1.1 Reagents and Consumables

2.1.1.1 Chemical Reagents

All chemicals used in this study (unless otherwise stated) were purchased from Sigma Aldrich (Arklow, Co. Wicklow, Ireland), Fisher Scientific (Ballycoolin, Dublin, Ireland), Bio-Sciences Ltd. (Dun Laoghaire, Co. Dublin, Ireland), Melford Laboratories (Ipswich, United Kingdom) or Amersham Biosciences (Piscataway, NJ). All solutions were prepared in Milli-Q purified water (Millipore, Tullagreen, Cork, Ireland) or molecular biology grade water (Sigma Aldrich (Arklow, Co. Wicklow, Ireland). All plastic-ware and consumables were purchased from Sarstedt Ltd (Co. Wexford, Ireland) or Fisher Scientific (Ballycoolin, Dublin, Ireland).

2.1.1.2 Molecular biology reagents and equipment

Molecular biology reagents used in DNA digestions, ligations and modifications were purchased from New England Biolabs (NEB, ISIS Ltd., Bray, Co. Wicklow, Ireland), Fermentas (Fisher Scientific, Ballycoolin, Dublin, Ireland) and Roche Products Ireland Ltd. (Naas Rd., Dublin, Ireland). DNA ladders were purchased from New England Biolabs (NEB, ISIS Ltd., Bray, Co. Wicklow, Ireland). Protein size ladders used in SDS-PAGE electrophoresis were purchased from Bio-Sciences Ltd. (Dun Laoghaire, Co. Dublin, Ireland). Protein A/G Sepharose 4 fast flow beads were sourced from VWR International Ltd. (Ballycoolin, Blanchardstown, Dublin 15, Ireland). Agarose gels were prepared using UltraPure Agarose (Invitrogen, Carlsbad, California, USA). Gels were run in Owl EasyCast B1 Mini Gel Electrophoresis system. After electrophoresis gels were visualised in trans-illuminator cabinet (AlphaInnotech 5400 or Syngene G-BOX) and images were captured using CCD camera. Protease inhibitor cocktail (PIC) was purchased from Roche.

2.1.1.3 SDS-PAGE and western blotting Reagents

Unless otherwise stated the NOVEX Sure Lock mini gel system (Life Technologies) with NUPAGE precast gels were used for SDS-PAGE gel electrophoresis. Protein

transfers to PVDF membrane purchased from GE Healthcare were carried out using the XCell II Blot Module (Life Technologies).

2.1.1.4 Antibodies

All antibodies used in this study are listed in Table 1 (primary antibodies) and table 2 (secondary antibodies) below. The table includes antibody source, host species, and working concentration for immunofluorescence (IF) and western blotting (WB) detection.

Table 2.1 Antibodies

Antigen	WB	IF	ChIP	Host	Source
CENP-A (Peptide)	-	-	1µL/10^6 cells	Rabbit	E. Giulotto
CENP-A (CREST)	-	-		Rabbit	E. Giulotto
CENP-A_horse	1/5000	1/200	1µL/10^6 cells	Sheep	Teri Masterson
CENP-C	1/500	1/100	1µL/10^6 cells	Guinea Pig	MBL

2.1.1.5Secondary antibodies

All secondary antibodies used in this study are listed in this table below

Table 2.2 Secondary antibodies

Reactivity	Conjugate	Host	Dilution WB	Dilution IF	Source
Anti Rabbit IgG	HRP	Goat	1:20,000	-	Jackson
Anti Sheep IgG	HRP	Goat	1:20,000	-	Jackson
Anti Guinea Pig IgG	Alexa488	Donkey	-	1:100	Courtesy of Dr. Elaine Dunleavy
Protein A	HRP	-	1:20,000	-	Merck Millipore
Anti Sheep IgG	TRITC	Rabbit	-	1:100	Jackson
Anti Guinea Pig IgG	HRP	Donkey	1:20,000		Courtesy of Dr. Elaine Dunleavy

2.1.1.6 Quantitative Real Time PCR (qPCR)

Quantitative real-time PCR was performed on the DNA Engine OPTICON2 – Continuous fluorescence detector (Courtesy of Prof. Brian McStay). The Fast SYBER Green Master Mix from Applied Biosystems was used to prepare reactions for qPCR. Three primer pairs purchased form Eurofins Genomics (Ebersberg, Germany) were used to in this study to detect enrichment inside or outside centromere domains. The primer probes are shown in Table 2.3 below.

Table 2.3 qPCR primer pairs

Probe	Chromosome (EquCab2.0)	Forward	Reverse	Product Length
Eca11	Chr11	CAGCAAGGCATTTCCAGTGA	CATGCAAGACAAGGAGGAACG	130 bp
PRKC	Chr19	TGGAGCAAAAGCAGGTGGTA	ATCGTCATCTGGAGTGAGCTG	116 bp
Eas30	Chr30	CACTACCCTGGCACTGCGA	TGGATGTCACGGTAGGCAATG	103 bp

2.1.2 Tissue Culture Materials

2.1.2.1 Cell lines

Skin fibroblasts in both horse and Donkey were used in this study and were kindly provided by Prof. Elena Giulotto, Pavia, Italy. The horse cells were primary cell lines and the donkey and hTERC and hTERT immortalised cell lines (Vidale et al., 2012).

2.2 - Methods - Wet Lab

2.2.1 Cell culture methods

2.2.1.1 Growth Conditions

Equid cell lines were cultured in DMEM F-12 media (Sigma Aldrich) supplemented with 2% non essential amino acids, 1% sodium pyruvate, 1% penicillin streptomycin, 1% L-glutamine, 10% foetal bovine serum, 10% horse serum, 400uM G418 (immortalised cell line only) and buffered to correct pH with sodium bicarbonate. Cell culture was carried out in a sterile tissue culture hood. Cells were typically seeded and cultured on 10cm or 15cm culture dishes and incubate at 37˚C at 5% CO_2 range. When cells reached 70-80% confluence, there were washed in sterile PBS (Bio-Sciences, Dún Laoghaire, Dublin), trypsinised in 1X Trypsin-EDTA for 2-3 minutes, resuspended in fresh complete media and split into two dishes.

2.2.1.2 Cryopreservation

For freezing down cells, detachment from a 70% confluent 10cm dish was carried out by incubation with 1X trypsin-0.25% EDTA for 3-5 minutes at 37˚C. Cells were re-suspended in complete media to inhibit the trypsin reaction and pipetted into 15mL tube. Cells were pelleted by centrifugation at 200*g* for 5 minutes and gently re-suspended in freezing media (Table 2.1). Cells were then transferred to a 1.5mL cryovial (Sarstedt). Cells were then transferred to a Nalgene freezing container which

cools at a rate of 1˚C/minute in -80˚C by use of isopropanol. After 2-3 days cells were then transferred to liquid nitrogen dewer for long term storage.

2.2.1.3Cell revival

Cell revival was carried out by thawing cryovial containing cells at 37˚C. Approximately 1mL of pre-warmed complete media was added to cryovial in a dropwise fashion. This cell suspension was then added to a 10cm tissue culture dish containing pre-warmed media subject to incubation with 5% CO_2. The cells were incubated overnight and allowed to adere to the dish and media was changed to recover DMSO.

2.2.2 Agarose Gel electrophoresis

2.2.2.1 Gel preparation & Running

Depending on resolution required, 0.7% to 1.5% agarose gels were prepared using UltraPure Agarose (Invitrogen, Carlsbad, California, USA) in 1X TAE buffer containing ethidium bromide at a concentration of 0.5ug/ml. 1-20µL of samples were diluted appropriately in 6XLoading dye (NEB) and run with 2Log Ladder (NEB). Gels were run in 1X TAE buffer for 40-80mins at 80-120V in Owl EasyCast B1 Mini Gel Electrophoresis system. After electrophoresis gels were visualised in trans-illuminator cabinet (AlphaInnotech 5400 or Syngene G-BOX) and images were captured using CCD camera.

2.2.3 DNA Purification

2.2.3.1 Phenol Chloroform Extraction and Ethanol Precipitation

Samples for DNA purification are normally made up to 300µl then 300µl of phenol chloroform isoamyl alcohol (25:24:1) is added to sample. The sample is then vortexed vigorously for approximately 5 seconds and centrifuged at top speed (~17,000g) for 2 minutes and the aqueous phase is transferred into a new tube. This step is repeated and, to precipitate the DNA, 1/10 volume of 3M NaCl is added along with three volumes of 100% ice-cold ethanol. After, sample is then stored at -80˚C for at least 20 minutes. The sample is then centrifuged at 17,000g for 15 minutes at 4˚C. DNA pellet is air-dried for 15-20 minutes and re-suspended in a suitable volume of molecular biology grade water or 1XTE.

2.2.3.2 QIAgen PCR purification kit

The QIAquick PCR purification kit (cat. 28106) was the primary method of DNA purification for ChIP DNA. Before starting the purification procedure with this kit a 1:250 volume of pH indicator to buffer PB was made up. Five volumes of buffer PB was added to one volume of PCR reaction and mix. If the colour of the mixture was orange or violet then 10µL of 3M sodium acetate, pH5 was added and mixed. The sample was then added to a QIAquick column and centrifuged for 1 minute into a 2mL collection tube. The sample was washed by adding 750µl of Buffer PE to the QIAquick column and centrifuged for 1 minute, the flow through was discarded and the tube was centrifuged again to remove residual wash buffer. The QIAquick column was added to a 1.5 mL microcentrifuge tube and 20-50µl of molecular biology grade water or 1XTE pH8 was added and allowed to incubate for at least 1 minute before recovering the DNA by centrifugation.

2.2.3.3 Quantitative Real-Time PCR

Quantitative real-time PCR was carried out on the DNA Engine OPTICON2 – Continuous fluorescence detector (Prof. Brian McStay). The qPCR reaction was made up as described in Table 2.4

Table 2.4 qPCR reaction components

Component	Volume (20µL/reaction)	Concentration in reaction
DNA (Input or ChIP)	1µL	-
Master Mix 2X	10 µL	1X
Forward/Reverse primers 2	2 µL (each)	0.2 µM
H_2O PCR grade	5µL	-

The reaction conditions were programmed into the OPTICON software as described in the Fast SYBER Green Master Mix instruction manual and shown in Table 2.5

Table 2.5 qPCR reaction conditions

Step	Temperature (°C)	Duration	Cycles
Polymerase activation	95	20	1
Denature	95	3	40
Anneal/Extension	60	30	

A description of the primer probes used in qPCR reactions are outlined below in Table 2.6

Table 2.6 qPCR reaction probes

Primer Probes (fw & Rev))	Description	Control
EAS30	CENP-A (centromeric) domain within a centromere in donkey genome	+
PRKC	(non-centromeric) which corresponds to a region on the PRKC housekeeping gene	-
ECA11	(Non-centromeric) region of the donkey genome	-

2.2.3.4 qPCR Calculation – Percentage recovery

The "percent input" method was used to calculate percentage recovery of ChIP DNA. The input sample represents the amount of chromatin used in the ChIP. Typically, 1% of starting chromatin is used as input. For example, if the starting input fraction is 1%, then a dilution factor (DF) of 100 or 6.644 cycles (i.e., log2 of 100) is subtracted from the Ct value of diluted input. Percentage ChIP realative to input was calculated as follows:

Adjusted Input = Ct Input – 6.644

Percentage recovery = $100*2^{(\text{Adjusted Input-Ct(ChIP)}}$

2.2.4 Protein Methods

2.2.4.1 Cell fractionation & extract preparation

Cell extract preparation was carried out during the native chromatin preparation protocol described in section 2.2.4.9. During the successive stages of the protocol the cell fractions named whole cell extract (WCE), cytoplasmic extract (CYTO), nuclear extract (NUC) and chromatin extract (CHR) were collected.

2.2.4.2 Protein sample preparation for SDS-PAGE

All cell or protein fractions were supplemented with NuPAGE LDS sample buffer to a final concentration of 1X and DTT to a final concentration of 100mM. Typically samples were made up to 10^4 cells/µL or 10^4 cell equivalents/µL (CE/µL). Samples were boiled for 5 minutes, quickly centrifuged and carefully loaded on gel for SDS-PAGE described in section 2.2.4.3.

2.2.4.3 SDS-PAGE and western blotting

Unless otherwise stated the NOVEX Sure Lock mini gel system was used fro SDS-PAGE gel electrophoresis and transfer system. The mini gels are precast gradient gels 4%-

12% and were run a 200V for 1 hour in either NuPAGE MES or MOPs running buffer. Samples are typically loaded according to cell number ($0.5x10^5$-10^5 cells per lane). The proteins were then transferred to PVDF membrane using the XCell II Blot Module. The gel and PVDF membrane were placed between specially arranged set of sponges and filter papers. Two sponges were placed on the bottom followed by a piece of filter paper. The gel was carefully placed on top of the filter paper with the membrane directly on top of the gel. Air bubbles were removed by running the gel knife along the filter paper that lies on top of the membrane. Another two pre-soaked sponges were then put on top of the build and the transfer box was closed before setting it into the transfer box, the transfer box was filled with NuPAGE transfer buffer until the buffer covered the sponges then the outside chamber was filled with distilled water. After the top was put on the transfer apparatus the transfer was run at 30V for 1 hour. Once the transfer was finished the membrane was blocked using 5% milk in TBS-Tween20 for 1 hour at room temperature on a rocker. The membrane was quickly rinsed in TBS-T then the primary antibody was made to the required dilution in 5% milk-TBST and allowed to incubate overnight at 4˚C (or 1 hour RT). After primary antibody incubation the membrane was washed 3 times in TBST for 5-10 minutes each time. A HRP conjugated secondary antibody was then made up in 3% milk-TBST and incubated for 1 hour at room temperature on a rocker. The membrane was then washed four times for 5-10 minutes each time using TBST. Detection was carried out using ECL were the membrane was incubated for 5 minutes with ECL solution and developed using the G-Box or x-ray film development.

2.2.4.4 Coomassie Blue Staining

SDS-PAGE gels were visualised after incubation with Coomassie blue solution (0.1% Coomassie R250, 10% acetic acid, 40% methanol) for two hours on a shaker at room temperature. Gels were then incubated with de-stain solution (20% methanol, 10% acetic acid) and a kim wipe until the excess dye was removed. Stained gels were then scanned using HP Scanjet Pro.

2.2.4.5 Crosslinked cell preparation

Cells were harvested by incubation with 1X trypsin – EDTA for 3-5 minutes at 37˚C. Trypsin reaction was inhibited by addition of complete media followed by elutriation of the cells to remove the majority from the culture dish. The cells are collected then

centrifuged at 200 *g* for 5 minutes. Cells are washed and re-suspended in ice-cold PBS and counted using a haemocytometer. Formaldehyde is then added to a final concentration of 1% and incubated at 26˚C for 9 minutes. The formaldehyde is quenched by the addition of glycine to a final concentration of 125mM at 26˚C for 15 minutes. Cells are then washed in ice cold PBS and pelleted in $1x10^7$ cell fractions, followed by freezing at -80˚C until further use.

2.2.4.6 Optimisation of the sonication procedure

Typically for a crosslinked-ChIP reaction whether the immunoprecipitated DNAs will be used in qPCR or ChIP-seq the DNA fragment sizes are recommended to be in the 400-600 bp range. This procedure was optimised by performing multiple successive rounds of sonication on the Diagenode Bioruptor UCD200. Donkey skin fibroblasts were sonicated at 10,20 & 30 cycles in 30secON/30secOFF cycle intervals. 30 cycles produced desired extent of sonication.

2.2.4.7 Protein A/G preparation

Protein A or G beads were prepared by making a blocking mix consisting of sonicated E.*coli* DNA (final concentration 75µg/mL), BSA 1mg/mL and 1 X PBS. The protein A or G beads were incubated for at least 2 hours at 4˚C on a rotator. Once the incubation was complete the protein A or G beads were washed three times in ice cold PBS with centrifugation at 3500rpm at 4˚C for two minutes between each wash. Beads were either used directly or stored in ice-cold PBS at 4˚C until ready for use.

2.2.4.8 X-ChIP (Formaldehyde Stabilised)

Approximately $1x10^8$ donkey skin fibroblasts were harvested and cross-linked as described in section 2.2.4.5. Cell pellets were thawed and washed two times with 1XPBS+Protease inhibitor cocktail (PIC) and re-suspended in "ChIP lysis buffer" (0.25% SDS, 50mM Tris-HCL pH8, 10mM EDTA pH8 and PIC) into $2x10^7$ aliquots of 650ul. Cells were sonicated using a Branson sonifier 450 probe sonicator – 24 cycles of 10seconds bursts at ~40% amplitude. 15ul of sample was taken for DNA analysis to check the efficiency of sonication. Samples were pooled together and made up to 10mls of "ChIP dilution buffer" (0.5% Nonidet P-40, 10mM Tris-HCL pH7.5, 2.5mM $MgCl_2$, 150mM NaCl and PIC). Chromatin was precleared in 200ul of pre-blocked protein G beads (beads blocked using sonicated *E.coli* DNA) for 2hrs at 4˚C. Samples

were recovered and 1% of sample volume was saved as input (100ul). Samples were split into 8x1.5ml tubes (1250ul each) and the required amount of antibody was added to each pre-cleared chromatin sample and incubated overnight at 4˚C. The following day 40ul of pre-blocked protein G beads were added to each ChIP sample and incubated for 4hrs at 4˚C to capture antibody. The supernatant was aspirated and beads were washed 5X in "ChIP wash buffer" (0.1% SDS, 1% Triton-X100, 2mM EDTA pH8, 150mM NaCl and 20mM Tris-HCL pH8) and the 1X "ChIP final wash buffer" (0.1% SDS, 1% Triton-X100, 2mM EDTA pH8, 500mM NaCl and 20mM Tris-HCL pH8). Immunocomplexes and input chromatin were eluted by adding 240ul of "ChIP elution buffer" (1% SDS, 100mM $NaHCO_3$, 40ug/ml RNase A) and incubating at 37˚C for 1hr. Proteinase K was then added to a concentration of 150ug/ml and incubated at 55˚C for 2hrs followed by de-crosslinking step and 65˚C overnight. Immunoprecipitated DNAs were cartridge purified using QIAquick® PCR Purification Kit. ChIP and Input samples were eluted in 100ul of molecular biology grade water.

For the CENP-C-Immort2 ChIP, cell pellets were re-suspended in "Buffer A" (100mM Tris-HCL pH8, 10mM DTT) and incubated for 10 mins on ice followed by a 30˚C incubation with shaking for 15min. Cells were centrifuged at 5000 rpm for 3mins at 4˚C and re-suspended in 1mL of "Buffer B" (10mM HEPES pH7.5, 10mM EDTA, 0.5mM EGTA, 0.25% Triton X-100). Cells were centrifuged at 5000 rpm for 3mins at 4˚C and re-suspended in 1mL of "Buffer C" (10mM HEPES pH7.5, 10mM EDTA, 0.5mM EGTA, 200mM NaCL). Cells were centrifuged at 5000 rpm for 3mins at 4˚C and re-suspended in 600μL of "Buffer D" (50mM Tris-HCL pH8.0, 10mM EDTA, 1% SDS, PIC) and divided up into 200μl aliquots. Sonication was performed on the Diagenode Bioruptor UCD200. Donkey skin fibroblasts were sonicated at 30 cycles in 30secON/30secOFF cycle intervals. Samples were then diluted in "IP Buffer" (15mM Tris-HCL pH8.0, 1.2mM EDTA, 180mM NaCL, 1.2% Triton X-100, PIC). The ChIP experiment was carried out as outlined above from here on.

2.2.4.9 Preparation of Native Chromatin by Micrococcal Nuclease

Donkey skin fibroblasts were harvested using 1X trypsin-EDTA and washed in complete media (DMEM-F12-HAM). Cells were then washed twice and re-suspended in non-complete media (DMEM-F12-HAM) followed by counting using a

hemocytometer (VWR, Ballycoolin, Dublin). Cells were then washed twice with PBS (Lonza) followed by two washes with "Isolation Buffer" (3.75 mM Tris-HCL pH8.0, 20 mM KCL, 0.5 mM EDTA, 0.5 mM DTT, 0.125 mM Spermidine, 0.05 mM Spermine and 0.1 mM PMSF). The remaining cell pellet was re-suspended in five volumes of "Isolation Buffer" supplemented with 0.1% Digitonin. Cells were then allowed to swell on ice for 10mins. Swollen cells were then subject to ~20-25 compressions using a tight pestle dounce homogenizer 15ml (Wheaton). During this step 1-2 ul of homogenized sample was taken every five compressions and viewed under light microscope to observe the approximate percentage of free nuclei. Once satisfied that enough nuclei were isolated the sample was centrifuged at 650 *g* for 10mins at 4˚C. Samples were taken prior to centrifugation for whole cell extract (WCE). After centrifugation, the supernatant was sampled and used as cytoplasmic extract (CYTO) and the pellet as nuclear extract (NUC). Nuclei were washed in "Isolation Buffer-Digitonin" and once with "Nuclei Wash Buffer" (20 mM HEPES pH8.0, 20 mM KCL, 0.5 mM EDTA, 0.5mM DTT and 0.1mM PMSF). Nuclei were re-suspended in "Micrococcal Digestion Buffer" (15 mM HEPES pH8.0, 15 mM NaCl, 60 mM KCL, 0.5 mM Spermine, 0.15 mM Spermidine, 1 mM DTT, 1 mM PMSF and PIC) to ~$1x10^5$ cell equivalents/µl (CE/µl) and $CaCl_2$ was supplemented to 1 mM. 24 U/ml of S7 Nuclease (Roche) was added to each sample and incubated at 37˚C for 20mins. The digested chromatin was supplemented with 200 mM EGTA to 10 mM and mixed vigorously by pipetting. 3M NaCl was then added to a final concentration of 300 mM and mixed vigorously by pipetting. The supplemented samples were then centrifuged at 17,000 *g* for 15mins at 4˚C and the supernatant was kept as the soluble chromatin fraction (CHR). From here ~$1x10^6$ CE was taken to observe extent of micrococcal digestion. Chromatin samples were subject to phenol chloroform extraction and ethanol precipitation described in section 2.2.3.1 and prepared for gel electrophoresis described in section 2.2.2.1.

2.2.4.10 Titration of S7 nuclease digestion

In order to have a sufficient degree of S7 nuclease digestion a titration experiment was carried out. We first decided to have a uniform cell concentration for each digestion experiment in the future – $1x10^5$ CE/µL. We thought it would be more efficient and reproducible to keep cell concentration, temperature and incubation time constant and have enzyme concentration as the only variable. We were then able to empirically

determine the most suitable digestion conditions for our ChIP reactions. An example of one S7 nuclease titration experiment is shown in Figure 4.3.

2.2.4.11 Sucrose Gradient Fractionation of Mono- and Oligonucleosomes

Sucrose gradients were made by hand. Two solutions 5% and 28% sucrose were made up in *Sucrose Gradient buffer* (1X Tris-EDTA, 0.3 M NaCl, 0.2 mM PMSF, 1 mM DTT). Ten sucrose solutions were made (5%-28%) and carefully layered on top of one another (28% on bottom). When the gradients were made they were set in the cold room for 2 hours (until chromatin fractions were made). When the native chromatin was prepared using micrococcal nuclease (2.1.9) the sample (1 mL) was carefully layered on to the gradient in a dropwise fashion. Once the samples were weighed out to two decimal places they were loaded onto the Beckman Coulter SW32Ti ultracentrifuge rotor. The ultracentrifuge was set to 28000rpm for 18hours at 6˚C. No brake was applied to the centrifuge when the spin was over. When the rotor was stopped the sample was recovered from the rotor bucket and the gradient was collected from the bottom in 40ul aliquots. Aliquots were taken for DNA and protein analysis with the remainder used in ChIP. Protein samples were prepared as described in section 2.2.4.2 and DNA samples were prepared as described in section 2.2.3.1.

2.2.4.12 Immunoprecipitation (Native)

Chromatin samples or nucleosome fractions previously prepared natively (section 2.2.4.9 & section 2.2.4.11) were pre-cleared using pre-blocked protein A/G beads for 1-2hrs at 4˚C. The beads were then discarded and the chromatin was supplemented with affinity purified CENP-A antibody (1ul/10^6 CE) and allowed to incubate for 4 hours until more protein A/G beads were added and allowed to incubate for 16 hours overnight. The immunocomplexes were then washed five times with "*ChIP wash buffer*" (0.1% SDS, 1% Triton-X, 150 mM NaCl, 2 mM EDTA pH8, 20 mM Tris-HCL) then once with "*ChIP final wash buffer*" (0.1% SDS, 1% Triton-X, 500 mM NaCl, 2 mM EDTA pH8, 20 mM Tris-HCL). The immunocomplexes were then recovered by adding a suitable volume of "*ChIP elution buffer*" (100 mM $NaHCO_3$, 1% SDS) and incubating at room temperature for 15 minutes followed by a 15 minute incubation at 37˚C. Samples were then processed for DNA purification using QIAgen PCR purification kit

(2.2.3.2) or protein analysis by adding a suitable volume of Laemmli sample buffer (section 2.2.4.2).

2.2.4.13 Immunofluorescence

Donkey cells were grown on glass coverslips in four well dishes. When the cells reached ~70% confluence, the media was removed and cells were gently washed in PBS. Cells were fixed with 100% ice-cold methanol for 20mins at -20˚C. Fixed cells were washed twice for 2 minutes in PBS followed by 2x3 minute washes in PBS-TX (PBS, 0.1% Triton X-100). Cells were then blocked in 1% BSA PBS-TX for 15 minutes. The primary antibody was diluted in 1% BSA-PBS-TX and incubated on the coverslips for 1 hour at 37C. Cells were washed 2x10 minutes in PBS-TX and incubated with a fluorescently conjugated secondary antibody in BSA-PBS-TX at 37C for 1 hour. Cells were then washed once in PBS-TX, once in PBS and once in water before being air-dried and mounted on slides with Slowfade and DAPI. Cells were then imaged on a Deltavision Core system (Applied Precision).

2.3 Materials – Dry Lab

2.3.1 Hardware

Computer used for all bioinformatics analysis was an Apple iMac (late 2012) running iOS X Mountain Lion. 32 GB RAM with 3.4 Ghz Intel core I7 duo processor.

2.3.2 Software

Software packages include:

Table 2.7 Software

Software	Version	Application
Bowtie2	2.1.0	Short read aligner
SAMtools	0.1.19	Alignment file manipulator
BEDtools	2.22.1	Alignment file manipulator
Picard Tools	1.102	Quality Control
FastQC	0.10.1	Quality Control
Trimmomatic	0.33	Quality Control
Deeptools	2.0.1	ChIP-seq normalisation
MACS	2.0	Peak Calling
R	3.1.3	Plotting
IGV	2.3.36	Genome Browser
NCBI Genome Workbench	2.7.6	Genome Browser
NucleR	3.2	Nucleosome Positioning

2.4 Methods – Dry Lab

2.4.1.1File formats

SAM file (.sam) – Sequence alignment and mapping file. This is an output file from a short read aligner. It contains all the information from the alignment.

BAM file (.bam) – Binary alignment and mapping file. This is the binary version (not human readable) of the SAM file that contains all the same information except its in binary format.

PRN file (.prn) – This is a space delimited file. It only allows 240 characters per line and allows for data to be separated by tabs. These file types are useful with the UCSC genome browser.

BED file (.bed) – This is a browser extensible data file. The BED file is a tab-delimited file that normally defines a feature track e.g. for the UCSC genome broswer. It can have three to 12 columns the first three columns are normally the chromosome number and the start and end coordinates on that chromosome.

BEDGRAPH file (.bedgraph/.bg) – This is a file normally used in the UCSC genome browser or IGV. It's similar to a .bed file but it starts with a track line definition which defines the files visual properties. It is used for showing continuous data in a interactive viewer.

WIGGLE & BIGWIG file (.wig & .bigwig/.bw) – This format allows a user to display large amounts of continuous dense data (for example, data at single base resolution) and contained in an binary indexed format. The .WIG or wiggle file is a smaller text file version of a BIGWIG.

2.4.1.2 Quality control of ChIP-seq data using FastQC and Trimmomatic

Two different quality control suites were used on the ChIP-seq data. FastQC provides data on the number of reads and outputs several quality statistics, in a user-friendly interface. Trimmomatic is a command line tool that carries out the same quality statistics but also has the option to trim sequence tags if needed. Trimmomatic was chosen because of its usefulness in working with paired-end data.

2.4.1.3 Short read alignment using Bowtie2

In order to align short reads to a reference genome the first step is to build an indexed reference genome. The indexed genome is a binary form of genome that allows for a computationally efficient "search and cross-reference" during the alignment. The reference genome is converted to this index binary form using a Burrows-Wheeler transform. In order to run this protocol the following function (command 2.1) is executed:

```
bowtie2-build –f chr1.fa,chr2.fa…..chrN.fa equcab2.0_bowtie2_index
```
Command 2.1 Build index genome

In this instance bowtie2-build is the burrows wheeler transform function, -f equcab2.0_bowtie2_index is the output file name – the name of the newly indexed genome. Once the indexed genome is built it is then ok to proceed with aligning the FASTQ reads to the genome. The command used to align all of the datasets described in this thesis is as follows:

```
bowtie2-align –x bowtie2_indexed_genome –p8 -1 fastq.reads.r1.fastq -2
fastq.reads.r2.fastq –S aligned.reads.sam
```
Command 2.2 Align Reads - Paired-end mode

As shown in the example above bowtie2-align is the default alignment parameter (see section 2.3.2.1), -x calls the indexed genome to be used, -p8 calls the computer use all possible processers for a faster alignment, -1 & -2 denotes the first and second reads in the pair (paired-end mode), -S outputs the file as a SAM file and aligned.reads.sam is the name of the output file.

2.4.1.4 File conversions using SAMtools

Files outputted from bowtie2 are large non-binary formats that can be opened in text editors or viewed in the terminal window. To work with the data in this format is very computational intensive. In order to be less computational intensive and more efficient the SAMtools suite is used to convert these big data files to binary format. There are several functions regularly used for file conversions. The main functions used for these are outlined below:

```
samtools view –bSo aligned.reads.bam aligned.reads.sam
```
Command 2.3 Convert SAM to BAM

In the above command `view` is the conversion function `–bSo` says output (`o`) binary (`b`) from this SAM (`S`) format

```
samtools sort aligned.reads.bam aligned.reads.sorted
```
Command 2.4 Sort BAM file

The above command uses the `sort` function to sort the reads according to chromosome number rather that genomic location. This output is again less computationally intensive to work with

```
samtools index aligned.reads.sorted.bam
```
Command 2.5 Index BAM file

Like the `index` command in bowtie2 (Section 2.3.2) the index function above creates and index file of the BAM file that can be used to search for specific locations or intervals on the aligned genome.

2.4.1.5 Normalisation using deeptools

Deeptools is a suite of tools used for various methods of normalisation. The software suite contains a number of normalisation functions. The main function used in this work was "bamCompare". The "bamCompare" function tool compares two BAM files based on the number of mapped reads. Typically the genome is partitioned user defined bin sizes, then the total reads per bin for each BAM file is calculated and then a user defined summary value is outputted. In our case we chose to output the scores of a subtractive normalisation based on reads per kilobase million (RPKM) count which outputs a .bedgraph or .bigwig file.

```
bamCompare -b1 chip.sorted.bam -b2 input.sorted.bam --outFileName
normalised.bedgraph --outFileFormat bedgraph --scaleFactorsMethod readCount
--ratio subtract --normalizeUsingRPKM --binSize 10 --numberOfProcessors max
```
Command 2.6 DeepTools normalisation

2.4.1.6 Read count extraction using SAMtools

One method we employed to plot our CENP-A binding regions was by extracting the read counts across the genomic locations and imported these data into R. In R we plotted read count as a function of genomic location. In order to do this the `mpileup` function in SAMtools was performed. As described earlier in the section 2.4.1.1 a SAM or BAM file is a tab-separated file with numerous columns. The columns contain all the information regarding the alignment. We used the `mpileup` command to extract the reads that map back to our genomic locations of interest, which can be seen in. The command we used was as follows:

```
samtools mpileup –r chrN:1,000-100,000 aligned.reads.sorted.bam >
chrN_1000to100000.txt
```

Command 2.7 Read count extraction using mpileup

This command extracted the reads and pileup value that mapped back to the genomic interval between 1000 and 100000 on chromosome *N*. Here the output file contains the start and end locations of the read, the read itself and the value given for the amount of reads mapped back to all these positions.

2.4.1.7Read count filtering

When plotting the read count graphs in R a lot of information is extracted from the mpileup command. In order to just extract the information necessary for the plot the following command was used to do so:

```
awk ‘{print$2, $4}’ chrN_1000to100000.txt > chrN_1000to100000.prn
```

Command 2.8 Print specific columns of a file

This line of code prints the second and fourth column in the chr*N*_1000to100000.txt file and writes a new file called chr*N*_1000to100000.prn containing only these two columns – the genomic position and the read count value.

2.4.1.8 Data visualisation

Throughout this work data was visualised using a number of different platforms; UCSC genome browser, NCBI genome workbench, IGV and R. The original analysis carried out aligning the donkey ChIP-seq data to the horse reference genome was converted to an observable format and visualised on the UCSC genome browser. Following the construction of the hybrid genome "EquDonk2.0" the aligned data were visualised on IGV or plotted in R.

2.4.1.9 Plotting read count graphs using R

Alignment data in two formats - raw read count data and bedgraph data were plotted in R. The R package *Sushi* was used to plot bedgraph files. Example scripts for both methods are shown below in Command 2.9.

```
Raw alignment file

png(filename="file.png")
par(mar=c(0.1, 5, 0.1, 5), oma=c(4,0,4,0))
plot(file, type="h", lty=1, lwd=0.1, col="blue", main="" , ylab="", xlab="",
col.axis="black",axes=F, col.lab="black", col.main="black", fg="black", bg="black",
xlim=c(0, 185838109), ylim=c(0, 300))
box(which = "plot", lty = "solid", col="gray40", bty="L")
mtext("ECA11",side=3,col="black",line=5.5, cex=1)
axis(2, col = "gray40", col.axis = "gray40", col.ticks = "gray40", cex.axis=0.6)
dev.off()

Bedgraph file
library('Sushi')

png(filename="file.png", height =8, width= 8, units="in", res=600, bg="white")
par(mar=c(1.6,1.5,1.6,1.5), oma=c(0.9,0.5,1.4,0.7))

plotBedgraph(bedgraph,  chrom, chromstart, chromend, color= "gray", transparency=.9,
lwd = 0.01, linecolor = "black", range = c(0,6500))
mtext("Heading",side=3,col="black",line=0.5, cex=0.2)
mtext("Y-axis label",side=2,col="gray40",line=0.5, cex=0.2)
mtext("X-axis label", side=1, col="gray40", line=0.5, cex=0.2)
axis(1, col="gray40", col.axis="gray40",col.ticks="gray40", cex.axis=0.4, lwd=0.5)
axis(2, col = "gray40", col.axis = "gray40", col.ticks = "gray40", cex.axis=0.4,
lwd=0.5)
dev.off()
```

Command 2.9 Plotting in R

2.4.1.10 MACs peak calling to identify unique sequence centromeres in E. asinus

MACs (Model-based analysis for ChIP-seq) is a software designed for calling peaks in next generation sequencing data. MACs identifies peaks signal accumulation by (a) identifying the number of reads and (b) the statistical significance of the data.

```
macs2 callpeak -t chip.sorted.bam –c input.sorted.bam --call-summits -n subpeaks
```

Command 2.10 Peak calling using macs

2.4.1.11 Calculating the relative abundance of CENP-A at centromere domains

In order to calculate the relative abundance of CENP-A across the unique sequence domains, a read count extraction was carried out (section 2.4.1.6) across these domains. Unwanted data were then filtered out (section 2.4.1.7). The average CENP-A associated DNA was calculated by summing the read counts of each autosomal CENP-A domain and dividing by the number of domains present. The read count at each individual centromere domain was then plotted as a function of the average CENP-A associated DNA to give relative CENP-A abundance across each unique sequence centromere.

2.4.1.12 Construction of hybrid genomes

In order to construct the hybrid genome *EquCabAsi, de novo* assembly was carried out using ChIP and Input datasets from "Asino Nuovo" (Blackjack ChIP and Input reads used for construction of *EquCabAsi-Blackjack*). Along with this PCR and sanger sequencing was used on two of the centromere peaks and compare quality with the bioinformatics *de novo* assembly. (J.G.W McCarter) The list of assembled centromere sequences were then inserted to the corresponding regions of the horse genome. Once the regions were inserted the genome was then indexed (section 2.4.1.3) and all the datasets were realigned (section 2.4.1.3)

2.4.1.13 Centromere sliding – finding the median

A metric was designed in Chapter 3 to examine the movement of centromere domains using ChIP-seq. For this, a sliding window function that runs along a centromere domain, in order to calculate the median, was developed. First total ChIP signal was calculated by summing the total read counts aligned over the centromere region. 5%, 50% and 95% of that total signal was calculated and these values were used in the sliding window function to output the genomic coordinate that corresponds to each one of the pre-calculated values. The function is outlined below in Command 2.11.

```
##Isolate centromere region
awk '{if ($2>= boundary_start && $3<= boundary_end) print;}' chip.bedgraph >
centromere.chip.bedgraph

##From isolated centromere region – sum total signal & calculate 5%,50%,95%
of value

for file in *.bedgraph; do awk '{ sum+=$4} END {print
FILENAME,"\t"sum"\t"sum*0.05"\t"sum*0.5"\t"sum*0.95}' $file > "$(basename
"$file" .bedgraph).txt"; done

##sum from centromere start site and output the genomic coordinate at which
5%,50%,95% values are reached

awk 'BEGIN {s = 0;}{s += $4;if (s >= 5%_value ) {print $0;exit;}}'
centromere.chip.bedgraph > centromere.5%.bed;
awk 'BEGIN {s = 0;}{s += $4;if (s >= 50%_value  ) {print $0;exit;}}'
centromere.chip.bedgraph > centromere.50%.bed;
awk 'BEGIN {s = 0;}{s += $4;if (s >= 95%_value ) {print $0;exit;}}'
centromere.chip.bedgraph > centromere.95%.bed;
```

Command 2.11 Calculating positional median

2.4.1.14 Estimating nucleosome positions using "NucleR"

In Chapter 4, native ChIP-seq data was used to call nucleosome positions. The R package nucleR was used to do this. Command 2.12 below, outlines the routine used in nucleR to process datasets. Files are imported in .bam format. This is followed by a number of steps that calculate coverage and carry out a normalisation routine using ChIP and input reads. The following step is the detection of nucleosome dyad, which is done using a local maxima search. Nucleosome calls are determined by selecting the surrounding bases around the dyad position, and are scored based on the height and sharpness of the peak; giving high score to large and sharp peaks and penalizing fuzziness (Flores and Orozco, 2011).

```
##Calculate the coverage, directly in reads per million (r.p.m)
cover_expt_trim_cenpa_Mono1 = coverage(reads_trim_cenpa_Mono1)
cover_ctr_trim_cenpa_Mono1 = coverage(reads_trim_cenpa_Mono1_input)

##Control correction (ChIP:Input)
corrected_cenpa_Mono1 = controlCorrection(cover_expt_trim_cenpa_Mono1,
cover_ctr_trim_cenpa_Mono1, mc.cores=1)

##Call nucleosomes
cover_clean_corrected_trim_cenpa_Mono1 =
filterFFT(corrected_cenpa_Mono1[["chrN"]][start:end], pcKeepComp=0.02)

##Call peaks
peaks_corrected_cenpa_Mono1_cen_w =
peakDetection(cover_clean_corrected_trim_cenpa_Mono1, width =147,
threshold="25%", score= TRUE)
```

Command 2.12 Nucleosome positiong calling - nucleR

Chapter 3– Identification of Satellite-free centromeres in *E. asinus*

3.1 Introduction

The centromere is responsible for the accurate segregation of sister chromatids in mitosis. Commonly known as the primary constriction, due to the highly condensed nature of the underlying DNA, the centromeric chromatin landscape is defined by the presence of a specialised nucleosome context, the histone H3 variant, CENP-A nucleosomes (Palmer et al., 1987; Sullivan et al., 1994). An important question is how are CENP-A nucleosomes organised within the centromere. Early studies using light microscopy and fluorescence in-situ hybridisation on stretched chromatin fibers have shown that CENP-A is interspersed in clusters with canonical histone H3 nucleosomes and occupies approximately half of the condensed chromatin domain (Blower et al., 2002). CENP-A organisation has also been studied in ChIP (chromatin immunoprecipitation) experiments across various neocentromeres showing that CENP-A at human neocentromeres can occupy a genomic footprint of 130-460 kb (Alonso et al., 2003, 2007; Capozzi et al., 2008; Cardone et al., 2006; Chueh et al., 2005; Lo et al., 2001a, 2001b). Later analysis using electron microscopy approaches revealed that CENP-A binds one third of the chromatin domain however CENP-A nucleosomes occupy a marginally different portion of the associated DNA volume with figures in the range of 6-8%(Marshall et al., 2008b). These studies also indicated the presence of a three-dimensional higher order chromatin compartment at the centromere that is contributory to its function by physically permitting the CENP-A domain to interact with the kinetochore at the centromere-kinetochore interface (Marshall et al., 2008b).

The recent discovery of neocentromeres in other species led to new insights into centromere domain organisation. The orang-utan genome contains one neocentromere (confirmed by ChIP-chip analysis using CENP-A and CENP-C antibodies) on chromosome 12, which spans ~225 kb (Locke et al., 2011). Chickens possess three non-repetitive centromeres on chromosomes 5, 27 and Z and through ChIP-seq approaches they showed that the CENP-A domains occupy ~40 kb of DNA

which is marginally shorter than previously seen in clinical neocentromeres (Shang et al., 2010, 2013). Interestingly in these chicken studies, they showed that conditional centromere deletion led to neocentromere formation on both transcriptionally active and inactive genomic regions which is somewhat in contrast to previous studies which showed neocentromeres primarily occupied gene deserts (Cardone et al., 2006; Lo et al., 2001b; Wade et al., 2009). While CENP-A overexpression in chicken cells did not result in any expansion of the CENP-A at neocentromere domains (Shang et al., 2013), CENP-A overexpression in human cells showed increased CENP-A levels at satellite-containing centromeres (Bodor et al., 2014). This finding may suggest that satellite-containing centromeres are able to incorporate and house more CENP-A over neocentromeres, possibly due to their underlying repetitive DNA nature which contains the 17 bp "CENP-B box" motif that permits CENP-B to bind (Masumoto et al., 2004; Muro et al., 1992), and stabilise CENP-A nucleosomes (Fachinetti et al., 2015; Fujita et al., 2015).

The centromere on chromosome 11 of the domestic horse (*Equus caballus*) is completely devoid of satellite DNA and through ChIP-chip analysis CENP-A and CENP-C were shown to co-occupy two distinct domains (136 kb & 99 kb), over a ~400 kb genomic region which were separated by a 165 kb interval (Wade et al., 2009). Further studies in equids using five unrelated horses, that were subject to ChIP-chip and chromatin fiber FISH analysis revealed that the centromeres on chromosome 11 occupied different positions within the ~400 kb domain. In two of the horses, CENP-A was contained in two distinct domains as seen before and in three of the horse individuals CENP-A was contained in a single domain. The CENP-A domains were analysed using a SNP-based approach and revealed that the multi-domain CENP-A binding represented distinct positional alleles present on the two homologous chromosomes (Purgato et al., 2015). This arrangement was also seen in the single CENP-A binding domains and the authors established that the single domains were again individual CENP-A alleles that slightly overlapped one another (Purgato et al., 2015). This is the second instance of epiallelism shown in centromeres as previous studies identified CENP-A domains that can bind different juxtaposed α-satellite arrays (D17Z1-A & D17Z1-B) on either of the two human chromosome 17 homologues (Maloney et al., 2012). Together these studies introduced centromere allelism and

centromere sliding as an aspect of neo/centromere function whilst again reinforcing that centromeres are defined epigenetically in a DNA independent manner. Further studies in the genus *Equus* and in particular *E. asinus* (the domestic donkey), *Equus grevyi* (Grevy's zebra) and *Equus burchelli* (Burchelli's zebra) provided more insight into equid centromere evolution. In-situ hybridisation experiments with total genomic DNA and two of the most common equid satellites "37-cen" and "2PI" revealed that several chromosomes across the karyotypes were devoid of satellite DNA and while a number of the satellite containing chromosomes were present, some of these were devoid of satellites at primary constrictions (Piras et al., 2010).

E. asinus provides an opportunity to examine centromere architecture in a mammal. A useful system to investigate the molecular architecture of centromeres at a single base resolution toward this, a ChIP-seq approach was taken to identify and characterise the satellite-free centromeres in *E. asinus.*

3.2 Statement of effort

This experimental procedures and analyses presented in this chapter are derived from an extensive collaboration between the labs of K.F Sullivan and E. Giulotto. CENP-A ChIP preparations were performed by members of the E. Giulotto group. J.G.W McCarter provided assistance in preparation of one mule CENP-A ChIP. Bioinformatics operations (analysis and visualisation) mostly developed and applied by J.G.W McCarter, however this does not include *de novo* assembly.

3.3 ChIP-seq identifies 16 unique sequence centromeres in *E. asinus*

Chromatin immunoprecipitation followed by high throughput sequencing (ChIP-seq) was used to investigate centromeric chromatin structure in *E. asinus*. Two antibodies with affinity for horse CENP-A; a rabbit antiserum against CENP-A peptide (Purgato et al., 2015; Trazzi et al., 2009; Wade et al., 2009) and human autoantisera with a high titer for CENP-A (see Table 2.1) were used to capture CENP-A bound chromatin from formaldehyde stabilised donkey skin fibroblasts from the donkey "Asino Nuovo" (AN), prepared by sonication. As a control ChIP-seq experiments were also performed on horse skin fibroblasts showing distinct binding to the unique sequence centromere on chromosome 11 in Figure 3.1 below.

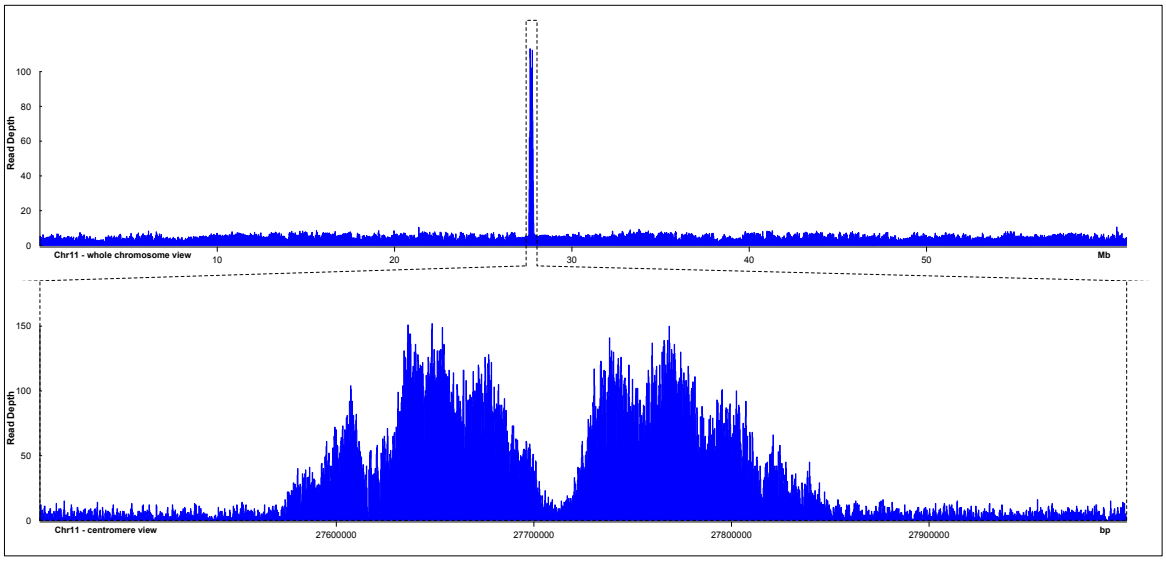


Figure 3.1 CENP-A ChIP-seq shows structure of horse centromere on chromosome 11
Whole chromosome view and centromeric view of CENP-A binding resolved by ChIP-seq with CENP-A immunoprecipitated DNAs, on horse chromosome 11. The domain occupies ~250 kb region of the genome and displays a multi-domain profile.

Libraries were prepared from immunoprecipitated DNA and subject to paired-end sequencing using an Illumina HiSeq2000 by IGA Technologies, Udine, Italy. A bioinformatics pipeline was developed in order to quality control the data and process the sequence reads according to current standards (Landt et al., 2012; Langmead and Salzberg, 2012; Li et al., 2009). ChIP-seq reads in the FASTQ format encode quality scores as part of the FASTQ structure (Cock et al., 2009). The quality score encoded is in the Phred format and is currently the *de facto* standard for marking read base qualities (Cock et al., 2009). Phred is defined in terms of probability of error (Ewing and Green, 1998; Ewing et al., 1998): $Q_{PHRED} = -10 \times \log_{10}(P_e)$. In order to check the

range of Phred scores across each base position in a read, *FastQC* was used and a whisker plot was generated to display these data (Figure 3.2). The per base quality scores for CENP-A (peptide antibody) ChIP and input libraries are displayed. The lower quartile (bottom of yellow box) and median (red line) values are above the Phred values 10 and 25 respectively, indicating that the FASTQ reads for both CENP-A datasets were of good quality for alignment.

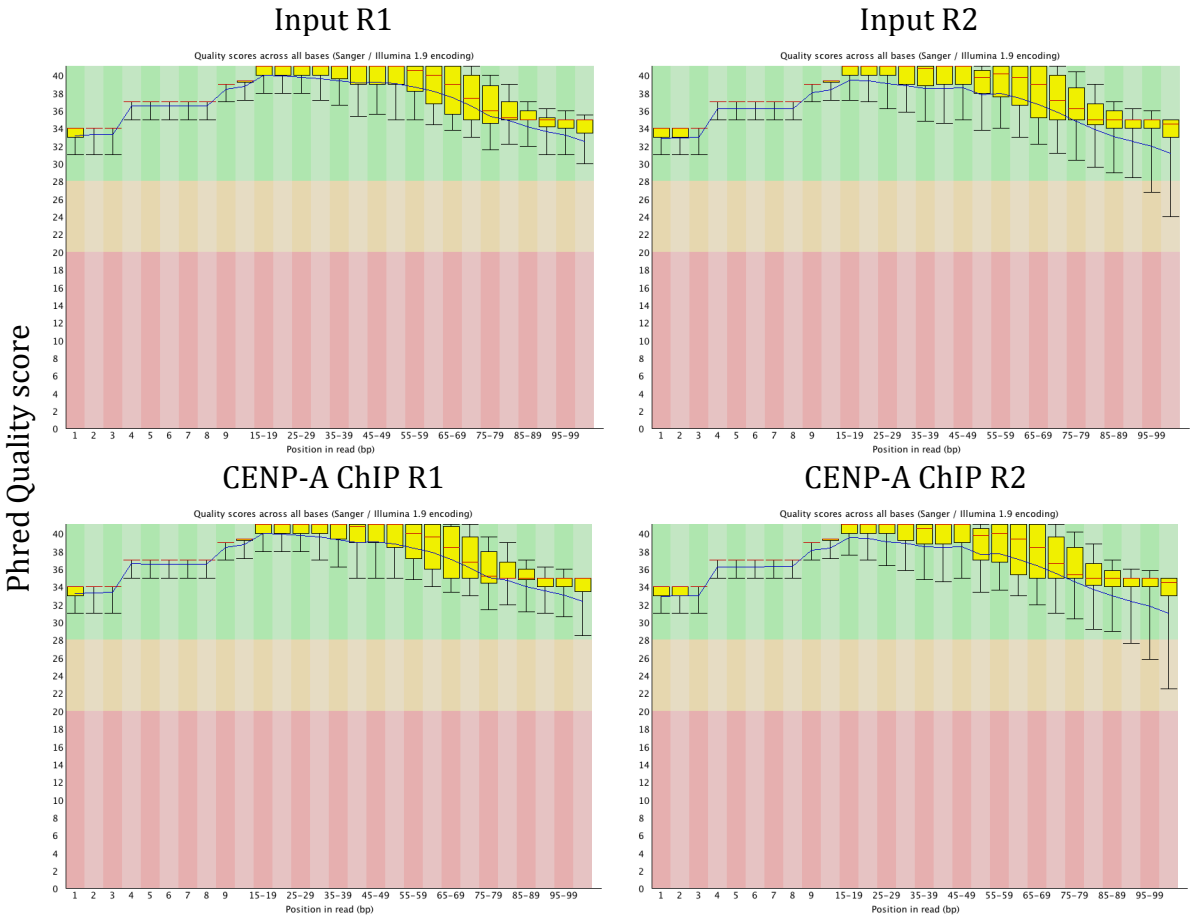


Figure 3.2 FastQC per base read quality metrics
Per base quality metrics for CENP-A (peptide antibody) ChIP and Input libraries. Plots generated using *FastQC*. Phred score is defined in terms of the estimated probability of error. Phred score of 30 is the equivalent to a 1/1000 chance of a called base being incorrect, by the equation $q = -10 \times \log_{10}(p)$.

In order to visualise the donkey CENP-A ChIP-seq data, reads were aligned to the horse reference genome (Wade et al., 2009). The horse and donkey species evolved from a common ancestor and are known to have highly similar genomes (in the order of 98-99%) (Orlando et al., 2013). The horse genome; *EquCab2.0* (September 2007 release) was downloaded from the NCBI genome database and indexed using Command 2.1 ChIP and Input reads in *FASTQ* format were aligned to the indexed horse genome using *Bowtie2* (Langmead and Salzberg, 2012). *Sensitive* and *paired-end* mode were used for the alignment. The outputted SAM file was then converted to its binary counterpart, BAM, using Command 2.3. In order to isolate true CENP-A signal across the genome, the peak caller *MACS* was used. *MACS* identifies peak signal accumulation by (a) identifying the number of reads and (b) statistical significance of the data (Zhang et al., 2008). Data are outputted in BED file format with four key parameters; *fold enrichment*, *pileup*, *p-value* and location. A set logical criteria was established in Microsoft Excel to filter regions in the genome that were more statistically significant to have CENP-A association. This was done using the "IF" and "AND" function in excel. The argument would set a threshold for reads that have a fold enrichment and pileup of more than ≥ 80 or a p value ≥ 8. Using this approach, 16 chromosomal locations were identified as shown in the table below. These chromosomal regions of interest, spanning between ~30-300 kb were then used to identify a set of BACs that spanned each region of interest. The BAC DNAs were then used in FISH experiments to show binding in horse and donkey chromosomes relative to the primary constriction (Figure 3.3) (E. Giulotto).

Table 3.1 MACS output – 16 chromosomal locations

EAS chr	ECA chr	Start	End	Span
Eas4	chr28	12,890,000	13,020,000	130,000
Eas5	chr19	4,974,000	5,131,000	157,000
Eas7	chr8	41,980,000	42,100,000	120,000
Eas8	chr20	26,440,000	26,530,000	90,000
Eas9	chr14	29,650,000	29,680,000	30,000
Eas10	chr25	8,614,000	8,836,000	222,000
Eas11	chr17	16,780,000	16,860,000	80,000
Eas12	chr9	31,980,000	32,270,000	290,000
Eas13	chr11	46,660,000	46,920,000	260,000
Eas14	chr13	7,262,000	7,521,000	259,000
Eas16	chr5	74,890,000	74,960,000	70,000
Eas18	chr26	22,380,000	22,510,000	130,000
Eas19	chr6	14,190,000	14,250,000	60,000
Eas27	chr27	19,730,000	19,870,000	140,000
Eas30	chr30	17,720,000	17,810,000	90,000
EasX	chrX	26,990,000	27,070,000	80,000

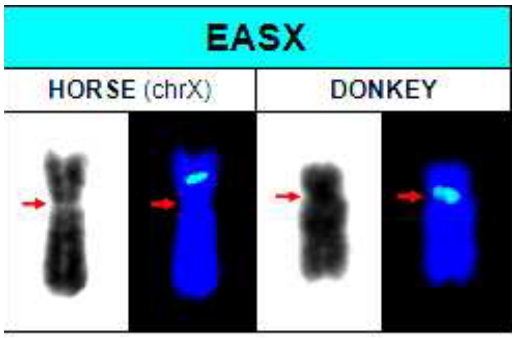


Figure 3.3 Cytogenetic conformation of ChIP-seq data
Figure adapted from Doctorate thesis of Federico Cerutti, "Molecular organization of epigenetically specified centromeric domains", (2014).

Shown in the figure is a FISH experiment performed on the chromosome X of the horse and donkey. In horse, the BAC (green) localises to a region on the chromosome that is not the centromere (red arrow). The same BAC hybridizes to the primary constriction region on donkey chromosome X indicating the region of centromere function.

In order to examine the fragment length between mapped read pairs, the *Picard tools* function *CollectInsertSizeMetrics* was performed on the alignment files and restricted to the CENP-A regions in Table 3.1. Shown in Figure 3.4 is the distribution of fragment lengths for the CENP-A ChIP and input datasets. The peptide antibody shows a range of ~150-580 bp with the majority of fragments between ~300-400 bp. A similar distribution is seen in the input dataset. The CREST dataset shows a broad distribution between ~150-600 bp with the majority of fragments in~200-400 bp range.

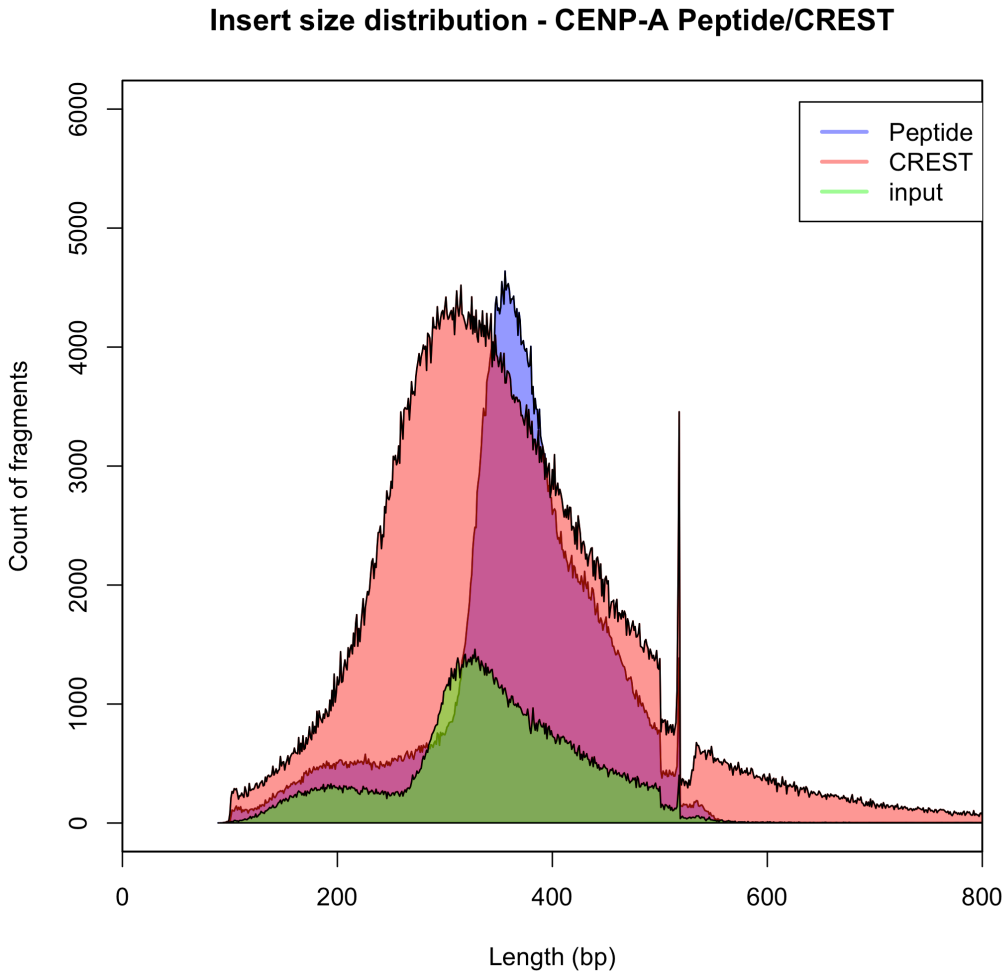


Figure 3.4 Fragment length distribution CENP-A
Fragment length distribution in the CENP-A datasets show input (green), CREST (pink) and CENP-A peptide antibody (blue)

A method to examine the quality of the immunoprecipitation is by measuring enrichment of signal associated with centromeres. This is done using FRiP (Fraction of reads in peaks), which calculates the percentage of reads in a dataset that fall within significantly enriched regions (Landt et al., 2012). FRiP analysis was performed on the CENP-A regions from Table 3.1 and the data show that both CENP-A datasets are well above the desired 1% value indicating the immunoprecipitations were successful.

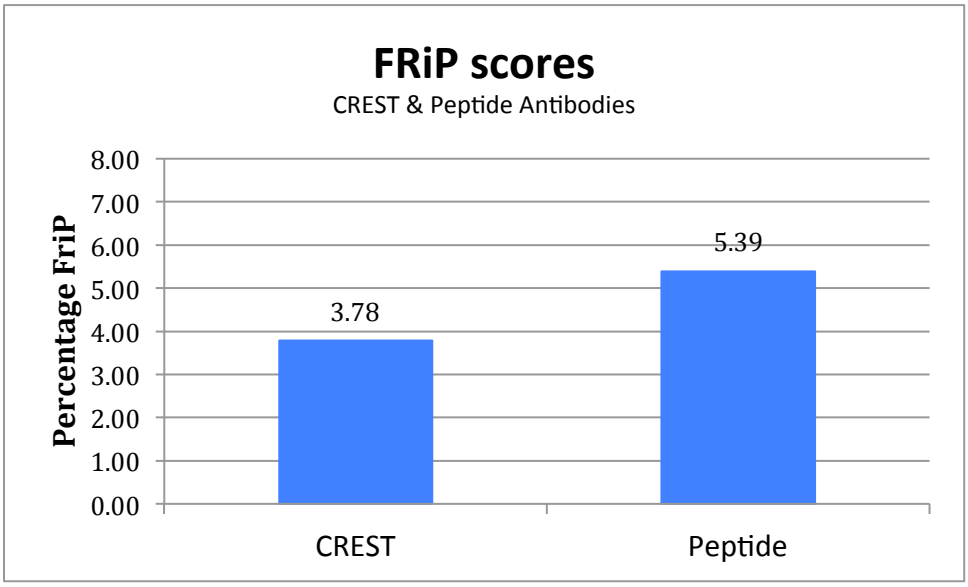


Figure 3.5 Quality control – FRiP
FRiP analysis shows 3.78-5.39% of CENP-A signal lies within the centromere regions, compared to the whole genome, indicating successful CENP-A ChIP.

In order to visualise the regions of CENP-A binding, the UCSC genome browser was used. It contains a library of assembled genomes, which includes the horse reference genome *EquCab2.0*. Following the information provided by the UCSC genome browser at http://genome.ucsc.edu/goldenpath/help/customTrack.html the BED file output data was converted for use in the UCSC genome browser. Using the 16 CENP-A regions, UCSC genome browser track files were assembled and uploaded to the browser. These tracks included information that displayed approximate coordinates of CENP-A domain boundaries as well as enrichment profiles marking CENP-A binding across each chromosome. As an example, Figure 3.6 below shows the ECA8/EAS7 CENP-A domain across a 500 kb window. The location of BACs identified that corresponded to the sites of each CENP-A binding domain were converted to track files and uploaded. The "ECA BAC" bar (light blue) show the coordinates of these BACs, confirming the location of the CENP-A domains. "EAS-B4 pileup", "EAS-B4 log(p)" and "EAS-B4 fold

enrichment" are values taken directly from the MACs output BED file and their signal values are represented by colour intensity. "USP Segment Definitions" is a track that displays the boundaries of the CENP-A domains, constructed by extracting maximum boundaries from the MACs generated BED file. BIGWIG files were used to display peak structure in detail and the example shown in Figure 3.6 illustrates CENP-A binding on horse chromosome 8/donkey chromosome 7 (ECA8/EAS7) ("chr8ChIP"). This browser tool served as a resource for collating data from the original donkey ChIP-seq experiment. A catalogue of the UCSC genome browser figures can be viewed in the Appendix I-Table 8.1.

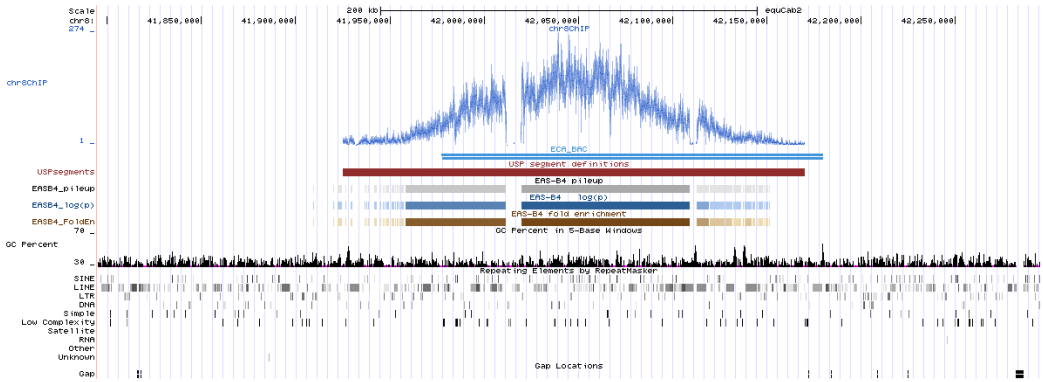


Figure 3.6 UCSC genome browser track files (ECA8/EAS7)
Image from UCSC genome browser describing the CENP-A domain that aligns to ~200 kb region on chromosome 8 in horse (chromosome 7 in donkey). CENP-A (bigwig) signal resembles a Gaussian-like distribution. Horse BAC ("ECA BAC") in light blue shows the localisation of the BAC used to identify the region in FISH. MACs peak calling output data "EAS-B4 pileup", "EAS-B4 log(p)" & "EAS-B4 fold enrichment" showing read count profile across the domain. "USP segment definitions" identify CENP-A boundaries over the region.

CENP-A domain properties were defined based on the shape of the distribution. Four types of CENP-A binding profile were resolved: "Gaussian-like", "Spike-like", "Multi-domain" and "Complex". The most common type of CENP-A profile among the selection exhibited a Gaussian-like distribution. Four of the profiles contained an intense, highly fixed signal. These appear to be regions of genomic amplification as identical amplification was observed in the input DNA. These were catalogued as "Spike-like" peaks. A selection of CENP-A profiles did not appear to have any typical shape and thus determined to be "Complex" and finally numerous CENP-A profiles exhibited "Multi-domain" binding. Each peak profile is displayed in Figure 3.7 and peak taxa are described further.

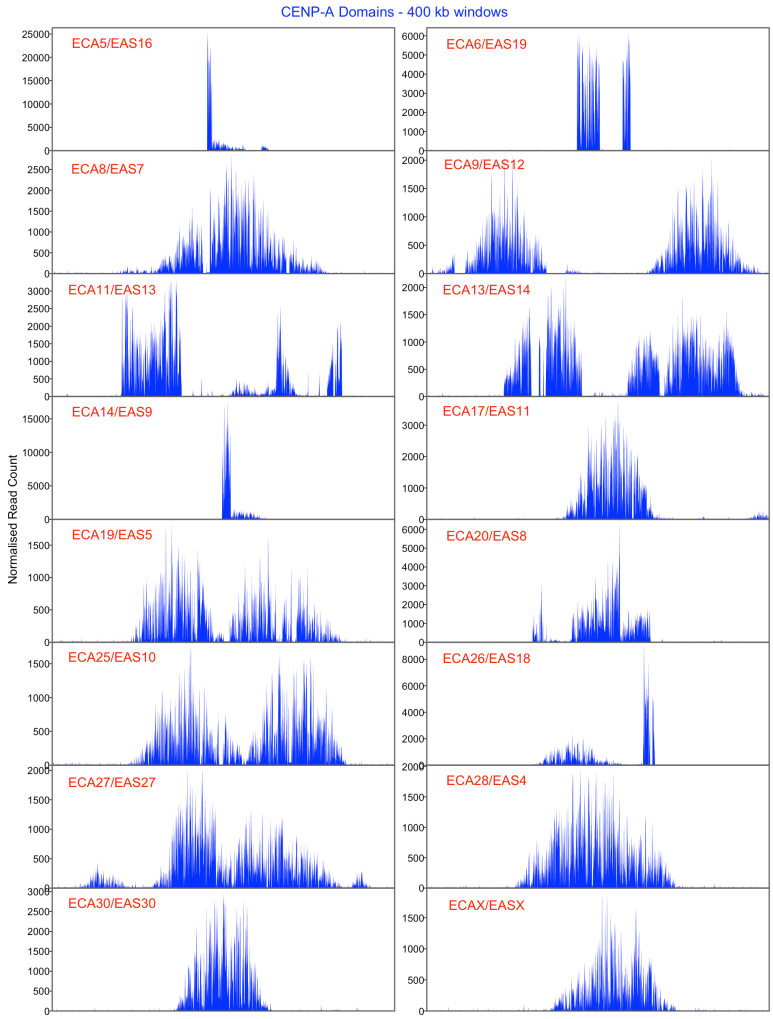


Figure 3.7 CENP-A binding profiles on *EquCab2.0*
16 CENP-A distributions obtained from CENP-A ChIP-seq reads mapped to the horse reference genome *EquCab2.0*. Horse chromosomes and corresponding donkey chromosomes are outlined in red. The CENP-A domains (400 kb window views) display four types of distributions, "Gaussian-like", "Spike-like", "Complex" and "Multi-domain".

CENP-A domains identified occupy a span in the range of ~60-320 kb and all 16 domains localized to satellite-free regions of the horse genome. The centromeres that exhibit Gaussian-like distributions are ECA8/EAS7, ECA17/EAS11, ECA28/EAS4, ECA30/EAS30 and ECAX/EASX. These domains show a balanced distribution with gradual even loss of signal approaching the domain boundaries. This signal profile is likely to represent a statistical distribution of CENP-A over a population of cells with the majority of cells containing CENP-A in the central region of the distribution. Some of the multi-domain centromeres also profile as multiple Gaussian-like distributions. Multi-domain CENP-A binding can be seen in ECA9/EAS12, ECA13/EAS14, ECA19/EAS10, ECA25/EAS10 and ECA26/EAS18. Each one of the domains are likely to be centromeres that occupy different positions on each of the chromosome homologues (Purgato et al., 2015).

Some of the spike domains show a highly fixed intense CENP-A signal, with a Gaussian-like subdomain (ECA5/EAS16, ECA14/EAS9). The intense signal may represent regions of genomic amplification. ECA26/EAS18 exhibits a multi-domain distribution; one of the domains is Gaussian-like (left domain) while the other is a spike domain (right domain). Another CENP-A domain catalogued as spike-like was ECA6/EAS19. This domain exhibits two intense signal spikes separated with a large gap with very abrupt boundaries.

The centromere domains catalogued as complex are the CENP-A profiles of ECA11/EAS13 and ECA20/EAS8. These distributions exhibit no shape comparable to previous domains and may be due to differences in the horse and donkey genomes.

These data are summarized in Table 3.2, which includes domain coordinates and span.

Table 3.2 CENP-A domain taxonomy

E.ca	E.as	Coordinates	Span (bp)	Taxonomy
5	16	74,873,953-74,934,671	**60,718**	**Spike/Gaussian** : Intense spike signal with subdomain broken Gaussian-like structure
6	19	141,80,841-14,252,211	**71,370**	**Spike** : Intense spike signal with large gap at central right
8	7	41,939,723-42,131,450	**191,727**	**Gaussian-like** : Gaussian like structure with numerous gaps throughout
9	12	31,944,937-32,253,365	**308,428**	**Multi-domain** : Multi-domain Gaussian-like structure with numerous gaps througout
11	13	46,713,711-46,844,122	**130,411**	**Complex/ Multi-domain** Multi-domain complex structured signal with gap in central and rightward region
13	14	7,222,815-7,487,414	**264,599**	**Multi-domain** : Multi-domain Gaussian-like structure with gaps in both domains
14	9	29,616,631-29,689,495	**72,864**	**Spike/Gaussian** : Intense spike signal with subdomain Gaussian like structure
17	11	16,742,696-16,855,544	**112,848**	**Gaussian-like** : Gaussian like structure with a selection of notches throughout
19	5	4,918,350-5,153,067	**234,717**	**Multi-domain** : Multi-domain Gaussian like structure with a numerous gaps throughout
20	8	26,417,453-26,509,636	**92,183**	**Complex** : Ultimately unstructured and not typical of neocentromere distribution
25	10	8,577,363-8,817,825	**240,462**	**Multi-domain** : Multi-domain Gaussian like structure with numerous gaps throughout
26	18	22,368,144-22,527,763	**159,619**	**Gaussian/Spike/Multi-domain** : Multi-domain Gaussian-like structure with spike-like signal intensity
27	27	19,634,011-19,957,570	**323,559**	**Gaussian-like** : Gaussian like structure with numerous sub-domain structures
28	4	12,870,911-13,058,463	**187,552**	**Gaussian-like** : Gaussian like structure with a selection of notches throughout
30	30	17,713,318-17,823,448	**110,130**	**Gaussian-like** : Gaussian like structure with a selection of notches throughout
X	X	26,936,344-27,073,478	**137,134**	**Gaussian-like** : Gaussian like structure with a selection of notches throughout – Haploid signal

While 16 satellite-free CENP-A domains have been identified in donkey, it remains possible that other satellite-free centromeres have yet to be identified. This observation comes from cytogenetic level analysis using FISH were it was established that at least 18 centromeres in donkey did not contain satellite signal (Piras et al., 2010). One potential reason why these centromeres haven't been identified in this study maybe that corresponding sequences are not present in the horse genome. A fully assembled donkey genome would be helpful in this matter. The donkey satellite-free centromere domains identified by ChIP-seq are indicated in Figure 3.8.

Collectively these data describe the identification of 16 satellite-free centromere domains in *E. asinus* which will prove as a useful genomic tool for investigating centromere structure and function.

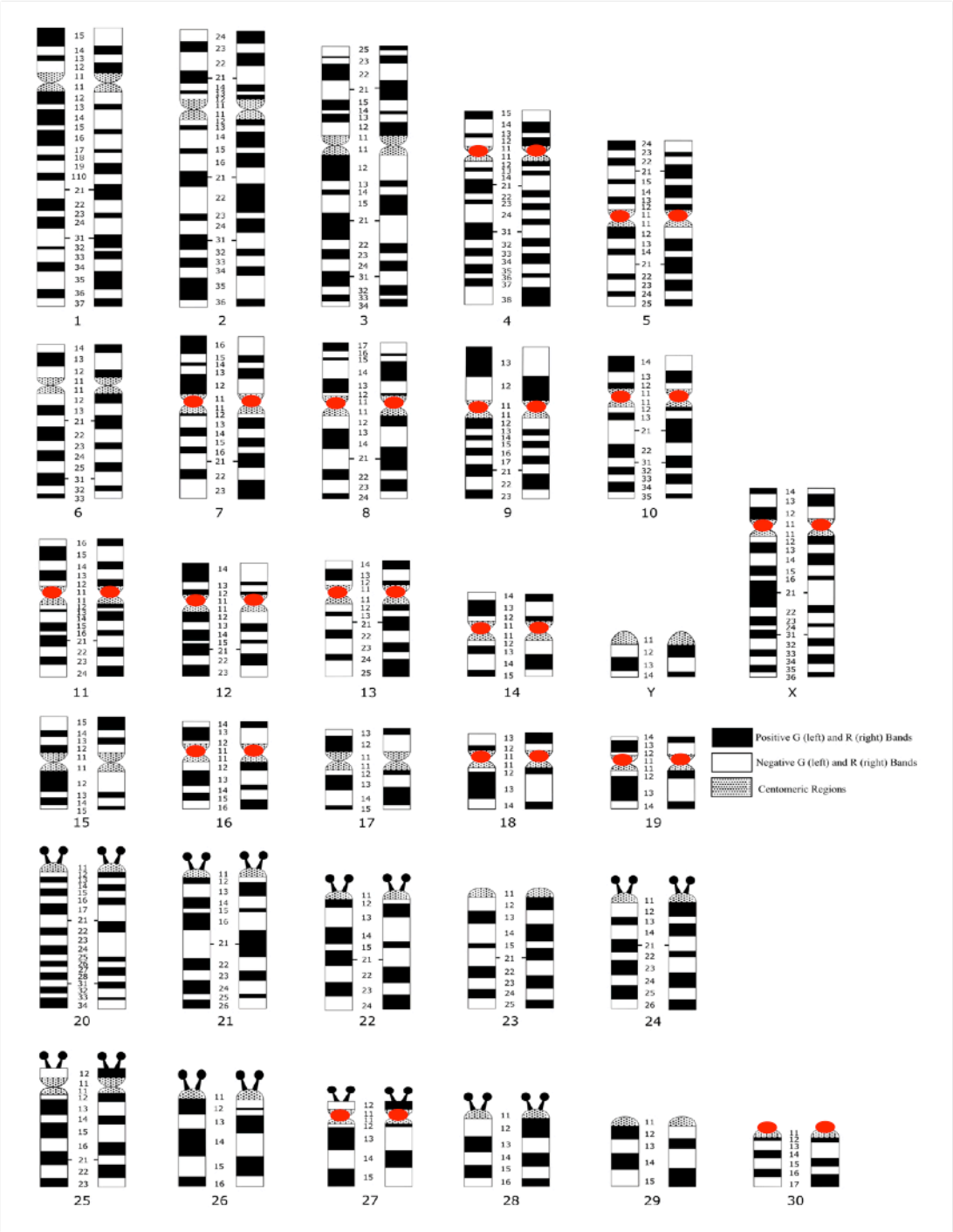


Figure 3.8 Donkey chromosomes containing satellite-free centromeres
Illustrated is a G-banded donkey karyotype adapted from (Di Meo et al., 2009) showing the donkey chromosomes containing satellite-free centromere regions (red) which were identified in this study.

3.3.1 Sequence determination of CENP-A binding domains in donkey

The ChIP-seq reads mapped back to the horse genome allowed for identification of 16 satellite-free CENP-A binding domains belonging to the donkey karyotype and their locations on the homologous horse chromosomes. A number of read-free gaps can be observed over the CENP-A binding domains aligned to *EquCab2.0*. Examples of these gaps can be seen in ECA6/EAS19, ECA8/EAS7 in Figure 3.7. Along with gaps, some of the peaks exhibited an unusual distribution e.g. ECA11/EAS13 (Figure 3.7) which was not observed in previous studies on neocentromeres (Shang et al., 2010, 2013; Wade et al., 2009). However, these anomalous features could be an artefact of the heterologous ChIP-seq mapping and possibly reflect rearrangements between the two genomes. The horse and donkey genomes have an identity of ~98% (Orlando et al., 2013) and in order to determine whether these gap features were a real representation of the alignment or due to heterologous mapping where endogenous donkey centromere DNA sequences differed from corresponding regions in the horse, native donkey centromere sequences were determined, by the Giulotto lab. Here, DNA across the centromere domains were constructed through *de novo* sequence assembly using the "Asino Nuovo" ChIP-seq reads. The newly assembled donkey centromere sequences provide us with a tool to examine donkey centromeres with higher accuracy.

3.3.2 Construction of EquCabAsi

In order to construct a more accurate genomic tool for examination of donkey centromeres, a hybrid genome was assembled, using the sequences determined by the Giulotto lab. Simply, defined segments of *EquCab2.0* were replaced *in silico* by the new donkey centromere sequences. The coordinates of the genomic regions in *EquCab2.0* that were replaced by newly refined donkey centromere sequences are outlined in Table 3.3 below. This genome is referred to as *EquCabAsi*. ChIP-seq reads were realigned and there were a number of observable differences in each of the domains that provided us with redefined peak distributions, which we believe accurately represent the CENP-A binding domains of these chromosomes. Revised CENP-A locations are detailed in Table 3.4.

Table 3.3 EquCabAsi - "Asino Nuovo" insert coordinates

Chromosome		Start	End
EAS4	ECA28	12,857,498	13,067,693
EAS5	ECA19	4,870,728	5,193,891
EAS7	ECA8	41,899,430	42,119,519
EAS8	ECA20	26,350,299	26,600,000
EAS9	ECA14	29,550,027	29,850,000
EAS10	ECA25	8,536,970	8,900,000
EAS11	ECA17	16,699,392	16,923,707
EAS12	ECA9	31,929,515	32,336,449
EAS13	ECA11	46,559,001	46,974,920
EAS14	ECA13	7,178,188	7,596,941
EAS16	ECA5	74,800,027	75,010,729
EAS18	ECA26	22,328,608	22,543,315
EAS19	ECA6	14,124,005	14,305,040
EAS27	ECA27	19,665,636	19,975,899
EAS30	ECA30	17,699,336	17,872,653
EASX	ECAX	26,900,001	27,139,288

Table 3.4 Revised CENP-A locations - *EquCabAsi*

E.ca Chr	E.as Chr	Peak 1 Start : End		Span (kb)	Peak 2 Start : End		Span (kb)
5	16	74,873,953	74,934,671	60.7	-	-	-
6	19	14,180,841	14,252,211	71.3	-	-	-
8	7	41,939,723	42,131,450	191.7	-	-	-
9	12	31,944,937	32,064,497	119.5	32,126,050	32,253,365	127.3
11	13	46,713,711	46,844,122	130.4	-	-	-
13	14	7,222,815	7,300,170	77.3	7,343,791	7,487,414	143.6
14	9	29,616,631	29,689,495	72.8	-	-	-
17	11	16,742,696	16,855,544	112.8	-	-	-
19	5	4,918,350	5,031,108	112.7	5,028,889	5,153,067	124.1
20	8	26,417,453	26,509,636	92.1	-	-	-
25	10	8,577,363	8,701,423	124	8,701,423	8,817,825	116.4
26	18	22,367,198	22,483,635	116.4	-	-	-
27	27	19,634,011	19,957,570	323.5	22,494,438	22,527,763	33.3
28	4	12,870,911	13,058,463	187.5	-	-	-
30	30	17,713,318	17,823,448	110.1	-	-	-
X	X	26,936,344	27,073,478	137.1	-	-	-

Some centromeres exhibited large structural alterations after the donkey centromere sequences were assembled. For example, Figure 3.9 shows the alterations that occurred in ECA11/EAS13. Here, the rightmost patch of signal was translocated to the left side of the CENP-A domain. The mid region gap is a 110 kb interval that is not present in donkey thereby resulting in the signal at each side of the gap to join.

All centromeres exhibited DNA sequence alterations, in the form of insertions, deletions or rearrangements. The redefined alignments are displayed in Figure 3.10. These data shows the extent of sequence differences at the centromere regions in donkey compared to corresponding regions in horse. Spike peaks were shown to be genuinely amplified in this analysis e.g ECA14/EAS9 or ECA26/EAS18 (Figure 3.10-D, -F). Whether these alterations occur more frequently at centromere domains compared to other non-centromeric locations is currently unknown .

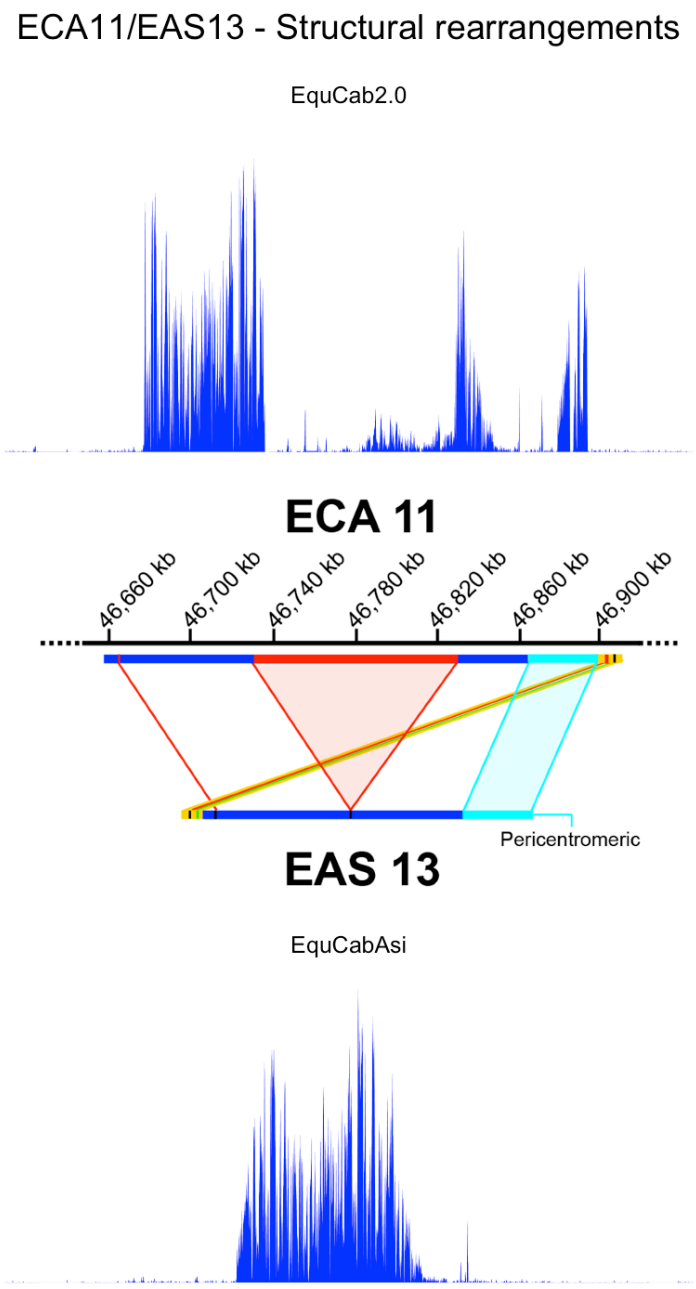


Figure 3.9 Structural alterations in centromere domains
Example of chromosomal rearrangement events at donkey centromeres when compared to corresponding regions of horse genome. Rearrangements included ~110 kb deletion at the donkey centromere EAS13 compared to homologous binding region in horse ECA11 (left).

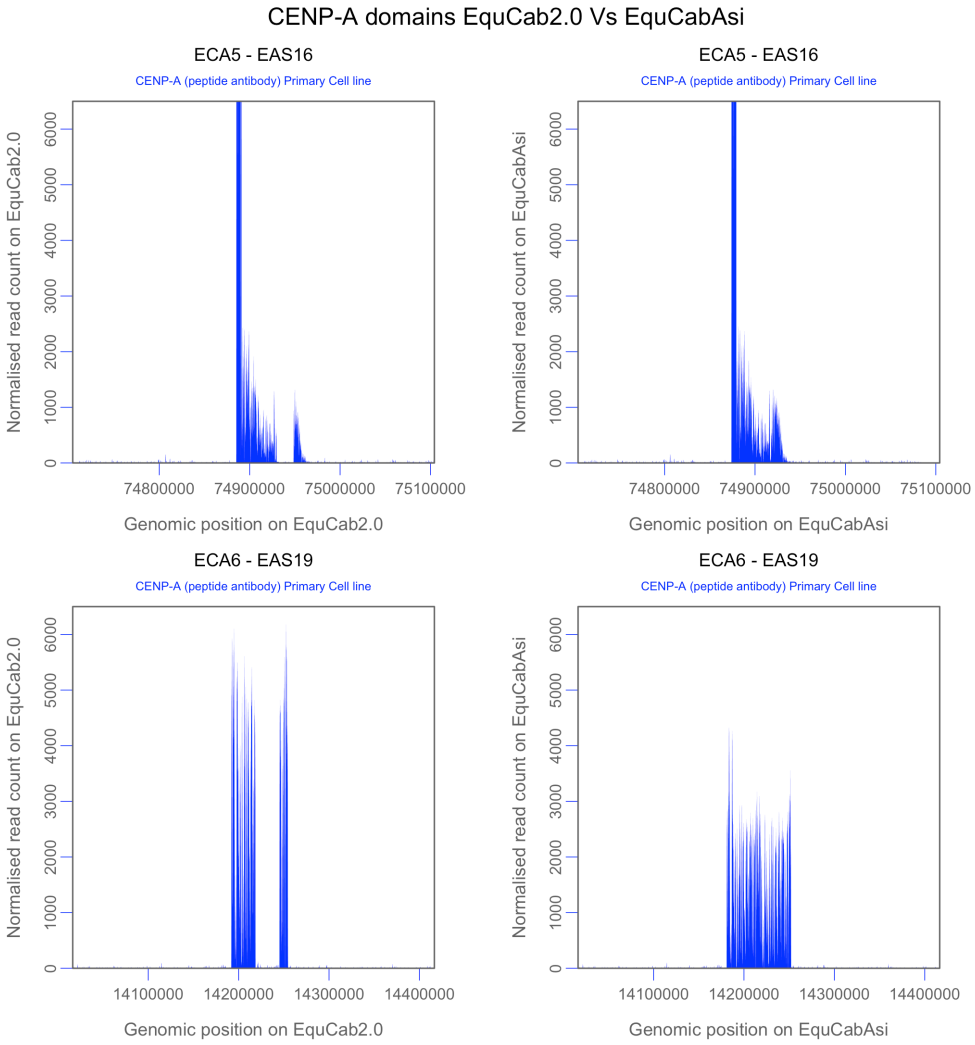


Figure 3.10 – A CENP-A domains - hybrid genome comparasion
CENP-A distribution compared in two reference genomes (500 kb windows). Panels on the left are centromeres aligned to the horse reference genome *EquCab2.0*. Panels on the right are centromeres mapped back to *EquCabAsi*. ECA5/EAS16 and ECA6/EAS19 are shown above. Prominent gaps are observed in both centromere domains aligned to the horse genome and these regions are covered in the corresponding hybrid genome

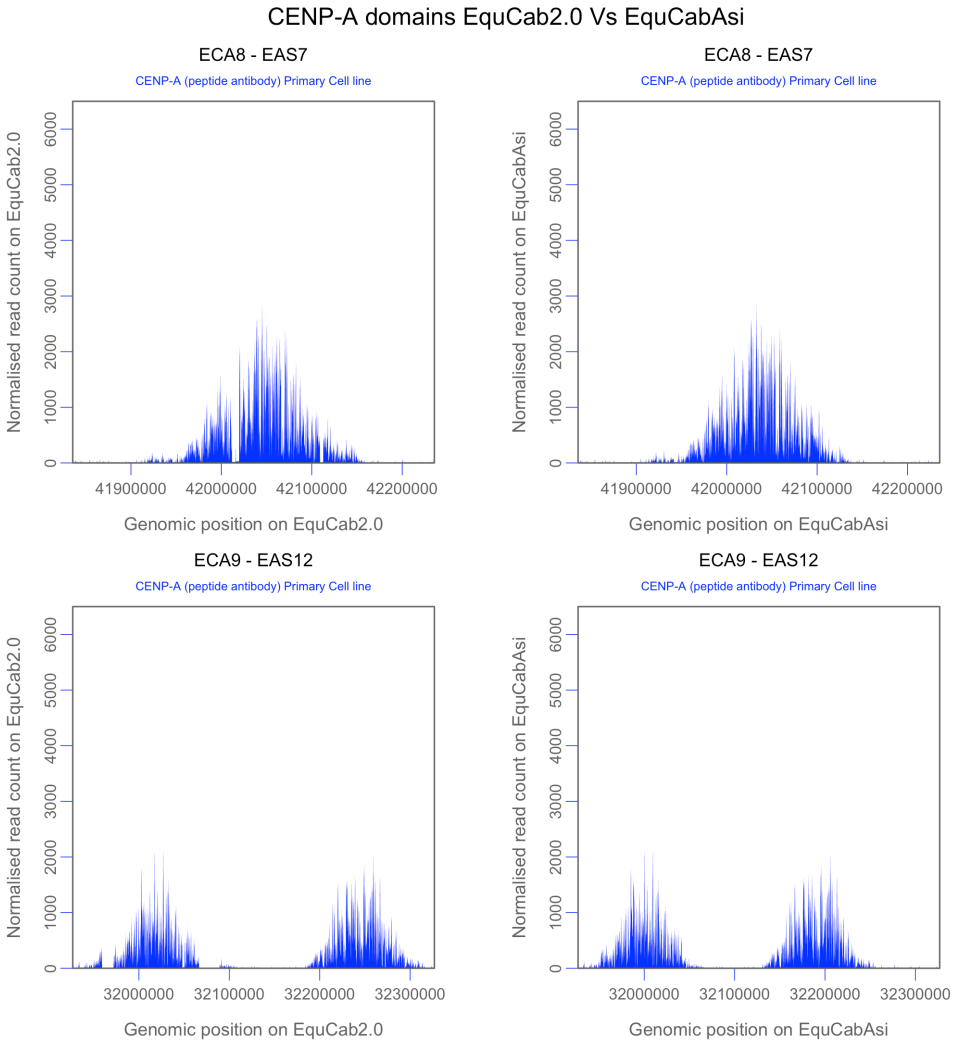


Figure 3.10-B CENP-A domains – hybrid genome comparison
ECA8/EAS7 and ECA9/EAS12 display a number of gaps in the horse genome however the signal is mostly conserved at both centromeres in the two reference genomes.

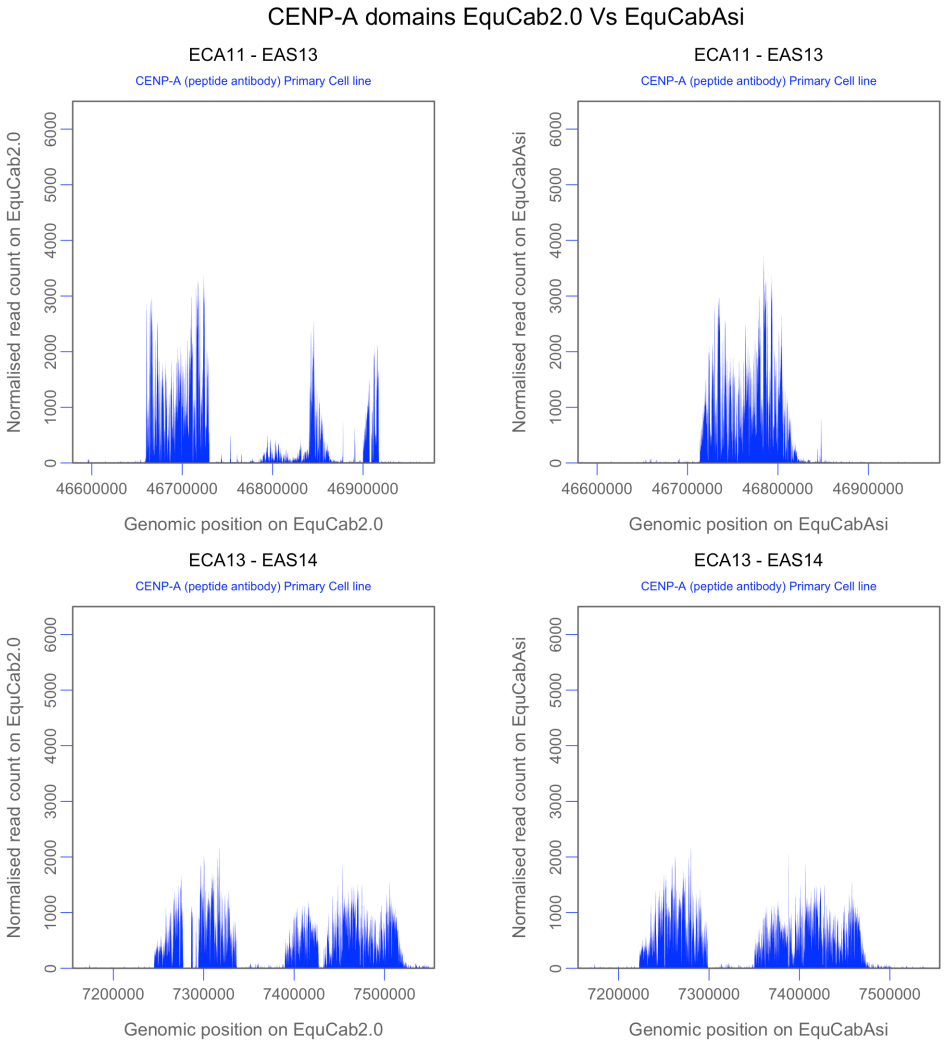


Figure 3.10-C CENP-A domains – hybrid genome comparison
ECA11/EAS13 exhibits a complex structure with patches of signal spaced between large gaps in *EquCab2.0*. However in *EqCabAsi* the CENP-A signal forms a Gaussian-like distribution as seen in other centromeres. ECA13/EAS14 shows a multi-domain centromere profile with a noticeable gap in the left domain. This gap is covered in the *EquCabAsi* alignment.

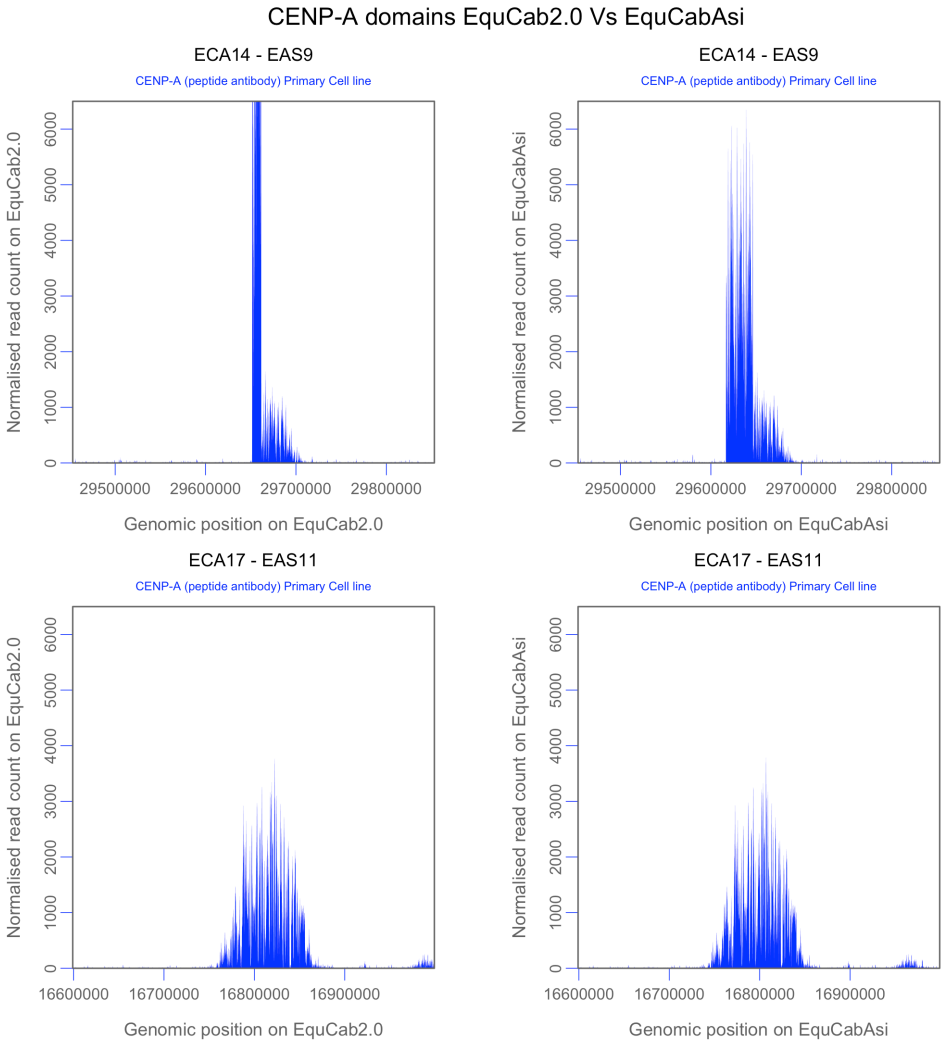


Figure 3.10-D CENP-A domains – hybrid genome comparison
ECA14/EAS9 forms a "spike-like" peak structure with a Gaussian-like sub-doamin in *EquCab2.0* however in *EquCabAsi* the spike signal is reduced. This is due to copy number of the DNA sequence underneath the spike. The DNA sequence under the spike is present in three copies in *EquCabAsi*. ECA17/EAS11 exhibits a very conserved distribution in both *EquCab2.0* and *EquCabAsi.*

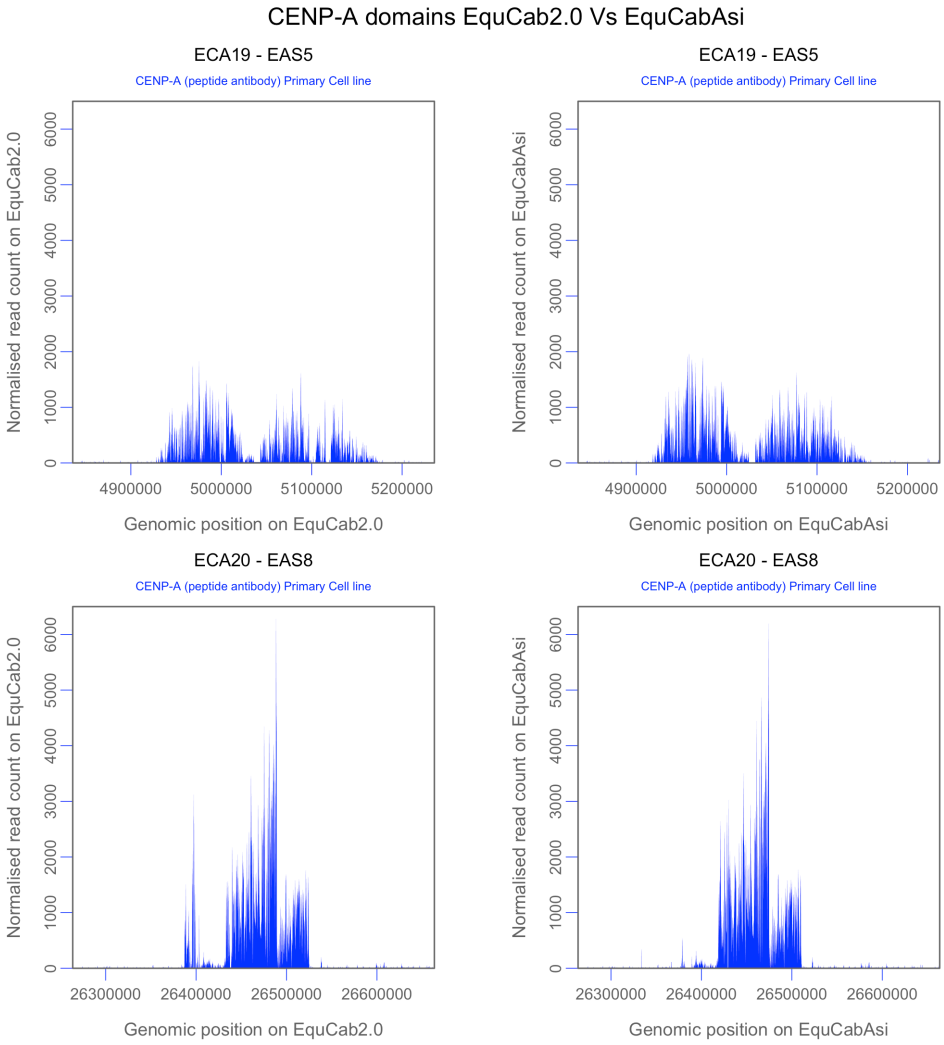


Figure 3.10-E CENP-A domains – hybrid genome comparison
ECA19/EAS5 exhibits minor gaps throughout the multi-domain structure in *EquCab2.0*. In *EquCabAsi*, these gaps are covered. ECA20/EAS8 exhibits a "complex" structure in both genomes with a signal spike to the left of the peak in *EquCab2.0*. This spike in signal is not present in the CREST dataset (Appendix I-Figure 8.1) or in the *EquCabAsi* alignment possibly suggesting an experimental artifact.

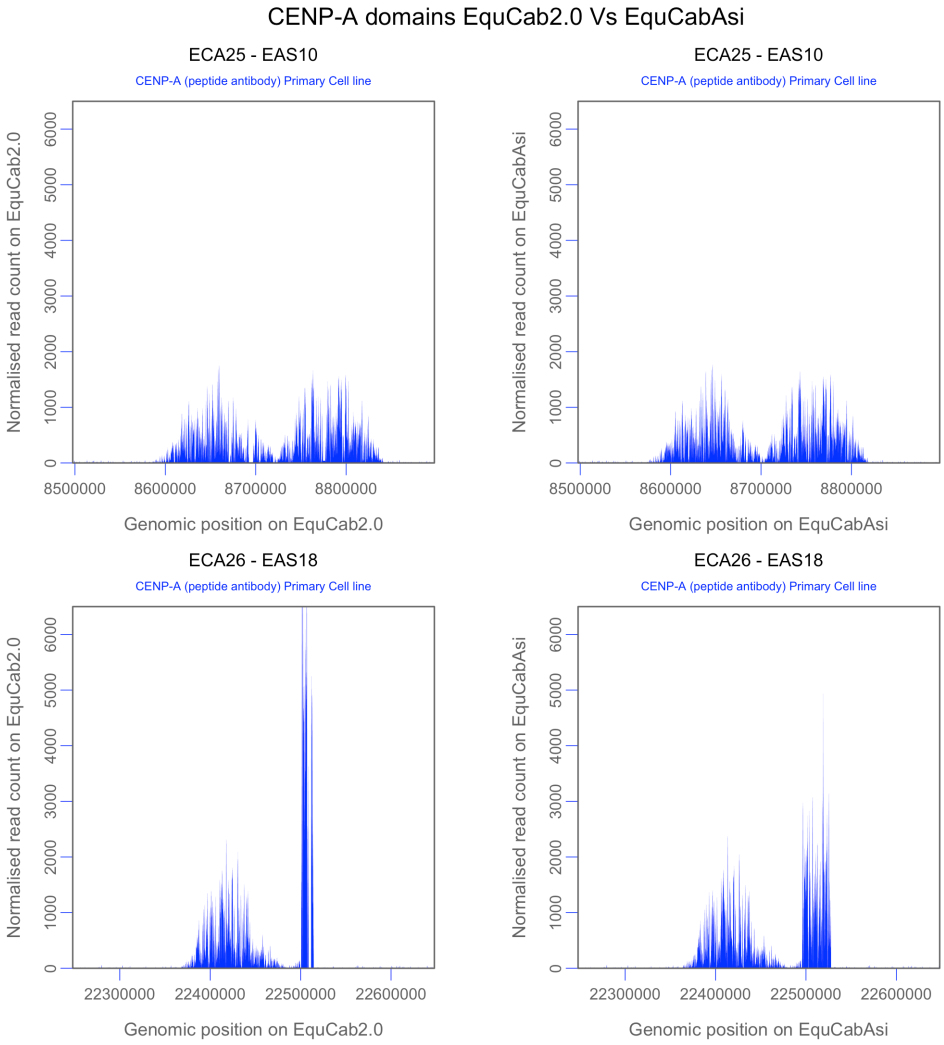


Figure 3.10-F CENP-A domains – hybrid genome comparison
ECA25/EAS10 exhibits a multi-domain structure with some noticeable minor gaps in *EquCab2.0*. These minor gaps are covered in *EquCabAsi*. ECA26/EAS18 profiles a multi-domain structure with one Gaussian-like distribution and a spike distribution. The spike peak is present in three copies in *EquCabAsi*.

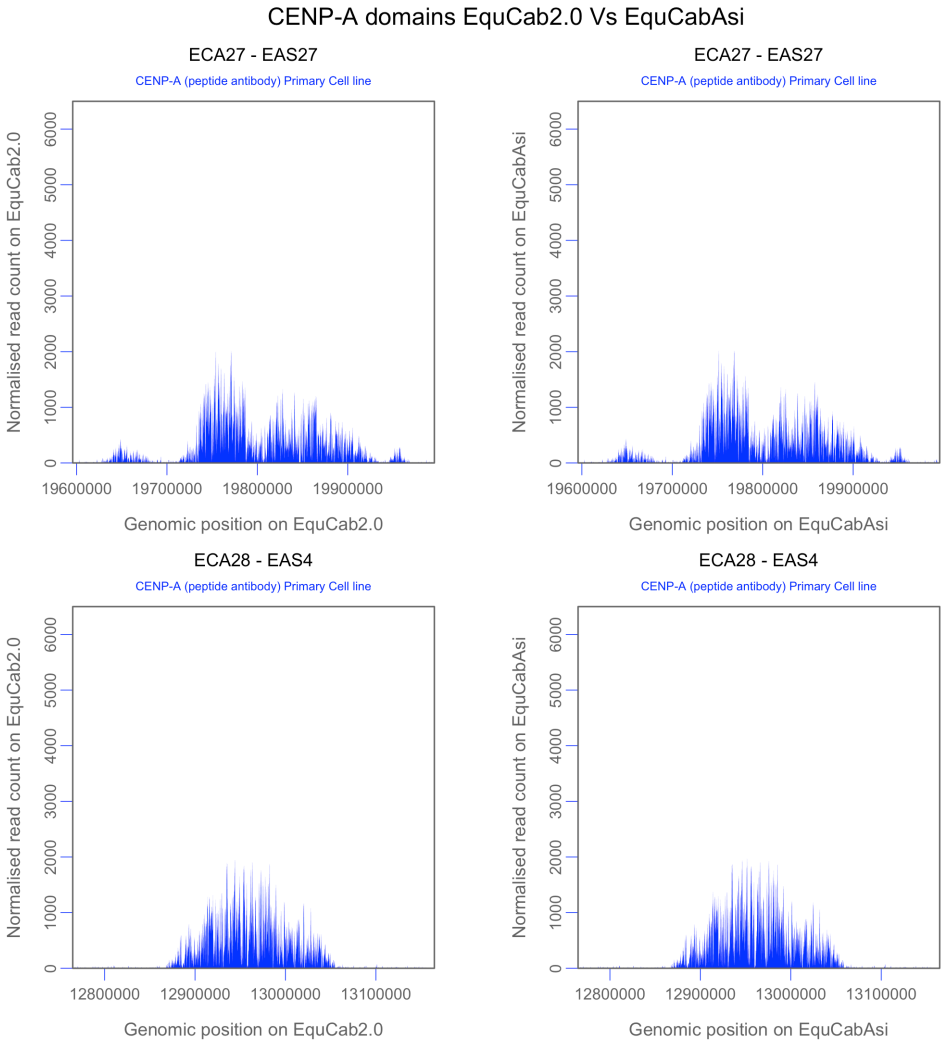


Figure 3.10-G CENP-A domains – hybrid genome comparison
ECA27/EAS27 exhibits a large domain that resembles an unresolved multi-domain structure. Very minor differences are seen when comparing alignments in both genomes. ECA28/EAS4 shows a conserved profile in both alignments.

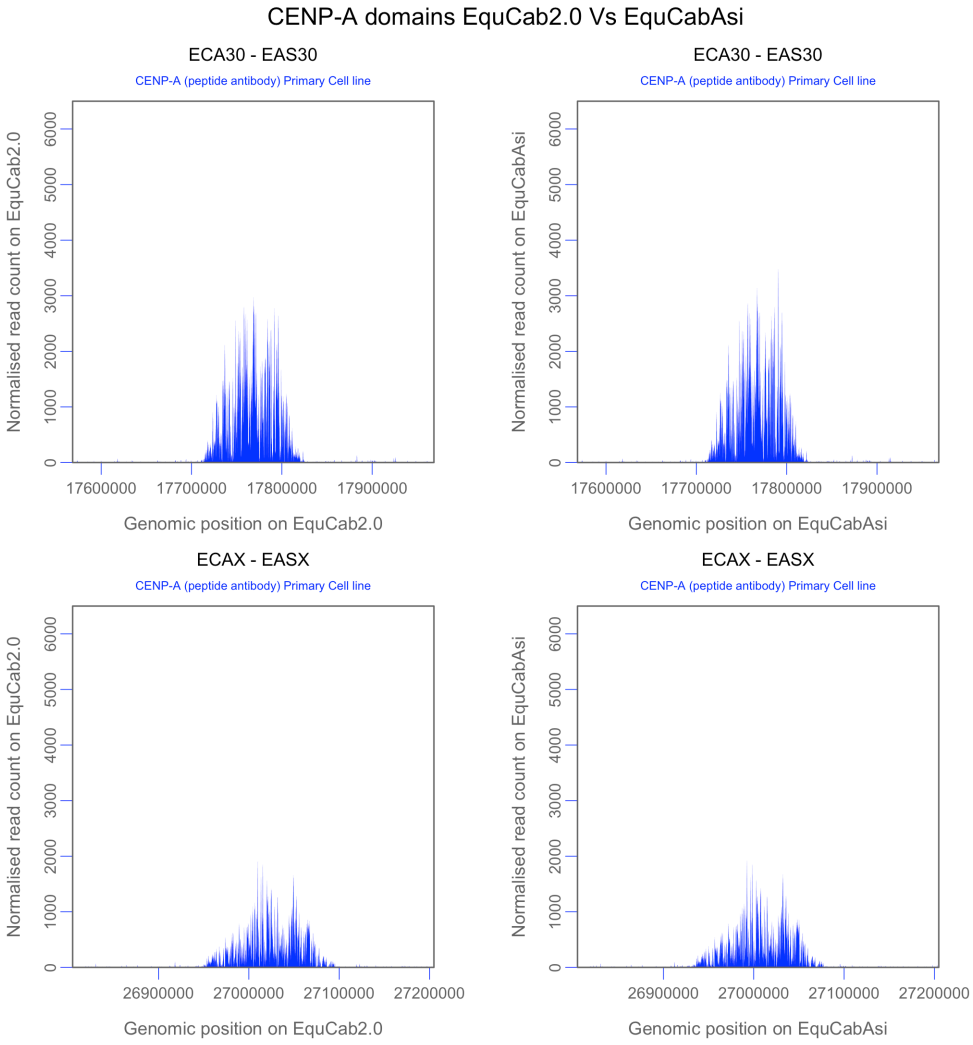


Figure 3.10-H CENP-A domains – hybrid genome comparison
ECA30/EAS30 and ECAX/EASX show very conserved CENP-A distributions in both alignments.

3.3.3 Structure of *E.asinus* CENP-A domains – positional alleles

In the centromere of horse chromosome 11, which was previously defined, a number of individuals were seen to possess two peaks of CENP-A binding. Through a SNP based analysis these peaks were shown to be centromere alleles where the centromere in each chromosome homologue occupied a slightly different position within a ~550 kb domain (Purgato et al., 2015; Wade et al., 2009). From this previous analysis (Purgato et al., 2015), it was expected that the multi-domain peaks corresponded to centromere positional epialleles. In similar analyses of these ChIP-seq data in the Giulotto lab, it was found that the multi-domain CENP-A profiles corresponded to centromere positional epialleles. Specifically, for those centromeres that exhibited two distinct Gaussian-like domains, each domain corresponded to the centromere of a different homologue. Centromeres that were not individually separated or appeared as a single domain along a genomic region were shown to be individual CENP-A alleles that were overlapping in genomic coordinates.

3.3.4 Inter-individual variation in centromere position examined by ChIP-seq

In order to examine variation of centromere position in unrelated individuals, skin fibroblasts from a second donkey "Blackjack" (BJ) were processed by ChIP-seq. The ChIP-seq reads were treated as before and originally aligned to the horse reference genome – *EquCab2.0*. All 16 chromosomes identified previously as satellite-free CENP-A domains showed similar CENP-A binding domains in comparable positions but exact locations differed from "Asino Nuovo"/*EquCab2.0,* at several centromeres. An example of this is shown in Figure 3.11, were the centromere region of ECA8/EAS7 in AN is a single domain structure and the corresponding centromere in BJ is a multi-domain structure. These data show that centromere sliding, originally observed in the horse, also occurs in the donkey. (All donkey comparison figures can be seen in Appendix I-Figure 8.3-Figure 8.9)

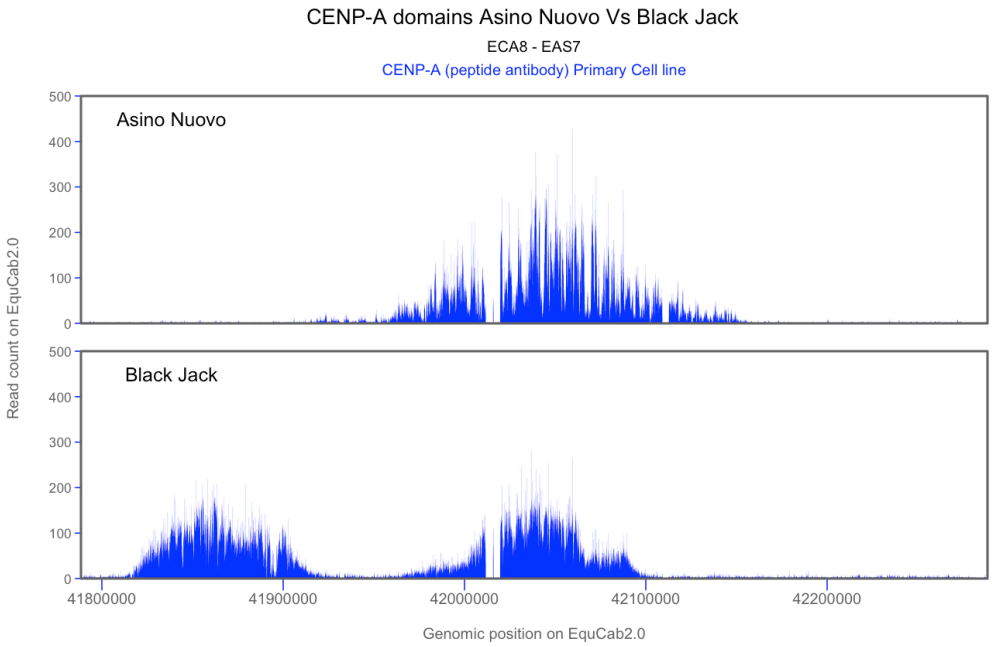


Figure 3.11–A CENP-A positional variation in unrelated individuals.
CENP-A domains are localised in different positions in two unrelated donkey individuals. ChIP-seq data from AN (top) and BJ (bottom) were aligned to *EquCab2.0*.

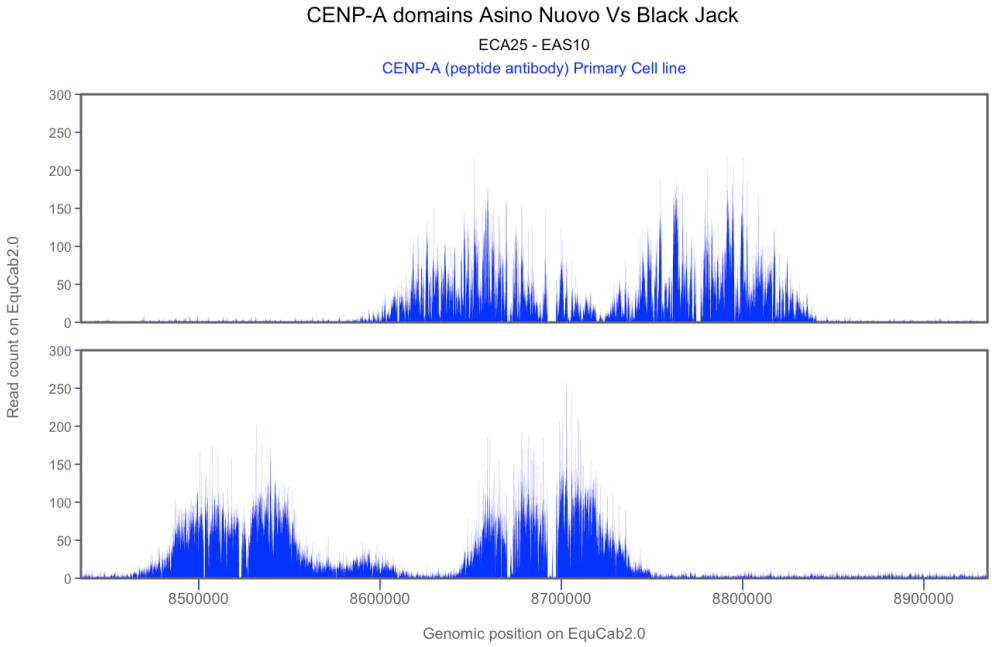


Figure 3.12 CENP-A positional variation in unrelated individuals.
CENP-A domains are localised in different positions in two unrelated donkey individuals. ChIP-seq data from AN (top) and BJ (bottom) were aligned to *EquCab2.0*.

3.3.5 Construction of Blackjack specific hybrid genome

The alternative CENP-A footprint in BJ prompted the assembly of DNA sequences from ChIP-seq data, analogous to that described for AN. A hybrid genome was constructed called "*EquCabAsi-Blackjack*". The coordinates of *EquCab2.0,* were replaced and newly refined newly refined "Blackjack" centromere sequences were inserted (Table 3.5). The locations and span of each BJ centromere are detailed in Table 3.6.

Table 3.5 EquCabAsi – "Blackjack" insert coordinates

Chromosome		Start	End
EAS4	ECA28	12,856,811	13,469,980
EAS5	ECA19	4,550,152	5,145,128
EAS7	ECA8	41,817,689	42,121,179
EAS8	ECA20	26,426,414	26,525,669
EAS9	ECA14	29,649,345	29,705,407
EAS10	ECA25	8,466,521	8,756,214
EAS11	ECA17	16,757,631	16,877,565
EAS12	ECA9	31,584,308	32,142,731
EAS13	ECA11	46,556,884	46,849,905
EAS14	ECA13	7,086,710	7,470,346
EAS16	ECA5	74,879,061	75,027,728
EAS18	ECA26	22,276,704	22,535,514
EAS19	ECA6	14,177,149	14312,673
EAS27	ECA27	19,710,712	19,915,943
EAS30	ECA30	17,636,535	17,835,931
EASX	ECAX	26,942,649	27,073,638

Table 3.6 Centromere locations and span - Blackjack

E.ca Chr	E.as Chr	Peak 1 Start : End		Span (kb)	Peak 2 Start : End		Span (kb)
5	16	74,873,953	74,934,671	60.7	-	-	-
6	19	14,196,664	14,331,529	134.8	-	-	-
8	7	41,798,262	41,932,073	133.8	41,939,158	42,091,468	152.3
9	12	31,567,773	31,728,407	160.6	31,976,192	32,106,220	130
11	13	46,538,312	46,661,580	123.2	46,688,576	46,767,361	78.7
13	14	7,090,371	7,187,946	97.5	7,236,734	7,355,199	118.4
14	9	29,551,886	29,582,767	30.8	-	-	-
17	11	16,754,379	16,867,227	112.8	-	-	-
19	5	4,544,394	4,669,514	125.1	4,933,407	5,063,934	130.5
20	8	26,417,453	26,509,636	92.1	-	-	-
25	10	8,455,349	8,632,866	177.5	8,623,529	8,775,312	151.7
26	18	22,279,231	22,404,532	125.3	22,494,925	22,528,671	33.7
27	27	19,731,572	19,926,402	194.8	-	-	-
28	4	12,901,848	13,000,193	98.3	13,221,140	13,371,835	152.3
30	30	17,643,181	17,810,449	167.2	-	-	-
X	X	26,936,344	27,073,478	137.1	-	-	-

3.3.6 Stability and transmission of centromeres across generations

A central question regarding the sliding or spreading of satellite-free centromeres across a genomic region is how or when this sliding occurs. A study was designed to investigate centromere localisation in the transmission of centromeres across generations. Sperm from the donkey BJ were used in *in vitro* fertilisation with three unrelated horses to produce hybrid mule embryos from which mule cultures were derived (D. Antczak, Ithaca, NY, USA, E. Giulotto). The hybrid mule cells generated, contained two haploid genomes, meaning single centromere alleles can be investigated in the progeny cell lines. Three conceptus mule cell lines (c1009, c1010 & c1011) were processed for ChIP-seq using the CENP-A antibody. The ChIP-seq reads were processed as before and then aligned to the *EquCabAsi-Blackjack* hybrid genome (Figure 3.13). These data display BJ on the top panel and each of the three concepti underneath.

One expectation is that for the BJ epialleles that are in different positions, the concepti offspring should show one of those alleles transmitted. This can be most clearly seen in the centromeres were the epialleles are discretely separated. An example of this can be seen below in ECA9/EAS12 (Figure 3.13-A), where two CENP-A domains are present in Blackjack. In this example the right most allele is observed in Conceptus1009 and Conceptus1011 but in Conceptus1010 the left most allele is transmitted. The data clearly show that the centromere epialleles in the paternal donkey (BJ) are independently assorted during transmission to mule offspring. ECA9/EAS12 shows a shift in position to the left from the leftmost allele but interestingly the CENP-A domain looks to have split into two domains when compared to its parent allele (Figure 3.13-A).

For most of the transmission events, centromeres alleles are transmitted with little movement/sliding in their position. Based on direct inspection the ECA17/EAS11 centromere has a very stable transmission pattern (Figure 3.13-B). Based on the same approach, the spike centromeres are all stably transmitted (Figure 3.13-C,D,E). Some of the centromere alleles, however, show changes in position. For example, ECA8/EAS7 shows a rightward shift in peak position in c1010, which shows that CENP-A binding is in a mid-region between the two parent alleles (Figure 3.13-F).

Collectively, these data show transmission of CENP-A domains from parents to offspring, using a ChIP-seq approach. We have shown that centromere positional variation is a property of centromeres. This suggests that sliding can take place, leading to the observed positional variation (Figure 3.13).

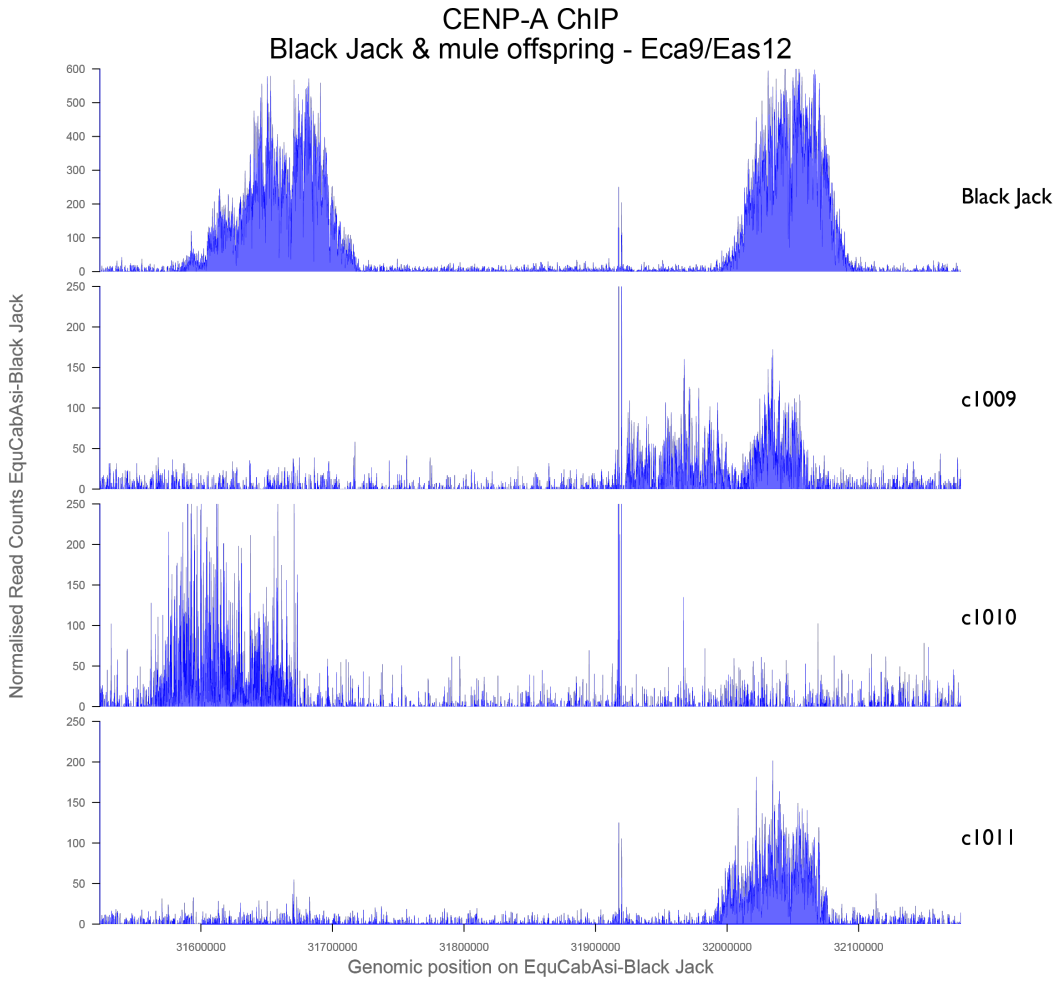


Figure 3.13 Blackjack and concepti CENP-A ChIP-seq
CENP-A ChIP-seq on Blackjack family – ECA9/EAS12 mapped back to *EquCabAsi – Blackjack*. The plot shows a varying transmission of CENP-A binding across the concepti offspring. The Blackjack CENP-A domain shows clear discrete alleles. The c1009 mule displays transmission of the rightmost allele, which has spread leftward. The c1010 mules transmission suggests the leftmost allele was passed down but a leftward shift of the CENP-A domain has taken place. The c1011 mule exhibits transmission of the rightmost allele.

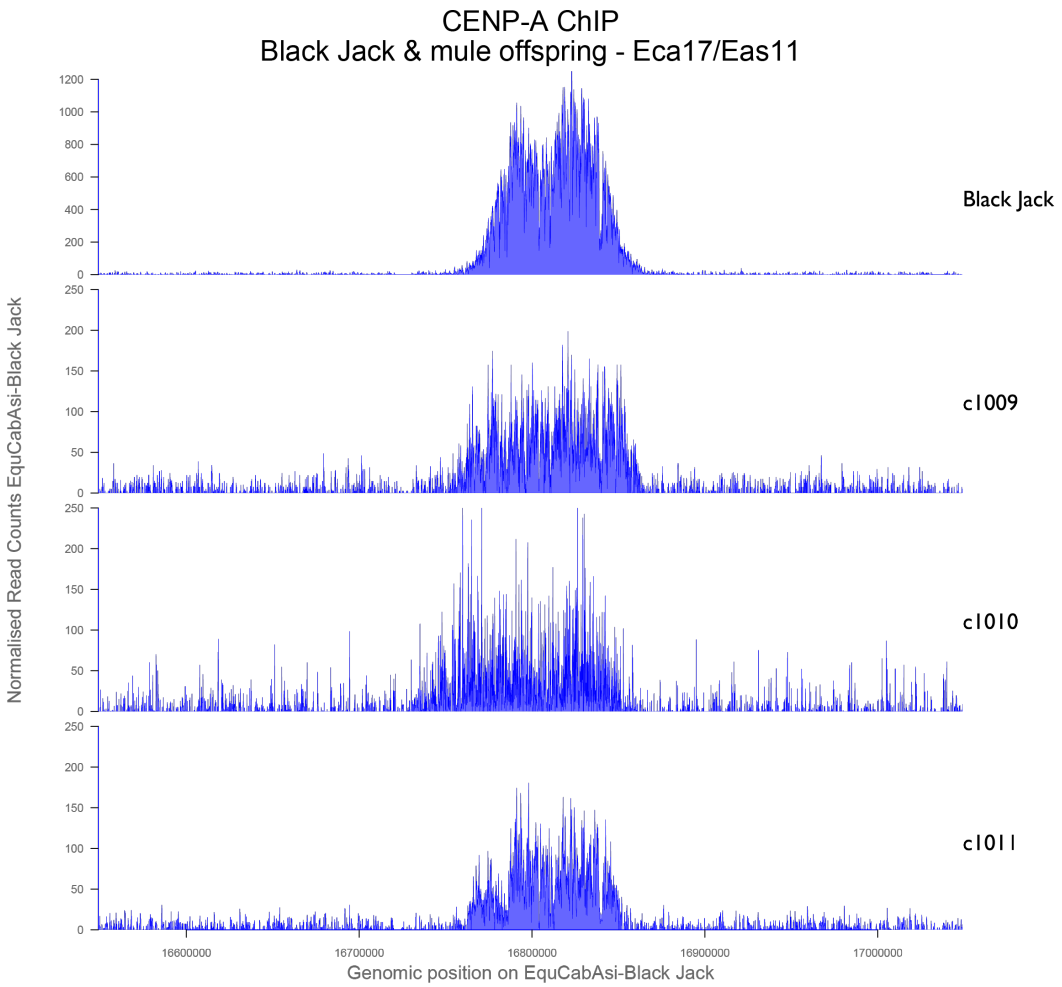


Figure 3.13-B - Blackjack and concepti CENP-A ChIP-seq
CENP-A ChIP-seq on Blackjack family – ECA17/EAS11 mapped back to *EquCabAsi* – *Blackjack*. The plot displays a more stable transmission of CENP-A binding across the concepti offspring. The Blackjack CENP-A domain shows no clear discrete alleles. The c1009 mule shows direct transmission of one of the Blackjack alleles. The c1010 mule shows transmission of the parent allele but the CENP-A domain appears to have taken a leftward shift. The c1011 mule displays a stable transmission.

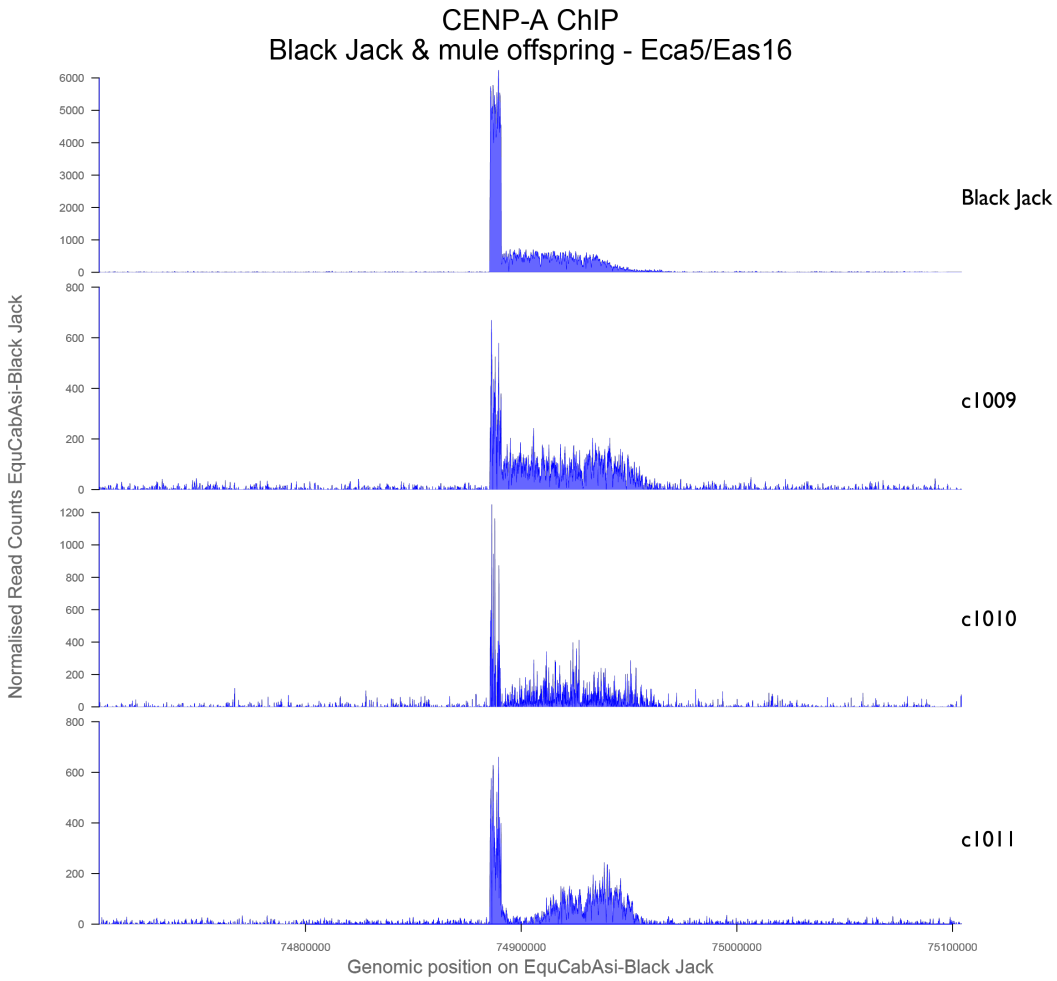


Figure 3.13-C – Blackjack and concepti CENP-A ChIP-seq
CENP-A ChIP-seq on Blackjack family – ECA5/EAS16 mapped back to *EquCabAsi* – *Blackjack*. The plot shows a highly fixed transmission of CENP-A binding across the concepti offspring.

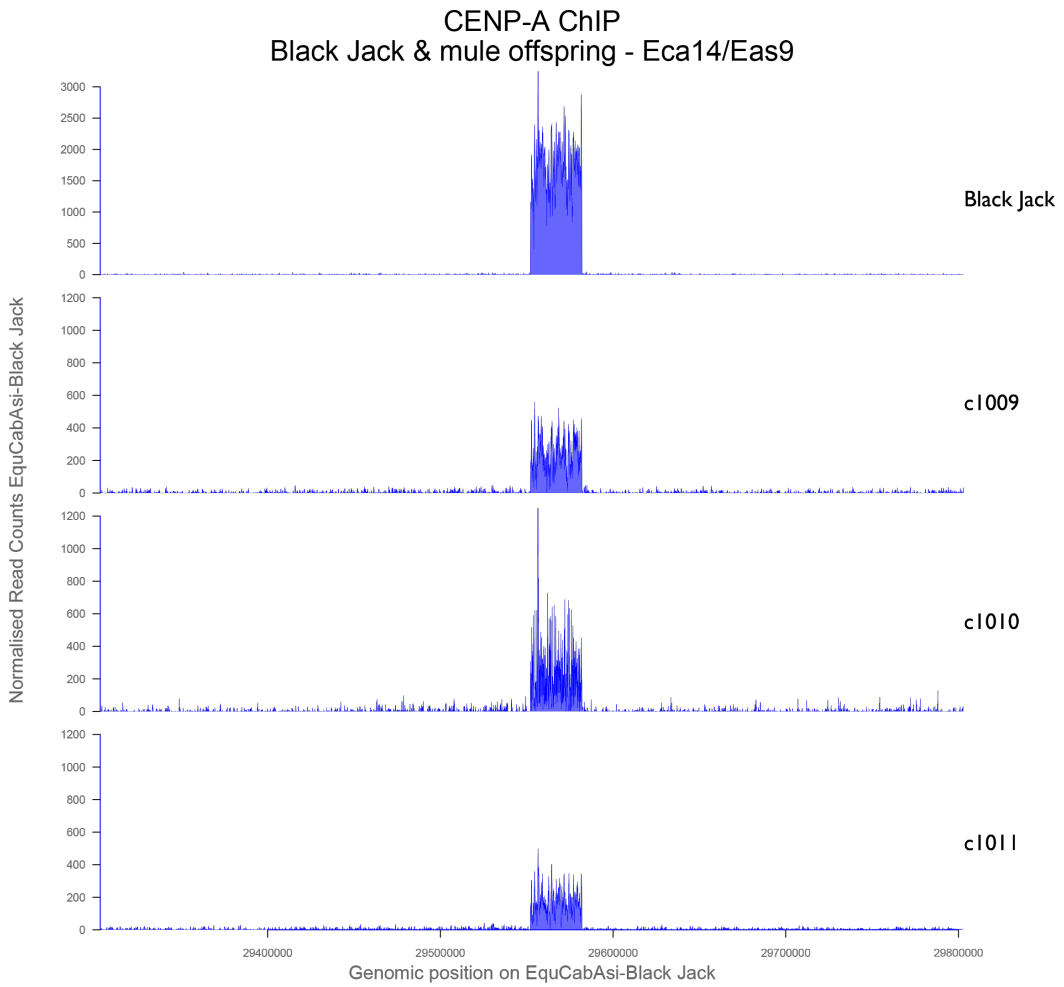


Figure 3.13-D - Blackjack and concepti CENP-A ChIP-seq
CENP-A ChIP-seq on Blackjack family – ECA14/EAS9 mapped back to *EquCabAsi* – *Blackjack*. The plot shows a highly fixed transmission of CENP-A binding across the concepti offspring.

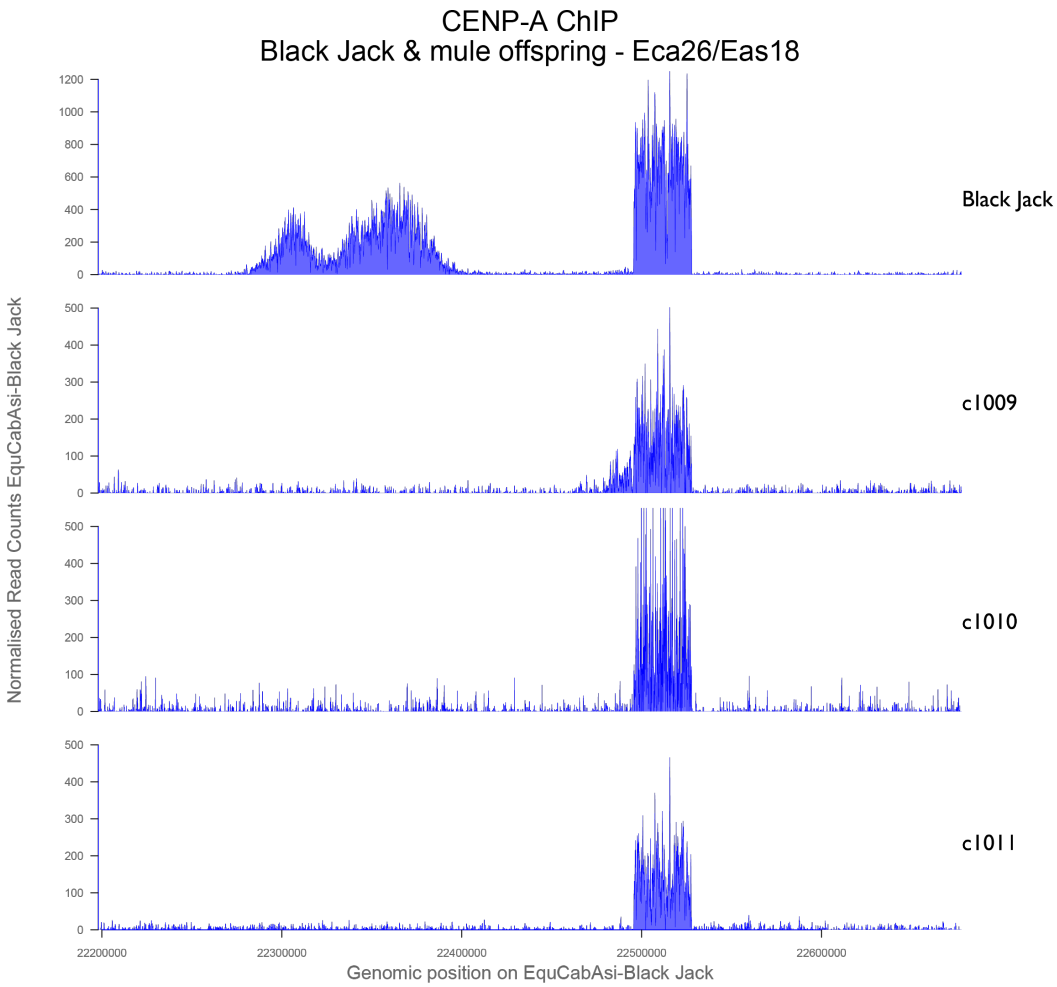


Figure 3.13-E - Blackjack and concepti CENP-A ChIP-seq
CENP-A ChIP-seq on Blackjack family – ECA26/EAS18 mapped back to *EquCabAsi* – *Blackjack*. The plot shows a varying transmission of CENP-A binding across the concepti offspring. The Blackjack CENP-A domain shows clear discrete CENP-A alleles: the leftmost displaying a Gaussian-like distribution and the rightmost highly fixed and spike-like. All concepti display stable transmission of the leftmost CENP-A allele

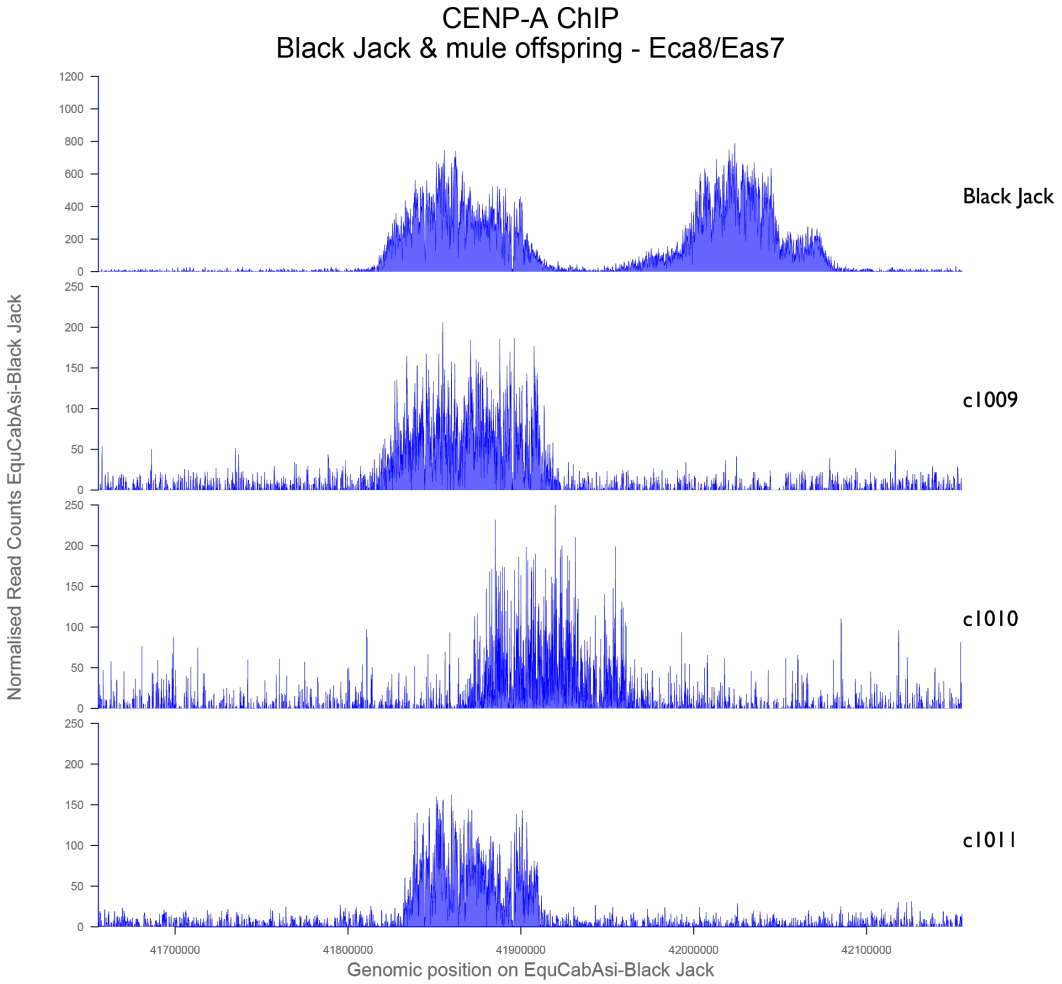


Figure 3.13-F - Blackjack and concepti CENP-A ChIP-seq
CENP-A ChIP-seq on Blackjack family – ECA8/EAS7 mapped back to *EquCabAsi – Blackjack*. The plot shows a varying transmission of CENP-A binding across the concepti offspring. The Blackjack CENP-A domain shows clear discrete alleles. The c1009 mule displays inheritance of the leftmost allele. The c1010 mules transmission suggests the leftmost allele was passed down but a rightward shift of the CENP-A domain has taken place. The c1011 mule exhibits transmission of the leftmost allele that occupies a slightly smaller span.

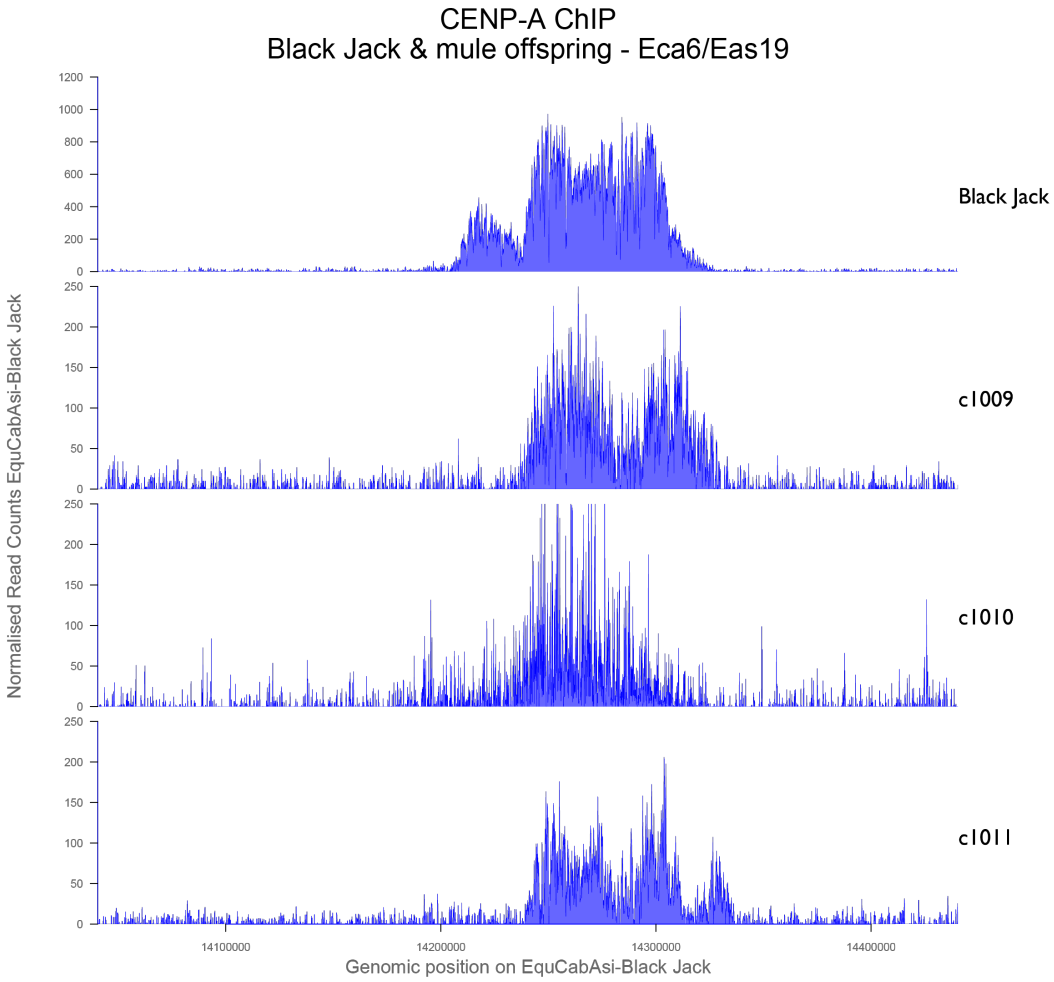


Figure 3.13-G - Blackjack and concepti CENP-A ChIP-seq
CENP-A ChIP-seq on Blackjack family – ECA6/EAS19 mapped back to *EquCabAsi – Blackjack*. The plot shows a varying transmission of CENP-A binding across the concepti offspring. The Blackjack CENP-A domain shows no clear discrete alleles. The c1009 mule displays a bipartite distribution. The c1010 mule shows transmission exclusively from the central region of Blackjack. The c1011 mule exhibits nearly a tripartite distribution indicating centromere movement through transmission.

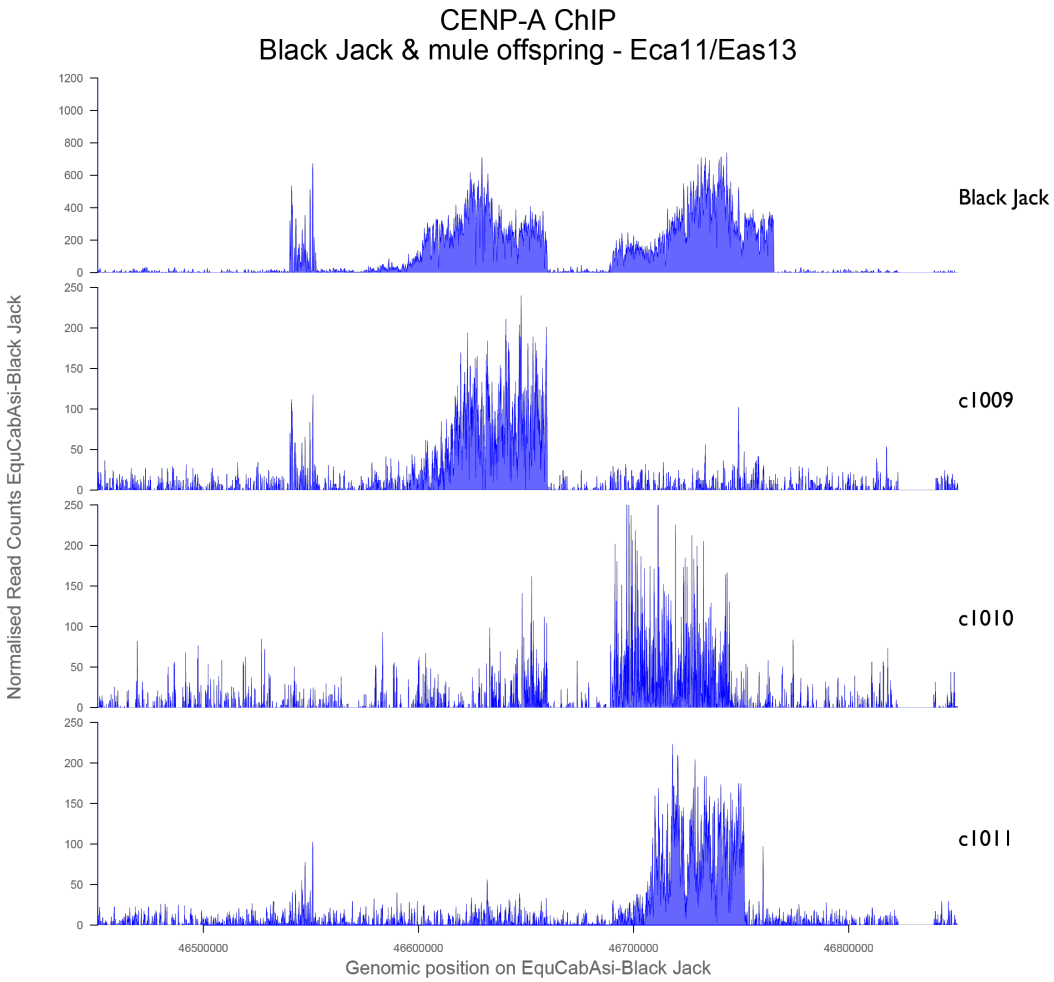


Figure 3.13-H - Blackjack and concepti CENP-A ChIP-seq
CENP-A ChIP-seq on Blackjack family – ECA11/EAS13 mapped back to *EquCabAsi – Blackjack*. The plot shows a varying transmission of CENP-A binding across the concepti offspring. The Blackjack CENP-A domain shows clear discrete alleles. The c1009 mule displays transmission of the leftmost allele. The c1010 mules transmission suggests the rightmost allele was passed down. The c1011 mule displays transmission of the rightmost allele.

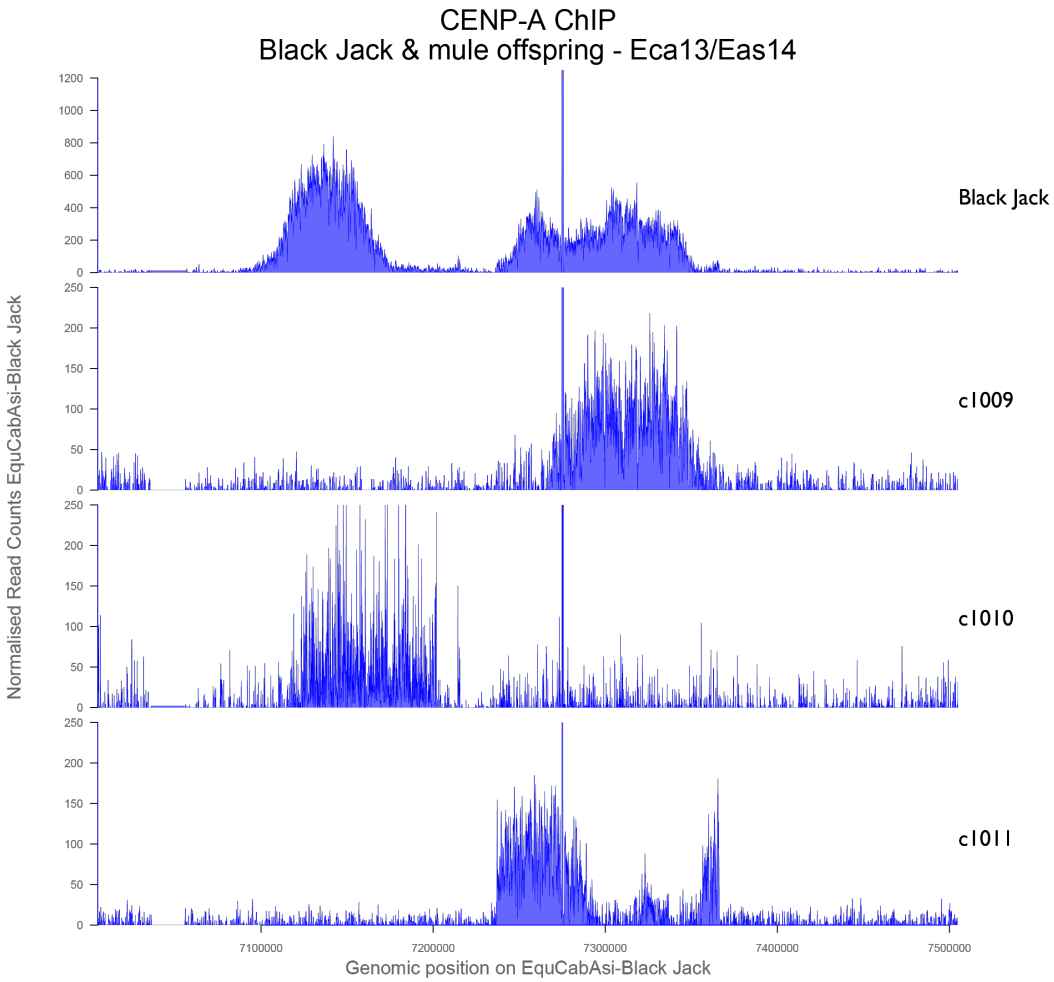


Figure 3.13-I - Blackjack and concepti CENP-A ChIP-seq
CENP-A ChIP-seq on Blackjack family – ECA13/EAS14 mapped back to *EquCabAsi – Blackjack*. The plot shows a varying transmission of CENP-A binding across the concepti offspring. The Blackjack CENP-A domain shows clear discrete alleles. The c1009 mule displays transmission of the rightmost allele that occupies a slightly smaller CENP-A domain. The c1010 mule shows transmission of the leftmost allele that has taken a slightly rightward shift. The c1011 mule exhibits a complex transmission of the rightmost allele in which shows a depleted CENP-A signal central and right.

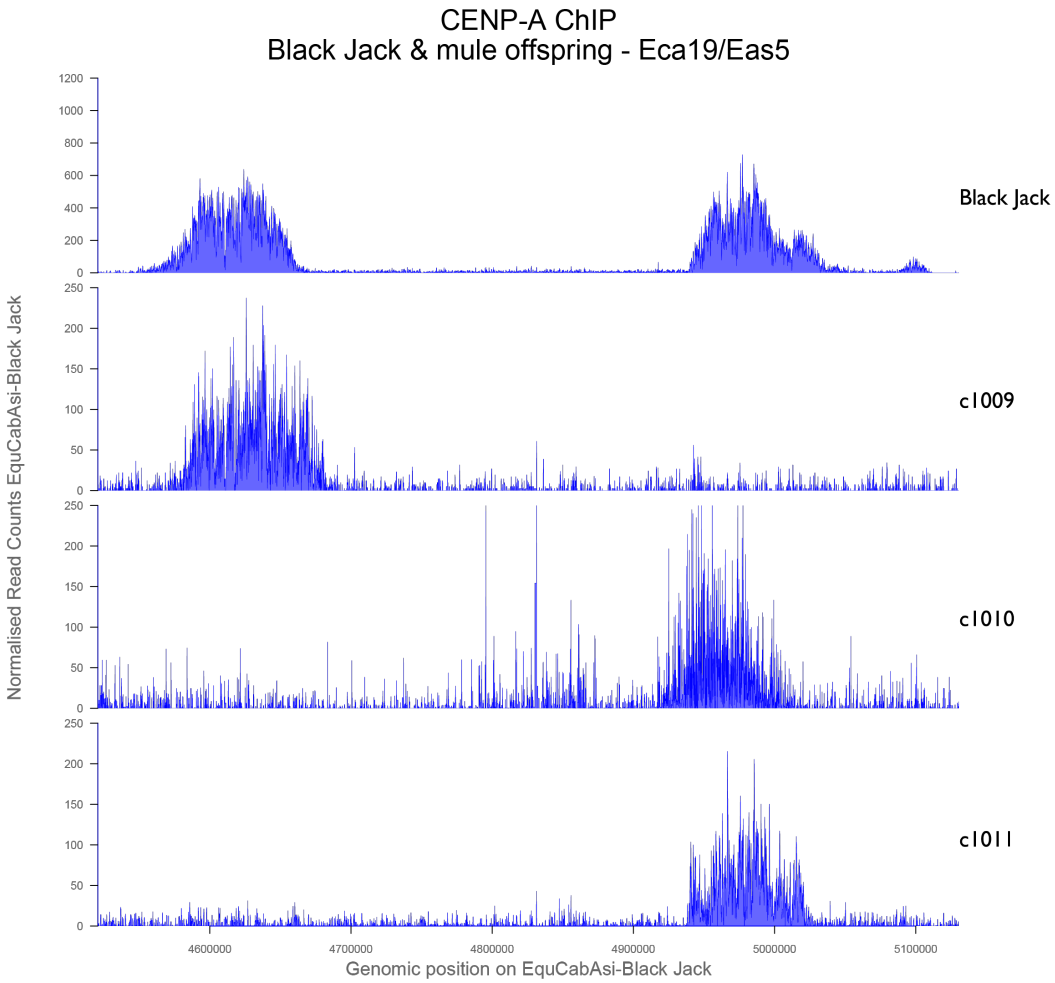


Figure 3.13-J - Blackjack and concepti CENP-A ChIP-seq
CENP-A ChIP-seq on Blackjack family – ECA19/EAS5 mapped back to *EquCabAsi* – *Blackjack*. The plot shows a varying transmission of CENP-A binding across the concepti offspring. The Blackjack CENP-A domain shows clear discrete alleles. The c1009 mule shows transmission of the leftmost allele and appears to exhibit a slight rightward shift. The c1010 mule shows transmission of the rightmost allele and the CENP-A domain appears to have taken a slight leftward shift. The c1011 mule displays a stable transmission of the rightmost allele.

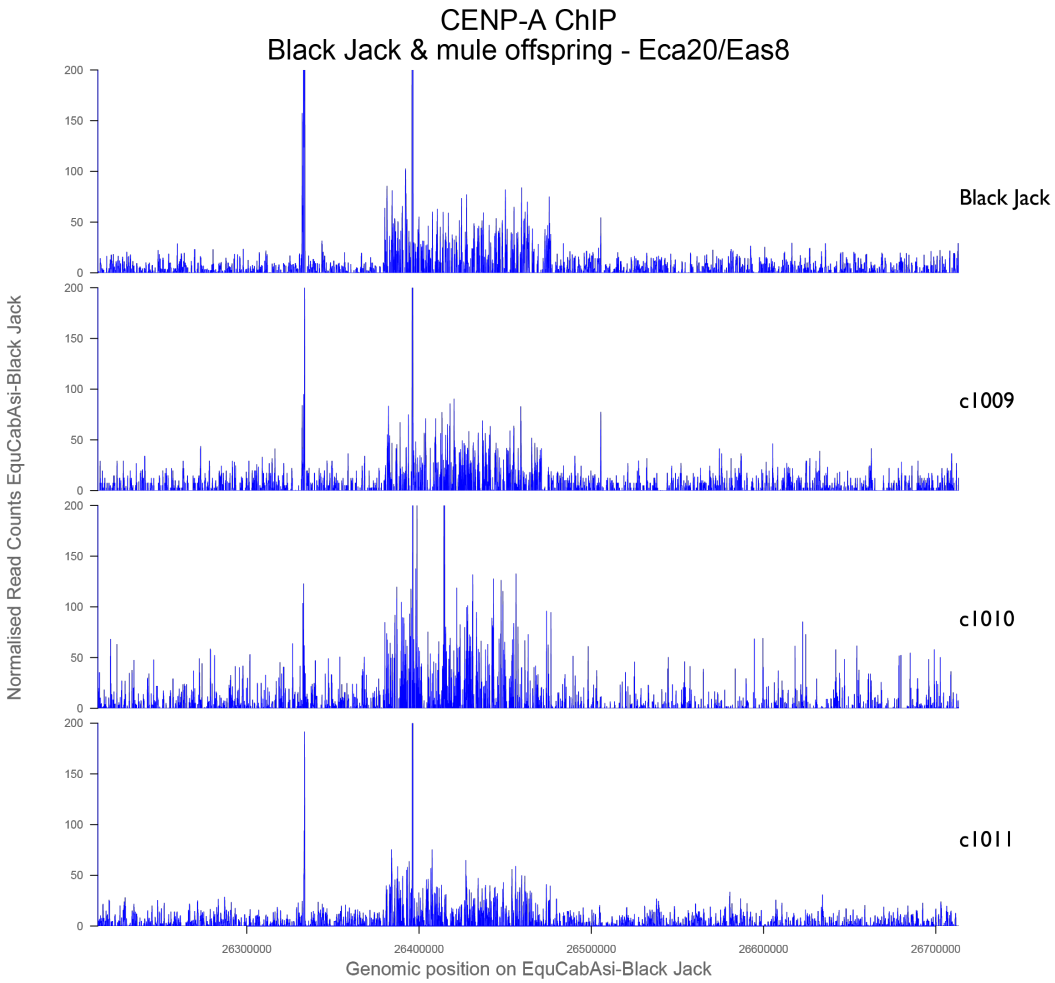


Figure 3.13-K- Blackjack and concepti CENP-A ChIP-seq
CENP-A ChIP-seq on Blackjack family – ECA20/EAS8 mapped back to *EquCabAsi* – *Blackjack*. The plot shows a stable transmission of CENP-A binding across the concepti offspring. The signal across all four individuals is of complex nature.

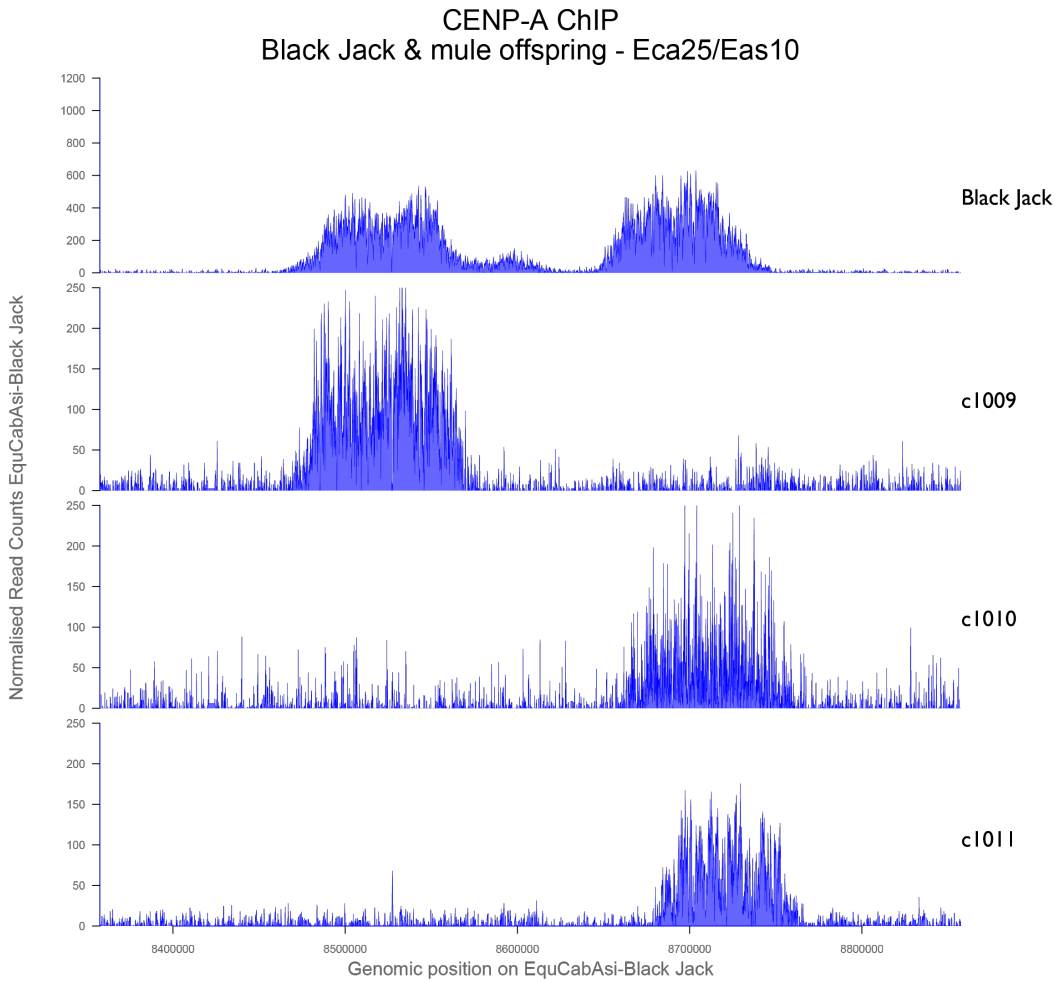


Figure 3.13-L - Blackjack and concepti CENP-A ChIP-seq
CENP-A ChIP-seq on Blackjack family – ECA25/EAS10 mapped back to *EquCabAsi – Blackjack*. The plot shows a varying transmission of CENP-A binding across the concepti offspring. The Blackjack CENP-A domain shows clear discrete alleles. The c1009 mule shows a stable transmission of the leftmost allele with. The c1010 mule shows transmission of the rightmost allele and the CENP-A domain appears to have taken a slight rightward shift. The c1011 mule displays a transmission of the rightmost allele that has taken a rightward shift.

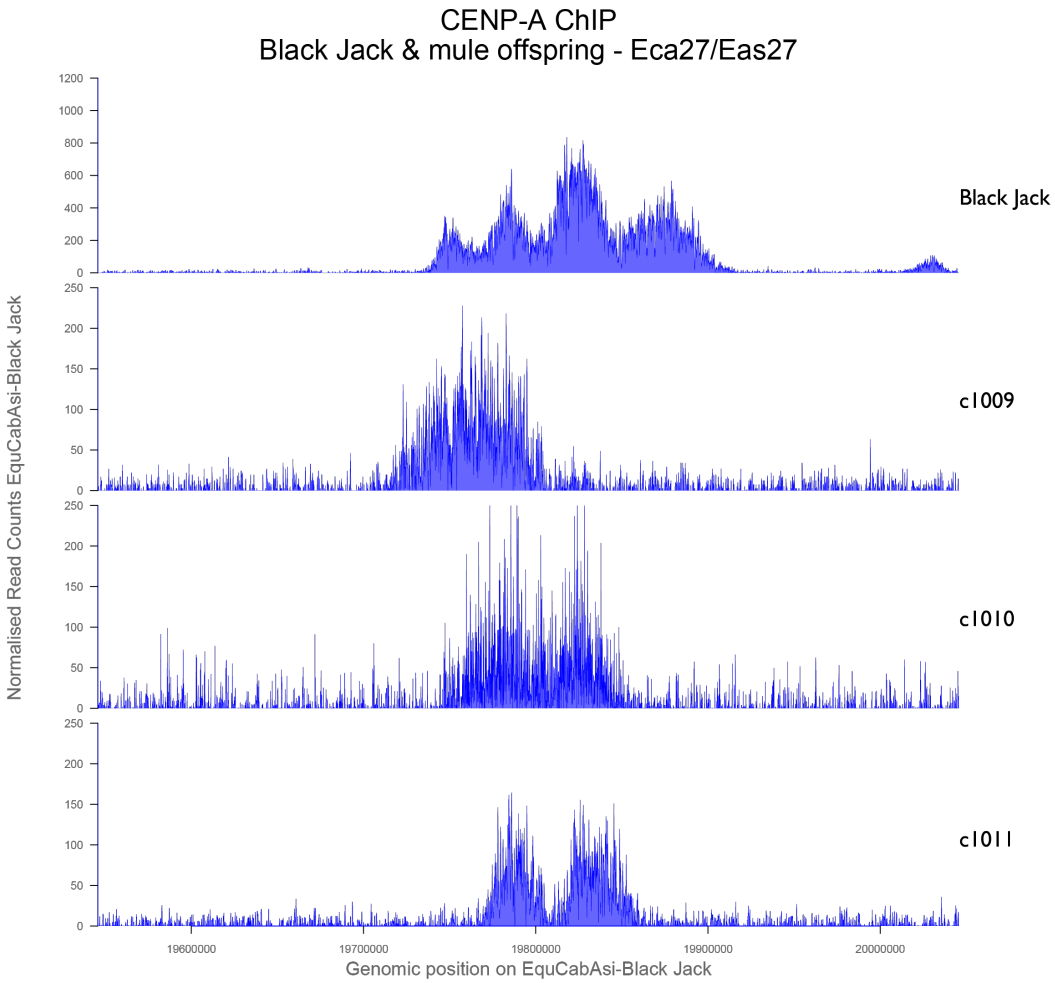


Figure 3.13-M - Blackjack and concepti CENP-A ChIP-seq
CENP-A ChIP-seq on Blackjack family – ECA27/EAS27 mapped back to *EquCabAsi – Blackjack*. The plot shows a varying transmission of CENP-A binding across the concepti offspring. The Blackjack CENP-A domain shows no clear discrete alleles but rather four subdomains of CENP-A binding. The c1009 mule shows a transmission of an allele that takes a leftward position. The c1010 mule shows transmission an allele that is taken a rightward shift compared to c1009. The c1011 mule displays a transmission an allele and has taken a bipartite distribution. None of the concepti display transmission of the rightmost subdomain.

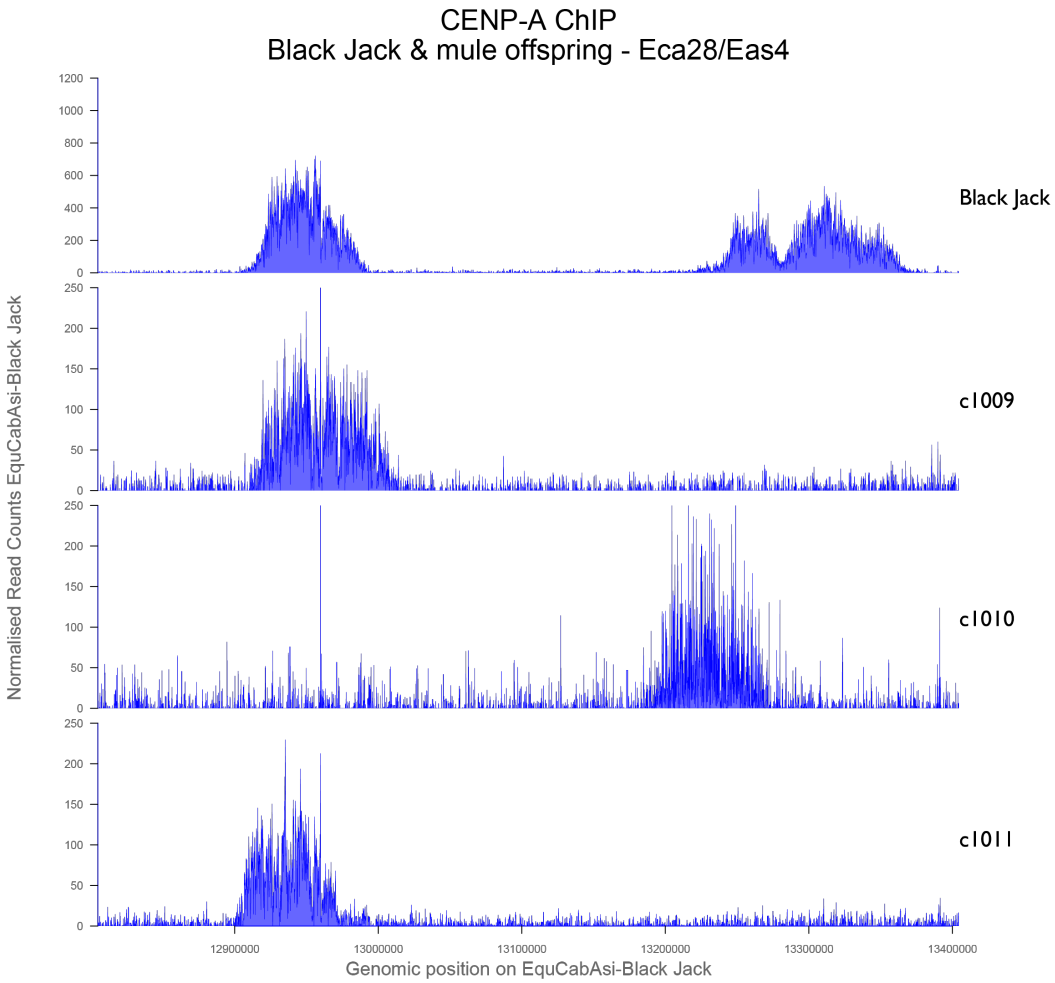


Figure 3.13-N - Blackjack and concepti CENP-A ChIP-seq
CENP-A ChIP-seq on Blackjack family – ECA28/EAS4 mapped back to *EquCabAsi – Blackjack*. The plot shows a varying transmission of CENP-A binding across the concepti offspring. The Blackjack CENP-A domain shows clear discrete CENP-A alleles. The c1009 mule shows a transmission leftmost allele with a slight expansion of CENP-A signal to the right. The c1010 mule shows transmission of the rightmost allele that is taken a clear leftward shift. The c1011 mule displays a stable transmission of the leftmost allele.

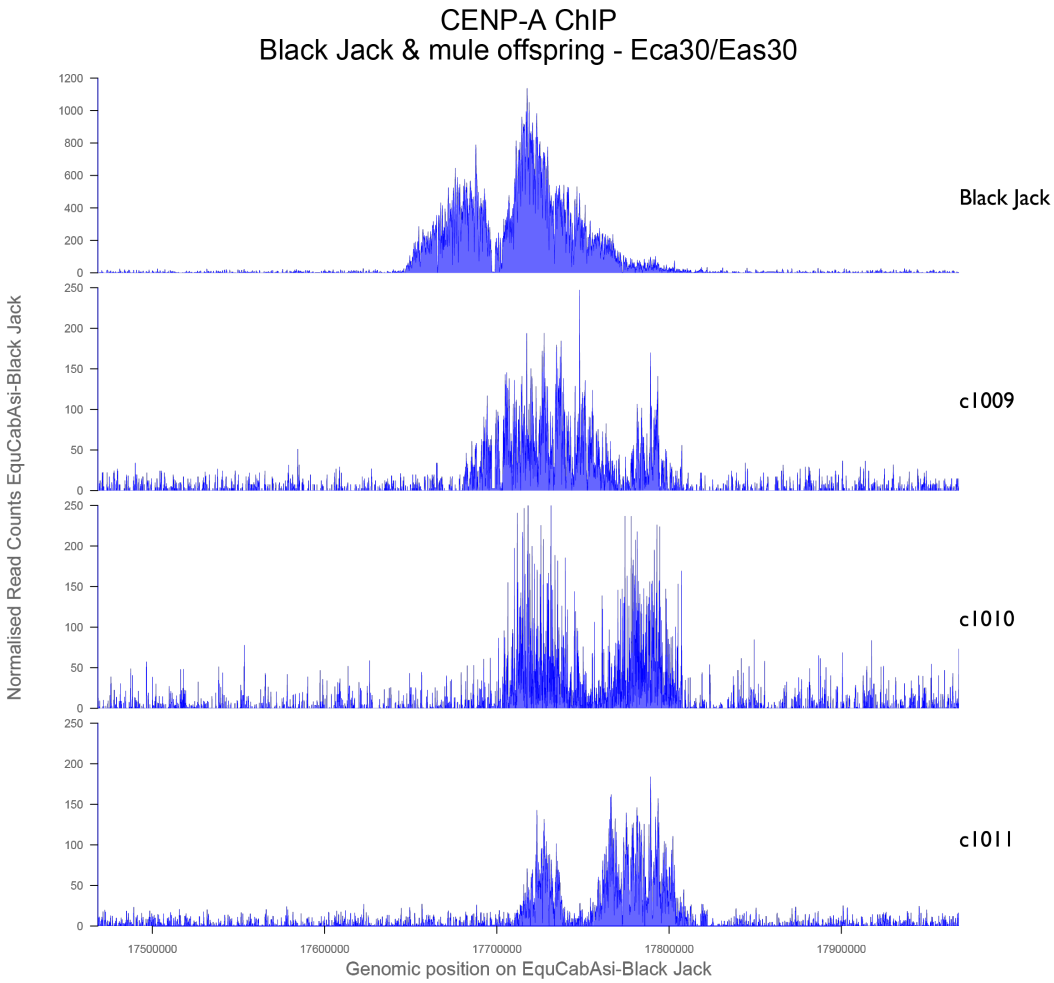


Figure 3.13-O - Blackjack and concepti CENP-A ChIP-seq
CENP-A ChIP-seq on Blackjack family – ECA30/EAS30 mapped back to *EquCabAsi – Blackjack*. The plot shows a varying transmission of CENP-A binding across the concepti offspring. The Blackjack CENP-A domain shows no clear discrete CENP-A alleles. The c1009 mule shows a transmission of an allele with a slight expansion of CENP-A signal to the right. The c1010 mule shows transmission of an allele that that has taken a bipartite distribution. The c1011 mule displays transmission of an allele that that has taken a bipartite distribution.

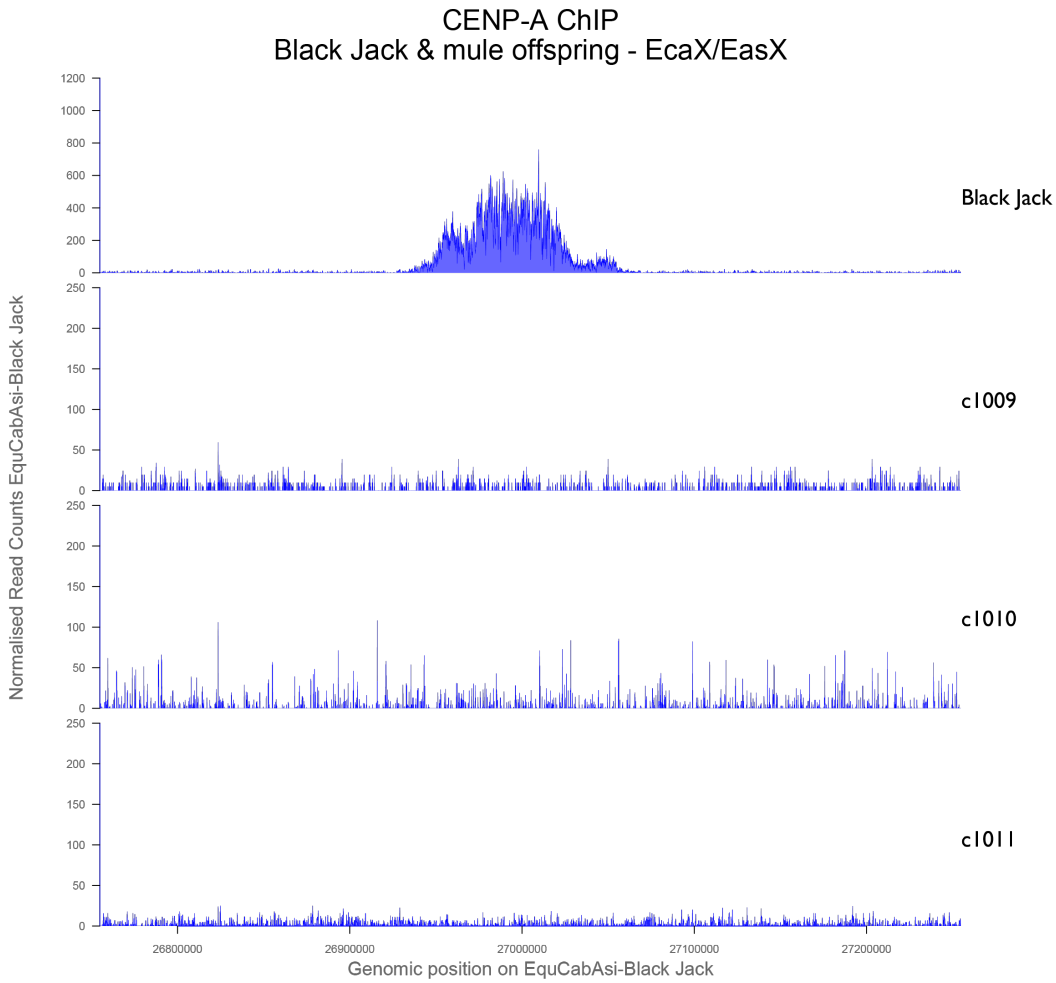


Figure 3.13-P - Blackjack and concepti CENP-A ChIP-seq
CENP-A ChIP-seq on Blackjack family – ECAX/EASX mapped back to *EquCabAsi* – *Blackjack*. The plot shows no transmission of CENP-A binding across the concepti offspring. The Blackjack CENP-A domain on the X chromosome is haploid as the parent donkey Blackjack is a male.

3.3.7 Analysis of centromere sliding between generations

In order to examine centromere sliding quantitatively, a metric was designed to calculate the range of sliding through transmission from parents to offspring. The Blackjack family ChIP-seq data (Blackjack, c1009, c1010 & c1011) were used in this experiment. The centromeres were characterised using read count distribution by identifying the median position within the distribution (50% of the read count to the left : 50% to the right) across each centromere domain in the parent and F1 progeny. The locations of the median positions on the parental centromeres were taken as a reference and the absolute difference, or as termed here "displacement", in positional median values for each centromere in each of the mule offspring was calculated. A graphical representation of the median analysis can be seen in Figure 3.14. Also shown are the points at which 5% or 95% of the integrated signal is located (from left to right). (All sliding figures can be seen in Appendix I-Figure 8.10-Figure 8.19)

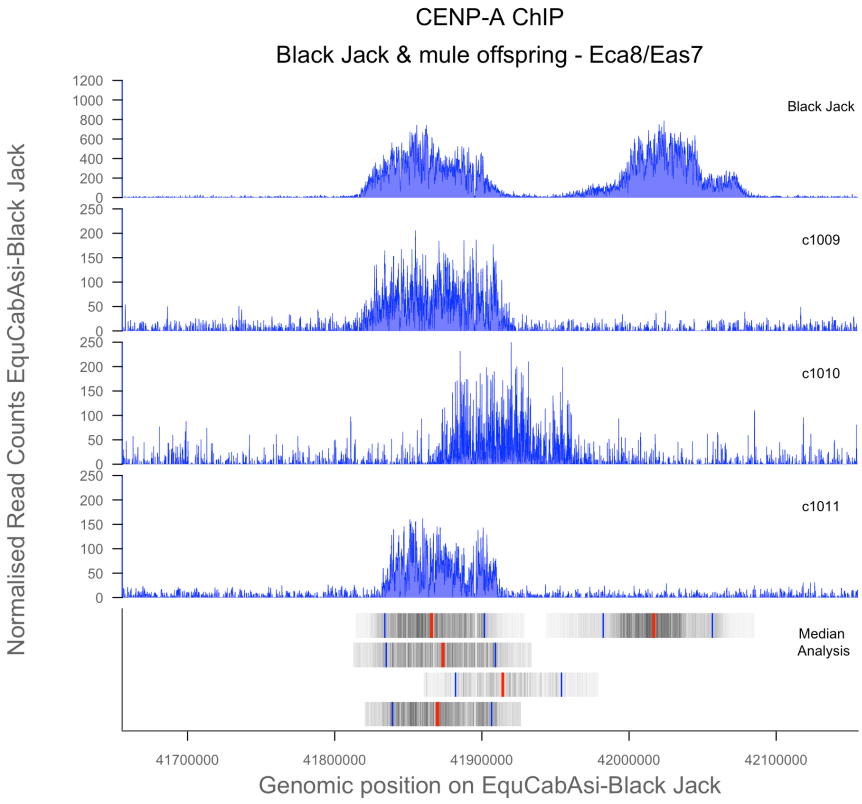


Figure 3.14 Centromere sliding analysis – Family data
Sliding analyses showing centromere movement across family datasets in ECA8/EAS7. Rightward shift of c1010 allele can be observed. Median analysis in bottom panel shows this clearly. Red bars in bottom panel denote peak centre of gravity calculated by determining the positional median. Blue bars represent centromere boundaries determined by calculating the exact positions at which 5% and 95% of the integrated ChIP signal are located. ChIP signal across each domain displayed as heatmap of read counts (grey).

This analysis included a subset of centromeres (ECA nomenclature: 6, 8, 9, 11, 13, 17, 19, 25, 27, 28, 30). The centromeres 5, 14, 20, 26 were excluded from the analysis due to their nature as stable positioned "spike" like centromeres. The X chromosome was also excluded, as this centromere is not present in the offspring. A displacement value between 1-78 kb was observed, with an average of 21.8 ± 20.3 kb displacement in the concepti centromeres, relative to the parental locus positions.

Table 3.7 Absolute displacement in family centromeres

Centromere	Displacement (bp)	Centromere	Displacement (bp)
cen17.c1009	1000	**cen25.c1010**	17000
cen19.c1011	1000	**cen13..c1011**	17000
cen25.c1009	4000	**cen19.c1010**	18000
cen11d.c1011	5000	**cen17.c1010**	19000
cen17.c1011	5000	**cen30.c1009**	19000
cen27.c1011	5000	**cen11d.c1010**	20000
cen6.c1010	7000	**cen13..c1010**	21000
cen8.c1009	7000	**cen25.c1011**	26000
cen11d.c1009	8000	**cen27.c1010**	26000
cen8.c1011	8000	**cen30.c1010**	46000
cen6.c1009	9000	**cen8.c1010**	50000
cen28.c1011	11000	**cen9.c1010**	52000
cen9.c1011	11000	**cen30.c1011**	57000
cen28.c1009	12000	**cen9.c1009**	57000
cen6.c1011	12000	**cen27.c1009**	62000
cen19.c1009	13000	**cen28.c1010**	78000
cen13..c1009	16000		

Table 3.8 Whole population measurements – family experiment

Measurment	Displacement (kb)
Average	21.8
Median	16
Standard Deviation	20.3
Avg+1SD	42.1
Avg+2SD	62.4

Only one centromere (cen28_c1010) was >2 standard deviations away from the average, the standard for significance at the $p<0.05$ level. Visual inspection of the ranked displacement across the offspring centromeres revealed a set of outliers with displacements 1 standard deviation or more than the average displacement. When the two groups are examined separately, it is clear that the highly displaced centromeres are distinct in their behaviour (Figure 3.15). The group 2 centromeres had an average displacement of 57.4 ± 10.4 kb compared to the group 1 centromeres showing an average displacement of 12.2 ± 7.1 kb (Table 3.9).

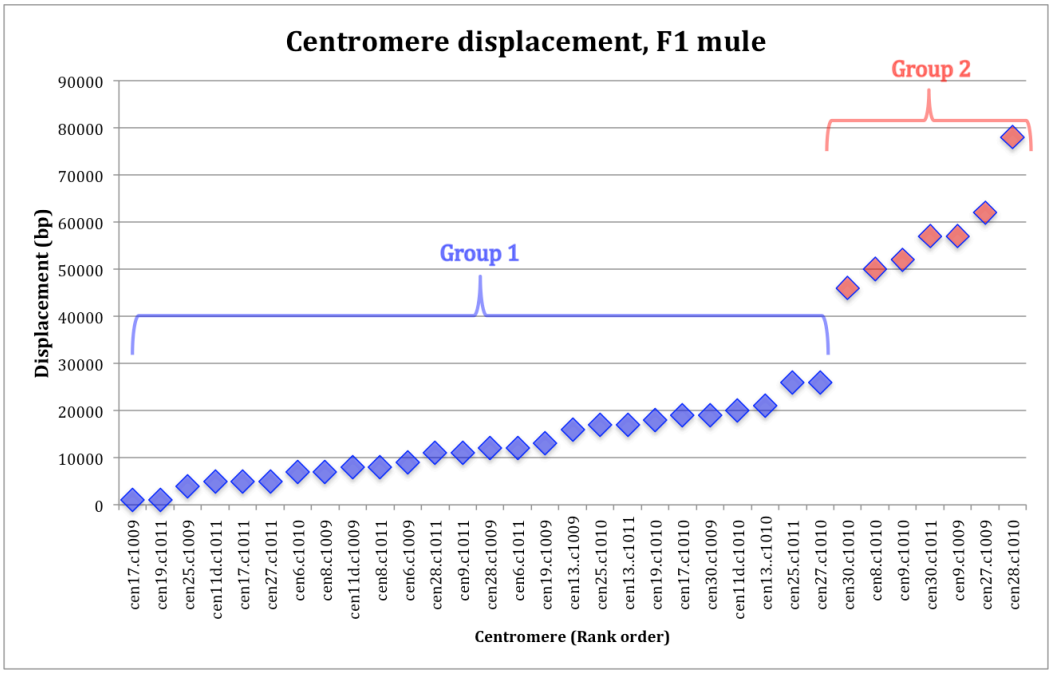


Figure 3.15 Centromere displacement groups
The absolute displacement of mule centromere domains, relative to the parent, ordered by rank. Two groups of centromeres are separated by level of displacement, low displacement (Group 1) and high displacement (Group 2).

Table 3.9 Centromeres grouped on displacement values

Group	Measurement	Value (kb)
Group 1	Avg	12.2
	Med	11.5
	SD	7.1
Group 2	Avg	57.4
	Med	57
	SD	10.4

Displacement analysis indicates the set of centromeres shown below in Table 3.10 as candidates for centromeres that exhibit movement between generations. Two of these, are the centromeres on ECA9 where two alleles can be seen in the parent. One conceptus, c1009, shows transmission of the rightward allele with a pronounced leftward shift of most of the CENP-A signal and a clear two-domain structure that is absent in the parent. The other is c1010, in which the rightward parental allele is transmitted but shows a leftward shift in c1010 of about 50 kb relative to Blackjack's allele position (see Figure 3.13-A). Another two high displacement centromeres are those on ECA30. The parental centromere shows a single broad CENP-A distribution without resolution of the alleles. The distribution of CENP-A in c1010 and c1011 show two domains. The rightmost peak is absent in the parental centromere (Figure 3.13-O).

Table 3.10 High displacement centromeres

Centromere	Displacement (kb)	Standard Deviations
cen30_c1010	46	1.0
cen8_c1010	50	1.2
cen9_c1010	52	1.3
cen30_c1011	57	1.5
cen9_c1009	57	1.5
cen27_c1009	62	1.7
cen28_c1010	78	2.5

The parental centromere in ECA28/EAS4 profiles two alleles as separate peaks. The mule offspring c1010 shows transmission of the rightward allele, which exhibits two subdomains of CENP-A binding in the parent. C1010 shows a single domain displaced to the left by 78 kb (Figure 3.13-N).

The parental ECA27/EAS27 peak exhibits a single broad distribution indicating that the alleles are not discretely separated. C1009 shows a substantial leftward shift, but this maybe because the single alleles are not clearly distinguishable and thus could represent the leftward allele of the parental homologue that transmitted this centromere. The distribution is Gaussian-like rather than the multi-domain structure of the parent, indicating that some rearrangement of CENP-A distribution within the centromere domain has occurred (Figure 3.13-M).

ECA8/EAS7 shows a well-resolved multi-allele distribution in the parent. In c1010, the leftward allele has been transmitted and a discrete displacement can be observed. The rightward shift is approximately 50 kb (Figure 3.14).

These analyses show that CENP-A domains transmitted from parents to offspring are generally stable, however, a subset of the domains show clear movement along the underlying DNA sequence.

3.3.8 Analysis of centromere sliding in single cell clones

Two possibilities that could explain when centromere sliding takes place are; (1) the process could take place during meiosis, forming the germ cells that gave rise to the offspring, or (2) during mitosis, either in the animal or in subsequent culture of fibroblasts. In order to test the mitotic stability of CENP-A domains, a similar measurement of sliding analysis was performed on a set of immortalized mule cell lines. Fibroblasts from c1009 were immortalized by transfection with telomerase and 6 single cell clones (named: 1, 3, 5, A, B, N) were isolated and cultured from the parental culture. These were grown for (~50) population doublings and processed for ChIP-seq analysis. As for the family analyses, centromere median positions were determined for 11 centromeres. The parental cell line was used as the reference and centromere displacement was determined relative to this. 10 centromeres were examined: ECA nomenclature 6, 8, 9, 11EAS, 11ECA, 19, 25, 27, 28 and 30. A graphical representation with the median analysis of the clone experiment along with CENP-A peak profiles of two centromere domains is shown in Figure 3.16 below. (All figures for Clone experiment can be seen in Appendix I-Figure 8.20-Figure 8.23)

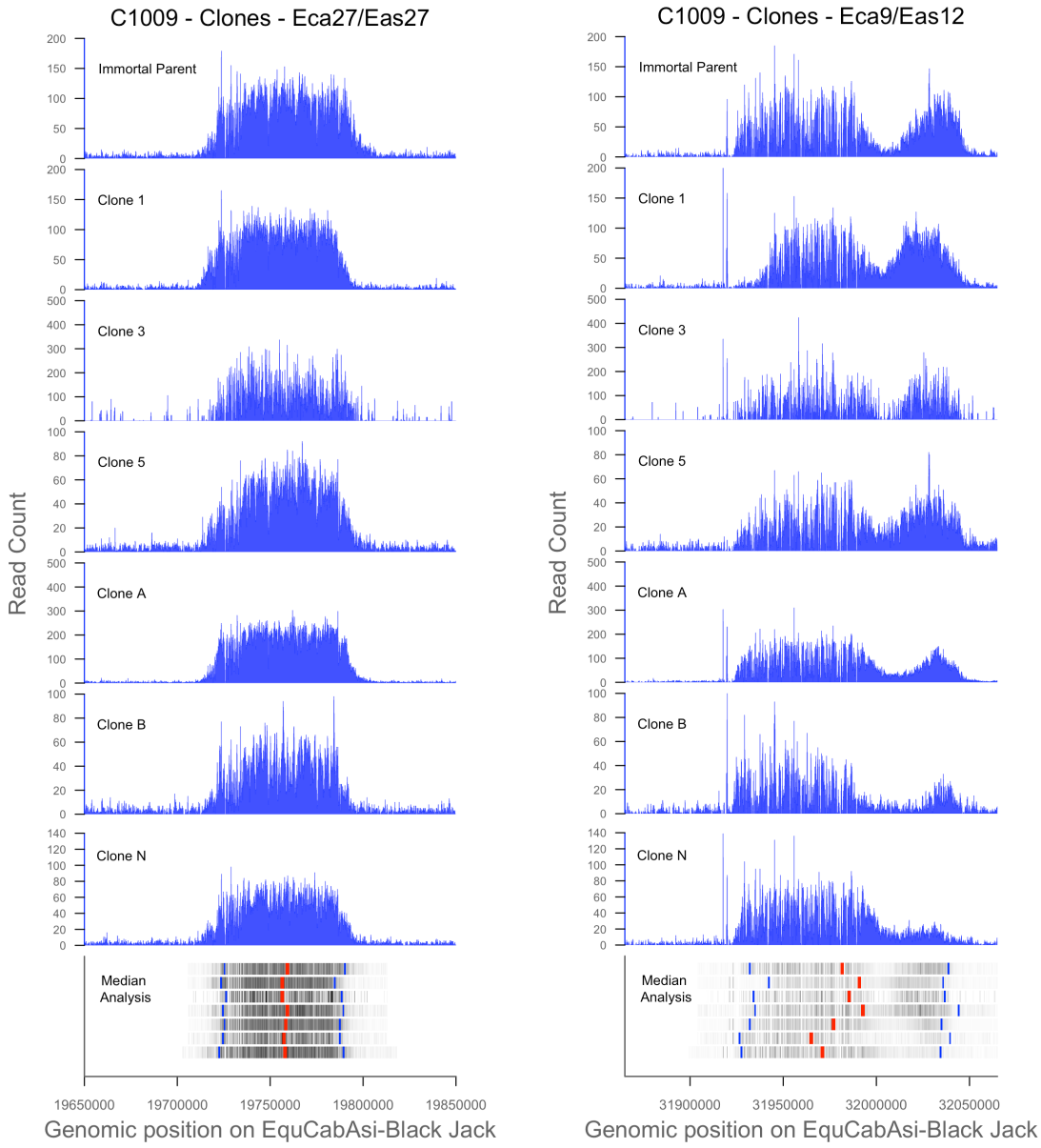


Figure 3.16 Centromere displacement in clonal cell lines
Displacement analyses showing varying levels of centromere movement. Panels display seven independent CENP-A ChIP-seq experiments, performed on asynchronous populations, derived from single cell clones including the parental cell line. Lower displacement observed ECA27/EAS27 (left) compared with ECA9/EAS12 (right). Red bars in bottom panel denote peak centre of gravity calculated by determining the positional median. Blue bars represent centromere boundaries determined by calculating the exact positions at which 5% and 95% of the ChIP signal are located. ChIP signal across each domain displayed as heatmap (grey).

Absolute displacement for each centromere domain is displayed by rank order in Figure 3.17 below. Displacement ranged from 0-18 kb across the sample set with an average displacement of 4.08 ± 3.47 kb (Table 3.11).

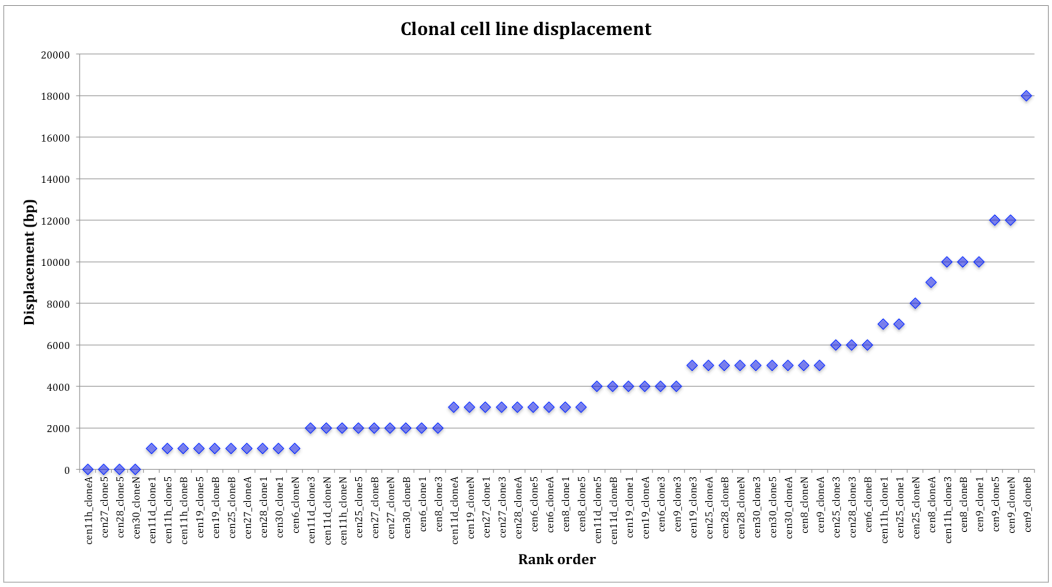


Figure 3.17 Absolute displacement - clone experiment
The absolute displacement of c1009 centromere domains in clonal cell lines, relative to the immortal parent, ordered by rank. Range of displacement is from 0-18 kb.

Table 3.11 Whole population measurements - clone experiment

Measurement	Displacement (kb)
Average	4.08
Median	3
Std	3.47

Analysis by chromosome revealed that most individual centromeres displayed the same displacement as observed for the population, in this experiment. One chromosome, however, showed higher displacement than the rest across the set of clones, ECA9/EAS12 (Figure 3.18). It could be that this centromere is more susceptible to sliding for some reason related to their DNA content or chromatin architecture. The centromere mapping to ECA9/EAS12 also showed high displacement in the family experiment.

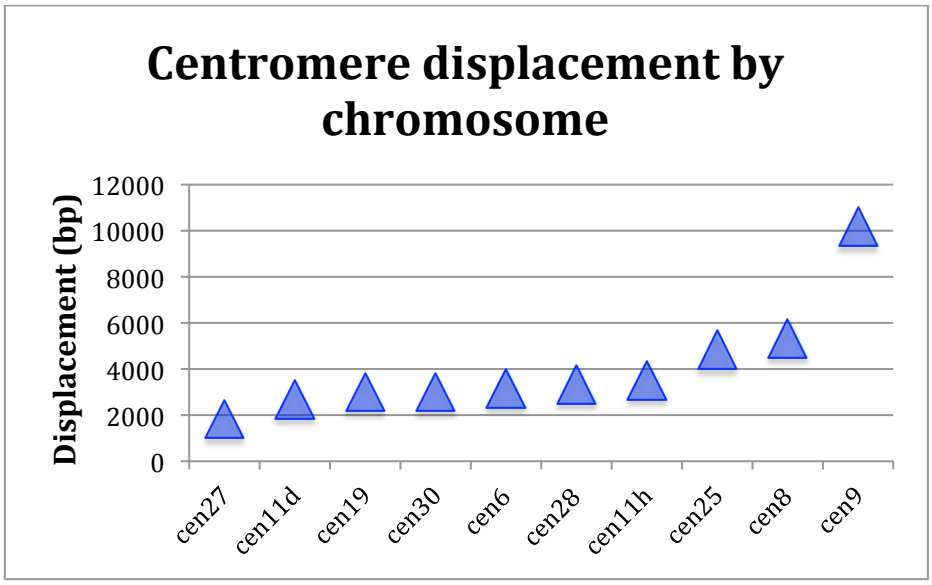


Figure 3.18 Centromere displacement by chromosome - clone experiment
Average displacement by chromosome shows uniform behavior across most centromere alleles measured. Cen9 (ECA9/EAS12) shows a higher average displacement with ~2 fold excess compared to others.

In order to compare the distribution of centromere displacement within each clone population, a whisker plot was generated. It revealed uniform behaviour in each of the cell lines (Figure 3.19).

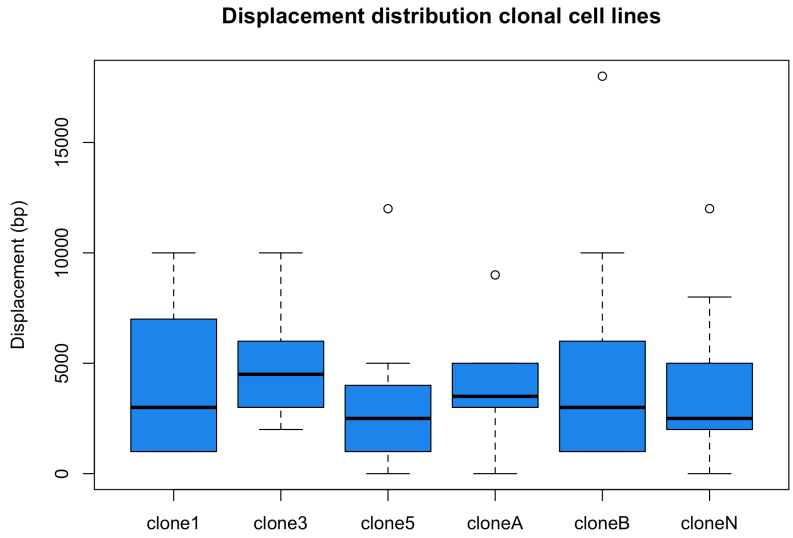


Figure 3.19 Average displacement across each clone
Boxplots showing the distribution of displacement across each clonal population. Data show similar behaviour across all clones with clone 3 showing a higher median displacement compared to the others.

The average displacement across each clone population ranges between 3.1 kb and 5 kb with the average of the total population being ~4 kb. These data also show the uniform behaviour across each cell line (Table 3.12). In an effort to examine variance between clonal cell lines an ANOVA (Analysis of variance) test was performed. The results displayed in Table 3.13 show p=0.88 indicating no significant variation in centromere behaviour in the clonal cells lines.

Table 3.12 Average displacement across clones

Clone	Average displacement (bp)	StdDev (bp)
clone1	3900	2948
clone3	4700	2238
clone5	3100	3360
cloneA	3800	2358
CloneB	5000	5119
CloneN	4000	3464
whole population	4083	3470

Table 3.13 ANOVA output

F	df	p-value
0.35	5.00	0.88

Collectively these results demonstrate that centromeres in cell lines subject only to successive rounds of mitosis display little evidence of centromere sliding compared to the transmission of centromere alleles from parental cell lines to F1 progeny. These results provide us with insights into the how regulation and maintenance of CENP-A chromatin domains are affected through transmission of centromere alleles.

3.3.9 CENP-A Abundance at satellite-free centromeres

The CENP-A ChIP-seq reads represent CENP-A associated DNA and we wanted to use this information to calculate the relative abundance of CENP-A at each one of the satellite-free centromeres. Using the previously identified coordinates for the Asino Nuovo CENP-A domains, the average integrated intensity for the autosomes within a given dataset was determined, by summing integrated count data across all autosomal centromeres and dividing by the number of autosomes. A normalization step was then performed, by dividing the integrated counts at each centromere domain by this average value. The X chromosome was treated separately, as it is present in haploid copy number in the male AN cells. The results shown in Figure 3.20 indicate that the levels of CENP-A are strictly maintained at the different centromeres. While the centromere regions occupied by CENP-A varied between ~57-320 kb, the amount of CENP-A associated DNA in the libraries from each centromere was remarkably constant. The notable exception is the X chromosome, which contained 50% of the signal detected at autosomes. As mentioned, this donkey is a male and so is only haploid. This confirms the sensitivity of the quantitation method. The error bars denote the standard deviation, which is a value of 0.12 and demonstrates the uniformity in the content of CENP-A across each domain. This quantitative metric was also applied to the Blackjack dataset and the results show uniform abundance of CENP-A across the entire dataset. Quantitative analysis performed on the individual alleles that could be discretely separated showed that ~50% of the CENP-A signal at the entire centromere domain was accounted for by each allele Figure 3.21. These results suggest the existence of a tightly maintained regulatory pathway for CENP-A.

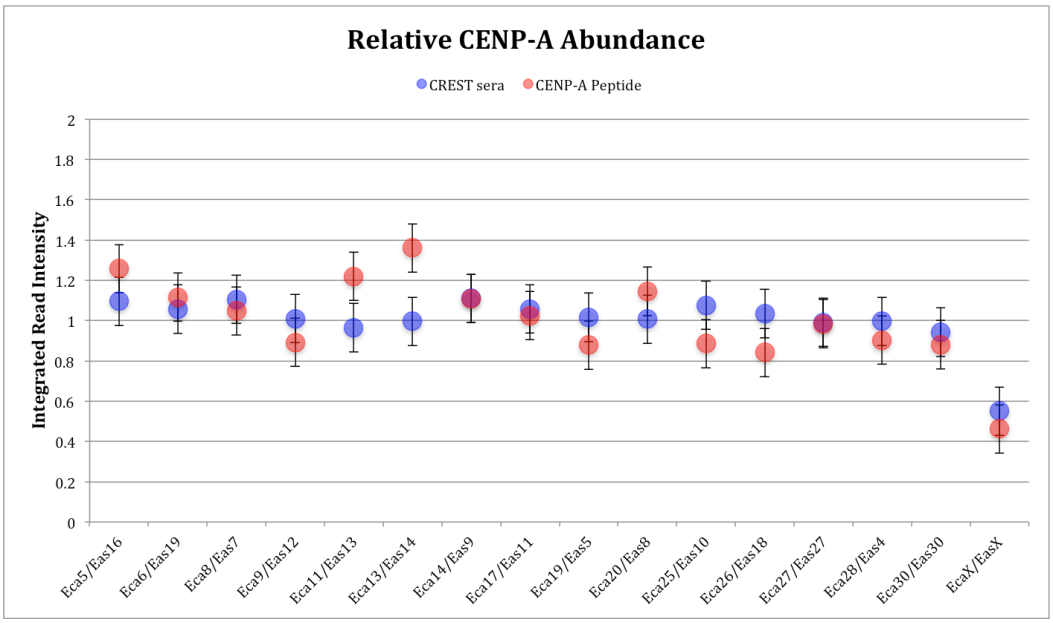


Figure 3.20 Relative CENP-A Abundance
Plot showing the relative abundance of CENP-A across satellite-free centromere domains. Data display uniform level of CENP-A in both the CREST sera and peptide antibody datasets.

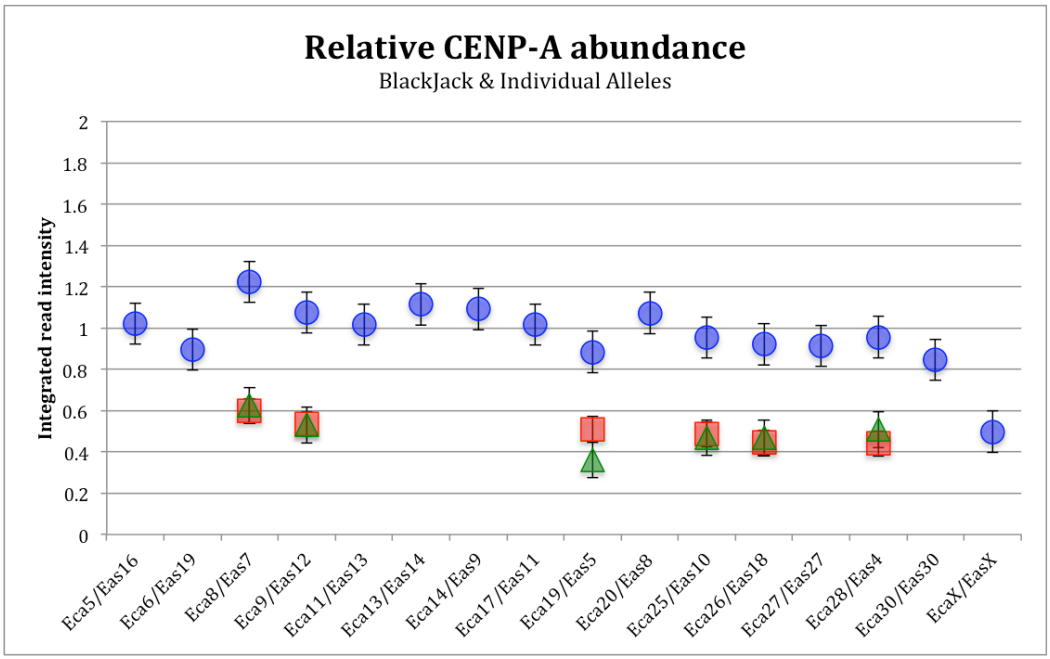


Figure 3.21 Relative CENP-A abundance - Blackjack
Black Jack data display uniform level of CENP-A across each centromere. Where the CENP-A alleles were separable, quantitative analysis shows ~50% of the total CENP-A signal at the entire domain is represented by each allele.

3.4 Concluding statement

Collectively, this chapter describes the identification and assembly of 16 satellite-free centromere domains in *E. asinus*, which will prove as a useful genomic tool for investigating centromere structure and function.

We also showed that centromere sliding, originally observed in the horse, also occurs in the donkey and that transmission of CENP-A domains from parents to offspring showed that centromere positional variation is a property of centromeres. Then, using a ChIP-seq approach on cell populations derived from single cell clones, we demonstrated that centromeres in cell lines subject only to successive rounds of mitosis display little evidence of centromere sliding compared to the transmission of centromere alleles from parental cell lines to F1 progeny. These results provide us with insights into the how regulation and maintenance of CENP-A chromatin domains are affected through transmission of centromere alleles. Using a quantitative approach, we calculated the relative abundance of CENP-A across the unique sequence centromere domains and showed very clearly a remarkable uniformity in CENP-A abundance across the set of centromeres, suggesting the existence of a tightly maintained regulatory pathway for CENP-A.

Chapter 4 - Analysis of CENP-A distribution within centromeres

4.1 Introduction

A central question in centromere research asks how the centromeric chromatin landscape is organised. While the importance of CENP-A in maintaining centromere integrity through its specialised protein configuration (Black et al., 2004), its deposition pathway and interactions with other proteins (Reviewed in - McKinley and Cheeseman, 2016), the organisation of CENP-A nucleosomes on the chromatin fiber is not well understood at a molecular level. Previous studies using three-dimensional deconvolution light-microscopy showed CENP-A forms a cylindrical-like structure at the primary constriction of metaphase chromosomes in flies and humans (Blower et al., 2002). Quantitative analysis on fly metaphase chromosomes showed that the CENP-A domains contained histones H2A, H2B but not H3 or phosphorylated H3. However, high-resolution immunofluorescence on stretched interphase and mitotic chromatin fibers of both flies and humans showed that CENP-A remains interspersed with histone H3 throughout the domain. CENP-A staining was said to be present on one-half to two-thirds of the entire centromere domain (Blower et al., 2002). Further experiments using electron microscopy on human alpha-satellite containing centromeres and neocentromeres revealed that CENP-A occupies a much narrower domain than previously observed, with CENP-A nucleosomes occupying in the range of 6-8% of the entire domain by volume (Marshall et al., 2008b). These results suggest that while the linear centromeric chromatin fiber consists of CENP-A nucleosomes interspersed with other nucleosome subunits, the three-dimensional architecture of the centromeric chromatin domain establishes a specific configuration in order to assemble kinetochore machinery.

Studying the structural properties of nucleosomes reveals more insight into the positions and configuration they occupy. Nucleosomes are the fundamental structural components of chromatin and the high conservation of canonical nucleosomes along with CENP-A across different species illustrates the defining role of epigenetics in eukaryotes (Malik and Henikoff, 2003). Canonical nucleosomes typically protect 147 bp of DNA which is wrapped around the octameric particle 1.65 times (Luger et al.,

1997). Structural properties at the amino terminal tail in the canonical histone H3 nucleosome aid the stability of the DNA binding at the entrance and exit paths of the nucleosome (Luger et al., 1997; Tachiwana et al., 2011; Tsunaka et al., 2005). *In vitro* reconstitution experiments have shown that CENP-A nucleosomes can be reconstituted on 121 bp of DNA and through MNase footprinting and ChIP-seq approaches CENP-A nucleosomes have been shown to protect DNA size fragments from 100-150 bp (Hasson et al., 2013; Tachiwana et al., 2011; Yoda et al., 2000). The difference in DNA protection between histone H3 and CENP-A nucleosomes is proposed to be caused by the structural differences in the two proteins as the CENP-A αN helix preceding the α1 helix of the HFD is at least 1 helical turn shorter thereby having less ability to bind entry and exit path DNA. These data suggest CENP-A DNA binding may be related to its function on chromatin.

Nucleosome positioning has been highlighted regarding its central involvement in structural, functional and regulatory processes of chromatin. Genome wide nucleosome positioning studies gave insights into the relationship between well-positioned nucleosomes surrounding transcription start sites (TSS) that impact accessibility to promoter regions (Jiang and Pugh, 2009). "Fuzzy" or delocalised nucleosome positions were also observed. These genomic regions have no discrete nucleosome position but tend to be occupied by multiple nucleosome configurations and are proposed to be delocalised in order for transcriptional machinery to assemble (Jiang and Pugh, 2009). Nucleosome positioning studies using tilling arrays in *S. pombe* showed that the central domain (cnt) of centromeric chromatin at all three of the *S. pombe* chromosomes is occupied by orderly positioned CENP-A^{Cnp1p} nucleosomes (Song et al., 2008). These results can be related to CENP-A positioning studies in human alpha-satellite containing centromeres which showed by MNase footprinting (Ando et al., 2002) and ChIP-seq (Hasson et al., 2013) that CENP-A nucleosomes can be positioned in phase between regions of CENP-B binding. CENP-B was also shown to stabilize CENP-A nucleosomes upon recruitment of CENP-C (Fachinetti et al., 2015); this happens concurrently with direct CENP-C recruitment through the carboxy terminal tail of CENP-A (Carroll et al., 2010; Fachinetti et al., 2013; Kato et al., 2013). These data provide good reason to ask whether CENP-A is regularly positioned in neocentromeres that are known not to contain CENP-B (Voullaire et al., 1993).

While the vital importance of CENP-A function is widely documented, little is known about native CENP-A nucleosome organisation at vertebrate centromeres. In order to gain more insight into the architecture of centromeric chromatin, CENP-A nucleosome distribution and the influence of this distribution throughout the centromere domain needs to be elucidated. This chapter describes a native ChIP-seq approach to directly examine CENP-A nucleosome distribution at satellite-free centromeres in *E. asinus*. These experiments aimed to determine CENP-A nucleosome positions across the satellite-free centromere domains and quantify relative CENP-A occupancy at these positions.

4.2 CENP-A antibody characterisation

In order to obtain a CENP-A antibody suitable for ChIP an antibody directed against horse CENP-A was produced (Teri Masterson). Antiserum obtained from sheep immunisation was characterised by western blot analysis in immortalised donkey cell fractions, shown in Figure 4.1. $1x10^5$ cells or cell equivalents (C.E.) loaded in each lane showed signal in all chromatin containing fractions with an prominent band corresponding to the size of CENP-A (16kDa) just above the 14kDa marker. In order to further assess the CENP-A specificity of the antibody, western blot analysis was performed on purified recombinant CENP-A protein and isolated mononucleosomes. These results showed a single band just above the 14kDa marker in both blots. In addition, the mononucleosome preparation showed a minor species just below 14kDa which may suggest protein degradation or may be signal coming from another isoform of CENP-A. These data are shown in Figure 4.1-B. In order to isolate CENP-A specific immunoglobulin, a sample of the antiserum was affinity purified using CENP-A coupled to cyanogen bromide-activated sepharose beads (Teri Masterson). Centromere specificity of the antibody was examined by immunofluorescence. Figure 4.1-C shows immunofluorescence performed using affinity purified sheep CENP-A antibody (Cy5-red) on mitotic immortalised donkey skin fibroblasts and donkey metaphase chromosomes (DAPI – blue) showing double dot staining indicating centromere specific binding. In order to validate the CENP-A antibody in immunoprecipitation, a ChIP-qPCR approach was taken. Native chromatin was prepared using isolated nuclei from $1x10^7$ immortalised donkey skin fibroblasts using 8U/mL of *S7 nuclease* as described in 2.2.4.9. A native ChIP experiment (See section 2.2.4.12) was performed using an affinity purified CENP-A antibody, DNA was purified from immunoprecipitates and used in a qPCR experiment with three primer pairs. As described in section 2.2.3.3 “EAS30” is a positive centromere probe that amplifies a genomic interval within the donkey centromere on chromosome 30.

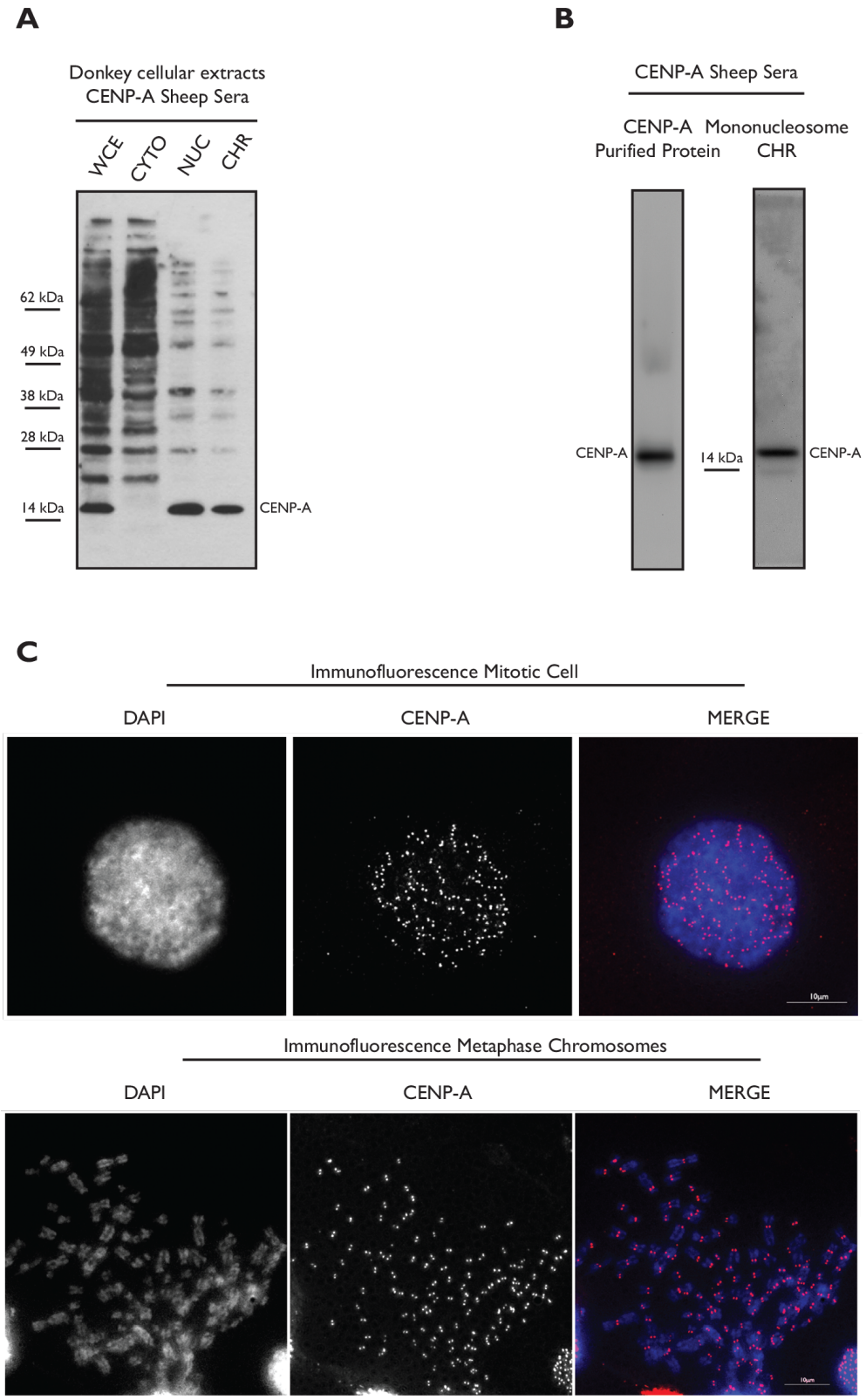


Figure 4.1 Characterisation of CENP-A sheep sera antibody
Western blot analysis using CENP-A sheep sera antibody (Teri Masterson) on donkey cellular extracts reveal affinity for a band correspond to the size of CENP-A (**A**). Further western blot analysis using CENP-A sheep sera detecting purified recombinant CENP-A protein and CENP-A from mononucleosomal chromatin fractions (**B** – left & right). (**C**) Immunofluorescence data shows a mitotic cell (top) with CENP-A sheep sera and on metaphase chromosomes (bottom – Teri Masterson) showing distinct foci representative of CENP-A binding

"ECA11" is a negative control that corresponds to a region of horse chromosome 11 that is centromeric in horse but not in the donkey. "PRKC" is a region of the housekeeping gene PRKC apoptosis WT1 regulator, which is non-centromeric and therefore not expected to bind CENP-A. The qPCR data displayed in Figure 4.2 show a 45% recovery in CENP-A associated DNAs with a 1-2% recovery in negative controls. Collectively these data verify CENP-A antibody that recognises and is effective in immunoprecipitation of donkey CENP-A and therefore is a suitable antibody for ChIP-seq.

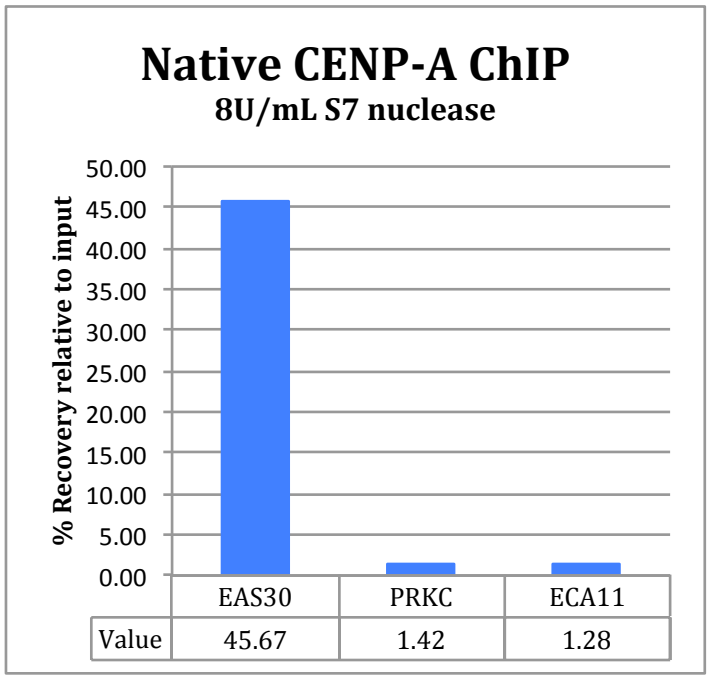


Figure 4.2 CENP-A ChIP-qPCR
ChIP-qPCR data from native CENP-A ChIP experiment showing % recovery of centromere associated DNA compared to non-centromere associated DNA.

4.3 Chromatin preparation, nucleosome isolation & immunoprecipitation

In an effort to obtain occupancy maps that resolve CENP-A nucleosome positions across donkey satellite-free centromere domains a native ChIP-seq approach was carried out. Concentrations of *S7 nuclease* (micrococcal nuclease) were titrated for a fixed time using isolated 10^7 nuclei from immortalised donkey skin fibroblasts (derived from "Asino Nuovo" cells in previous chapter) in order to identify conditions for effective chromatin digestion (Figure 4.3-A). Effective chromatin digestion was determined to be when ~50% of input material was reduced to mononucleosome fragments, minimising over digestion. Subsequently this method scaled up and performed on 10^8 isolated nuclei from immortalised donkey skin fibroblasts. Solubilised bulk chromatin was fractionated through a sucrose density gradient in order to isolate nucleosome arrays of different size classes: mono-, di-, tri- and oligonucleosomes. Samples from each fraction including the pre-fractionated bulk chromatin were treated with Proteinase K, column purified and run on an agarose gel to confirm separation of nucleosome arrays Figure 4.3. The chromatin corresponding to the mononucleosome fractions (denoted by red dot in Figure 4.3-B) were processed in a ChIP assay with the antibody against CENP-A. The mononucleosome (Mono1) ChIP sample was subsampled kept for SDS-PAGE/western blot analysis shown in Figure 4.3 below and show CENP-A signal in the "ChIP" fraction.

Another independent mononucleosome sample (Mono2) was prepared from a separate cell population as a replicate (Appendix II-Figure 8.24). A trinucleosome fraction (trinuc) was also processed with the aim of cross-referencing CENP-A positions from the mononucleosome data (Figure 4.3-B, sample-14).

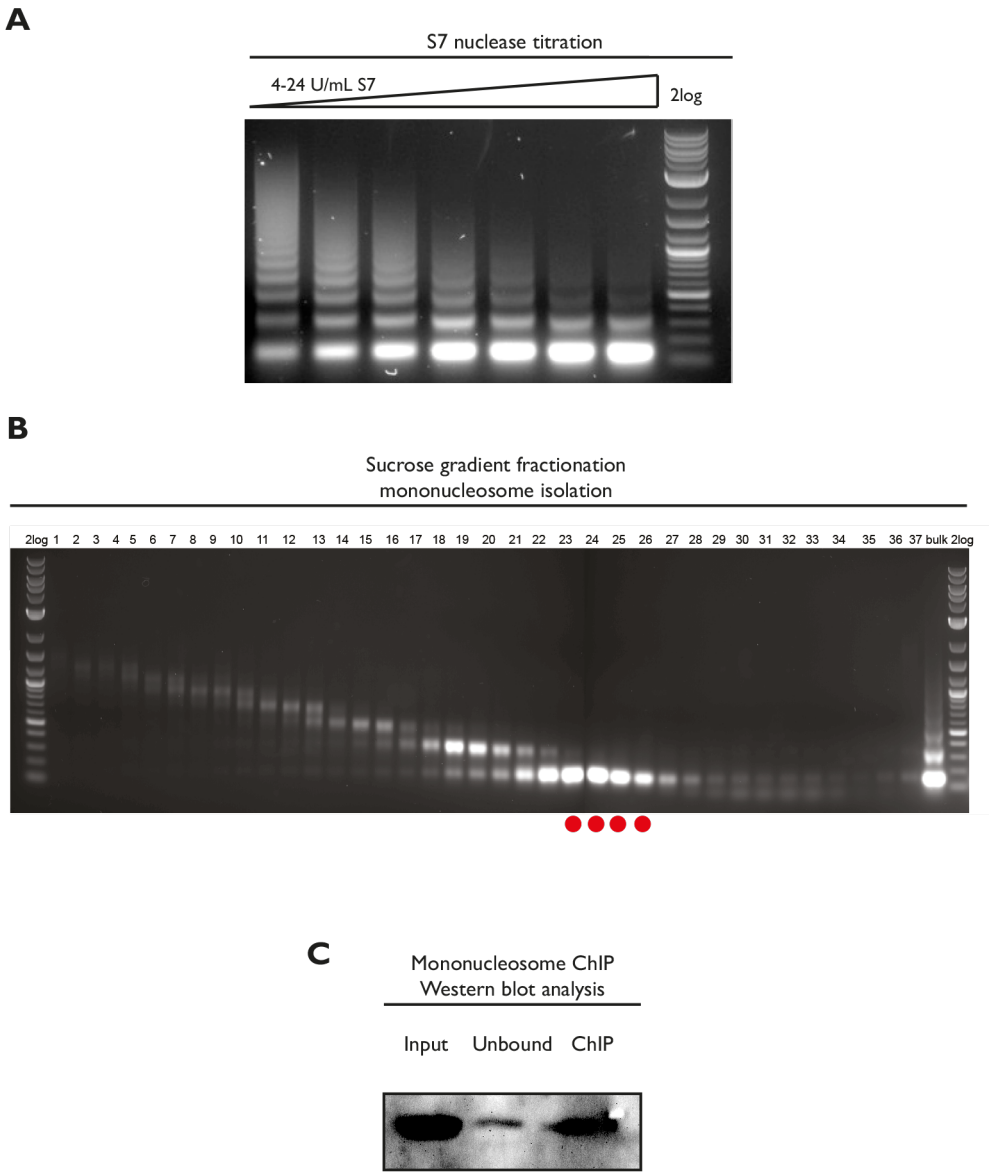
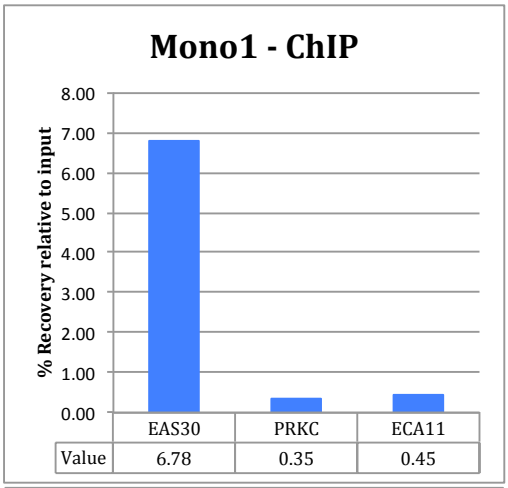
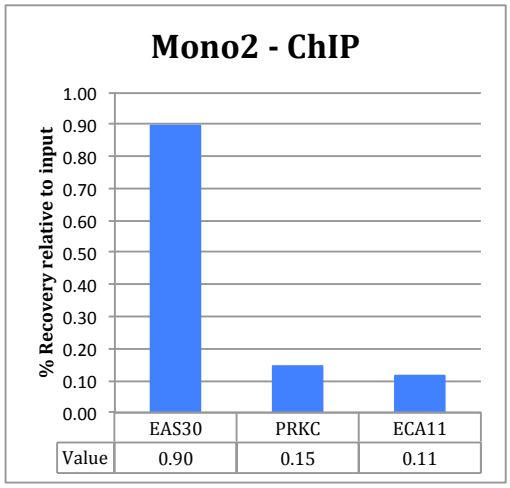


Figure 4.3 Chromatin preparation and western blot analysis
(**A**) S7 nuclease titration on donkey chromatin showing increasing extent of digestion resulting in higher concentrations of mononucleosomal fragments (4-24U/mL). (**B**) Sucrose gradient fractionation of donkey chromatin showing isolation of mono- and oligonucleosome DNAs with bulk chromatin sample as reference. Sample 14 was used in the Trinuc ChIP experiment (**C**) CENP-A mononucleosome immunoprecipitation monitored by western blot analysis showing enrichment of CENP-A in ChIP fraction.

4.4 Enrichment in centromere specific DNA in native immunoprecipitation

The purified DNA from the ChIP and Input samples from all the native fractions were prepared as described in section 2.2.3.3 for qPCR. The percentage recovery of centromere specific DNAs in the mononucleosome immunoprecipitation, the independent mononucleosome replicate and the trinucleosome sample are shown in Figure 4.4 below. The Mono1 ChIP showed a ~6.7% recovery in centromere specific DNA with a ~15.1 and ~19.3 fold enrichment with respect to negative probes (Table 4.1). A low recovery in CENP-A associated DNA was observed in the Mono2 sample (~0.9%) with an overall lower signal to noise when comparing fold enrichment with respect to the negative probes (Table 4.1). The Trinuc sample showed ~19.8% recovery in centromere specific DNA and a much higher signal to noise ratio (ECA11- ~40.2, PRKC- ~39). These results show immunoprecipitation of CENP-A in all three ChIP assays however a wide range of variability can be observed. Notably, CENP-A immunoprecipitation from mononucleosome samples tended to show lower recovery when compared to CENP-A immunoprecipitations using oligonucleosome samples (data not shown).





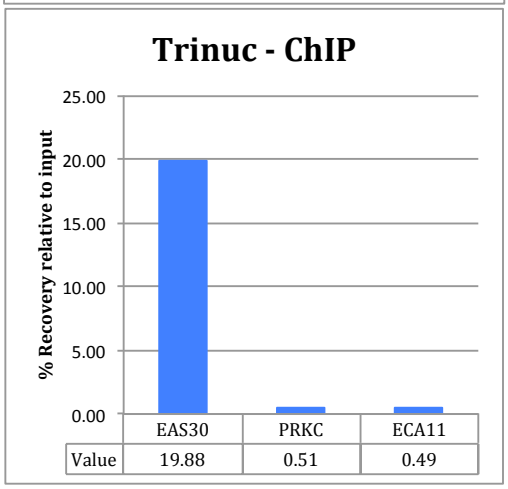


Table 4.1 Fold enrichment relative to negative primer pairs

"-"Probe	ECA11	PRKC
Mono1	15.14	19.38
Mono2	7.89	6.18
Trinuc	40.22	39.03

Figure 4.4 qPCR analysis of native CENP-A immunoprecipitations
Recovery of DNA assayed by qPCR is expressed as the percent recovered for centromeric DNA, EAS30, and two negative control regions, PRKC and ECA11. Three experiments used for subsequent DNA sequencing are shown: Mono1, Mono2 and Trinuc. The relative fold enrichment of centromeric DNA is shown in the Table 4.1

4.5 ChIP-Seq data

As described in the previous section, in order to identify CENP-A nucleosome positions across satellite-free centromeres in *E. asinus* two independent native chromatin preparations were used to isolate CENP-A mononucleosomal DNAs. In addition, a trinucleosome fraction was isolated with the aim of cross-referencing CENP-A positions from the mononucleosome data. DNA purified from immunoprecipitates and provided to *IGA Technologies* for library preparation and sequencing. In brief, 10-100ng of DNA was sent, end repaired, A-tailed and Illumina *Trueseq* adapters were ligated. DNA libraries were then subject to paired-end sequencing on the Illumina HiSeq 2500 platform.

In order to saturate CENP-A coverage the Lander/Waterman equation was used, a method for estimating genome coverage (Coverage (C)= Read length (L) x Number of reads (N)/Genome length (G)) (Lander and Waterman, 1988). We used a modified version of this equation in which "Genome length (G)" was replaced with "total centromere length" ($2.4x10^6$). It was that estimated $40x10^6$ reads would correspond to up to 1000X coverage across all target domains and so we predicted this to be sufficient for the CENP-A immunoprecipitates sent. Below, Table 4.2 outlines read output from the sequencing reaction. The quantity of read data obtained was sufficient. The reads obtained were of a desirable length between 100-130 bp, which would allow for near-complete coverage of CENP-A mononucleosome bound DNA fragments.

Table 4.2 Native ChIP-seq data summary

Dataset	Filename	Total Sequences	Seq. length
CENP-A Mono1 Input	1_AGAGGATG_L001_R1_001.fastq.gz	5442394	100
	1_AGAGGATG_L001_R2_001.fastq.gz	5442394	100
	1_AGAGGATG_L002_R1_001.fastq.gz	5616758	100
	1_AGAGGATG_L002_R2_001.fastq.gz	5616758	100
	1_AGAGGATG_L003_R1_001.fastq.gz	3850737	125
	1_AGAGGATG_L003_R2_001.fastq.gz	3850737	125
	1_AGAGGATG_L005_R1_001.fastq.gz	11209268	130
	1_AGAGGATG_L005_R2_001.fastq.gz	11209268	130
		52,238,314	
CENP-A Mono1 ChIP	2_ACGCTTCT_L001_R1_001.fastq.gz	4236136	100
	2_ACGCTTCT_L001_R2_001.fastq.gz	4236136	100
	2_ACGCTTCT_L002_R1_001.fastq.gz	2853330	125
	2_ACGCTTCT_L002_R2_001.fastq.gz	2853330	125
	2_ACGCTTCT_L005_R1_001.fastq.gz	13001153	130
	2_ACGCTTCT_L005_R2_001.fastq.gz	13001153	130
		40,181,238	
CENP-A Mono2 Input	4_AGTCAGGT_L001_R1_001.fastq.gz	1786181	100
	4_AGTCAGGT_L001_R2_001.fastq.gz	1786181	100
	4_AGTCAGGT_L002_R1_001.fastq.gz	1853772	100
	4_AGTCAGGT_L002_R2_001.fastq.gz	1853772	100
	4_AGTCAGGT_L003_R1_001.fastq.gz	4279091	125
	4_AGTCAGGT_L003_R2_001.fastq.gz	4279091	125
	4_AGTCAGGT_L005_R1_001.fastq.gz	14164253	130
	4_AGTCAGGT_L005_R2_001.fastq.gz	14164253	130
		44,166,594	
CENP-A Mono2 ChIP	5_TAGCAGGA_L003_R1_001.fastq.gz	2663627	125
	5_TAGCAGGA_L003_R2_001.fastq.gz	2663627	125
	5_TAGCAGGA_L005_R1_001.fastq.gz	20561732	130
	5_TAGCAGGA_L005_R2_001.fastq.gz	20561732	130
		46,450,718	
CENP-A Trinuc Input	6_CATGGATC_L003_R1_001.fastq.gz	5269484	125
	6_CATGGATC_L003_R2_001.fastq.gz	5269484	125
	6_CATGGATC_L005_R1_001.fastq.gz	18909418	130
	6_CATGGATC_L005_R2_001.fastq.gz	18909418	130
		48,357,804	
CENP-A Trinuc ChIP	7_CTCGAACA_L003_R1_001.fastq.gz	1684788	125
	7_CTCGAACA_L003_R2_001.fastq.gz	1684788	125
	7_CTCGAACA_L005_R1_001.fastq.gz	11600894	130
	7_CTCGAACA_L005_R2_001.fastq.gz	11600894	130
		26,571,364	

4.5.1 Quality control

Data obtained by Illumina sequencing were taken through a series of quality control steps. As described in Chapter 3, *FastQC* was used to carry out quality metrics on the data. The quality of sequence reads in FASTQ format is determined by the use of a Phred score, which is calculated by the methods described in (Ewing and Green, 1998). A score of 30 indicates a 1/1000 chance of the called base in the sequencing reaction being incorrect. The Phred quality scores for the Mono1 dataset are represented by a whisker plot (Figure 4.5), generated by the *FastQC* suite and clearly show the CENP-A ChIP and Input DNA were of high quality with the average Phred score of ~38 or higher across the entire 100 bp length of the sequence reads (Mono2 & Trinuc FastQC can be seen in Appendix II-Figure 8.25-Figure 8.26). The sequence reads for all three datasets were aligned using *Bowtie2* as before. The native experiments carried out in this study are aimed at identifying candidate CENP-A nucleosome positions and such it is important that the DNAs sent for sequencing are in the mononucleosomal size range. Using *Picard tools,* fragment size metrics were calculated for all aligned native datasets using paired end data. The fragment length distributions were constricted to the known centromere regions. The Mono1 data are represented by high-density histogram plots, shown in Figure 4.6**.** The histogram plots show that the majority of fragments are in the 150-200 bp range for the input sample and in the 120-150 bp range for the CENP-A ChIP sample. The peak of the distribution in the ChIP sample is 136 bp. These data are in agreement with a previous study which describe some CENP-A nucleosome populations that protect 120-150 bp (Hasson et al., 2013). The Mono2 fragment size data are shown in Figure 4.7 and a similar distribution is observed. However, an abundance of larger fragment sizes in this in the ChIP sample can also be observed, which may be reflective of the lower signal to noise seen in the qPCR (Figure 4.4). The Trinuc data displayed in Figure 4.8 show a wide distribution of fragment lengths which is not consistent with a typical trinucleosome sample (~450-550 bp), however a sharp peak at 524 bp is observed. These data indicate possible degradation at a particular stage of the sample processing.

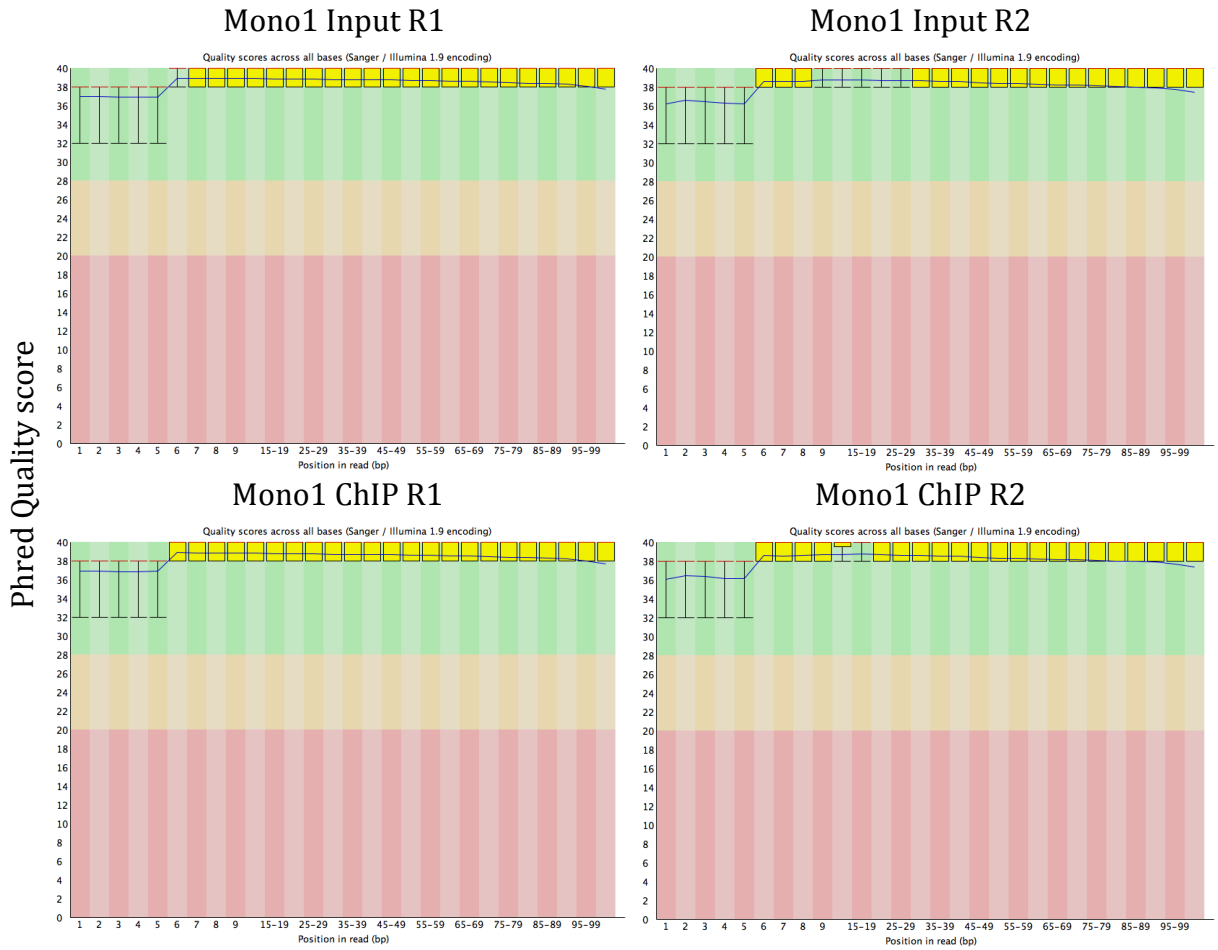


Figure 4.5 Per-base quality scores Mono1
Per base quality metrics for CENP-A ChIP and Input Mono1 libraries. Plots generated using *FastQC*. Phred score is defined in terms of the estimated probability of error. Phred score of 30 is the equivalent to a 1/1000 chance of a called base being incorrect, by the equation $q = -10 \times \log_{10}(p)$.

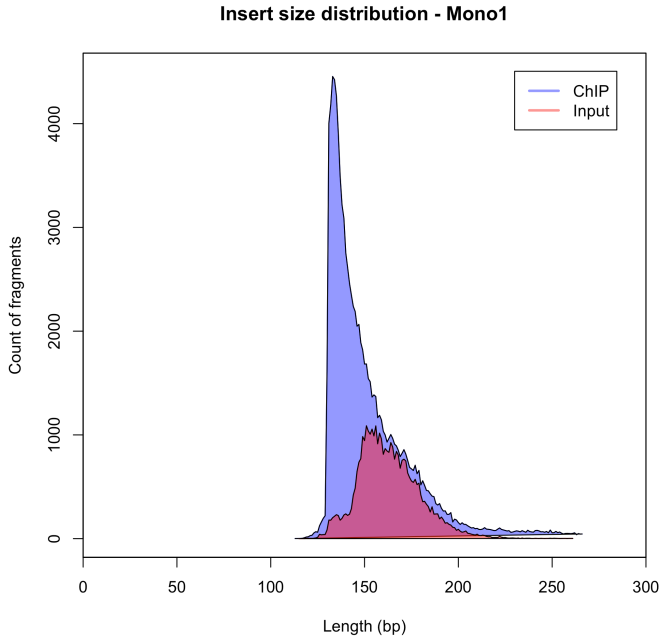


Figure 4.6 Fragment size distribution Mono1
Insert size metrics test performed on Mono1 dataset using *Picard tools*. Size distribution in ChIP sample (blue) shows that the majority of CENP-A fragments (nucleosome bound DNAs) are ~130 bp. This is in contrast to input sample (red) where the majority of the fragments are in the expected ~150 bp range. Size metrics calculated from reads mapped to centromere regions.

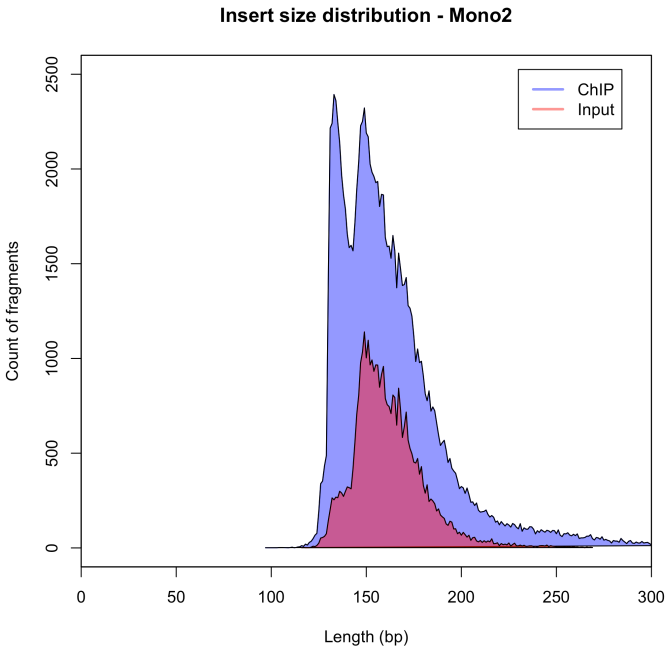


Figure 4.7 Fragment size distribution Mono2
Insert size metrics test performed on Mono2 dataset using *Picard tools*. Size distribution in ChIP sample (blue) shows that while an abundance of CENP-A fragments are ~130 bp, the ChIP sample also exhibits a broad peak around the 150 bp range. The input sample (red) the majority of the fragments are in the expected ~150 bp range. Size metrics were calculated from reads mapped to centromere regions.

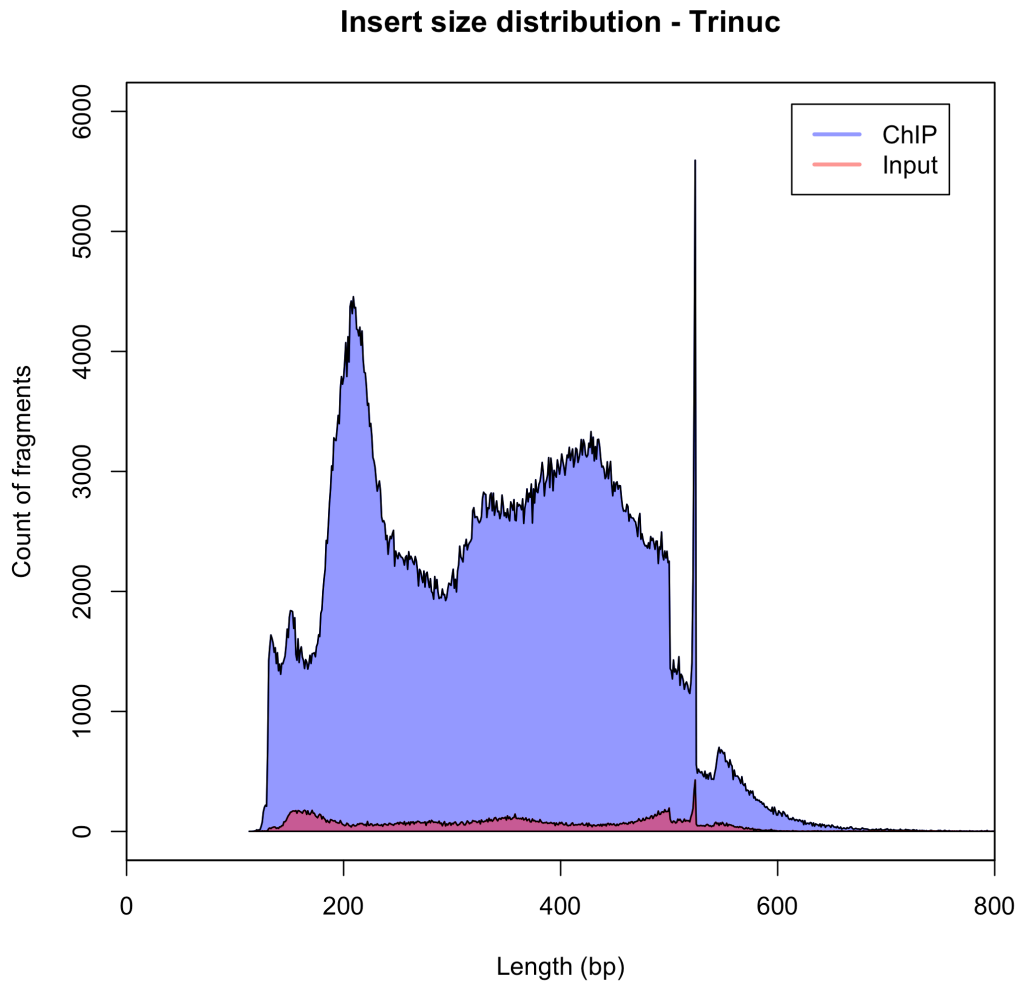


Figure 4.8 Fragment size distribution Trinuc
Insert size metrics test performed on Trinuc dataset using *Picard tools*. Size distribution in ChIP sample (blue) shows a wide range of CENP-A fragments between ~130 bp and ~600 bp, with a spike in signal at 524 bp. The spike presumably corresponds to trinucleosomal DNA. The input sample (red) also shows a broad distribution of fragments. Size metrics calculated from reads mapped to centromere regions.

A FRiP analysis (described in Chapter 3) was also carried out on the aligned native data. FRiP analysis performed on the Mono1 data indicate this ChIP experiment was successful as a value >=1% is observed. The Mono2 dataset however is slightly below the desired range which reflective of the qPCR data (Figure 4.2) and fragment length distribution (Figure 4.7). FRiP scores for the Trinuc fraction are displayed in Appendices and they show a high value of ~9.6%(Appendix II-Figure 8.27).

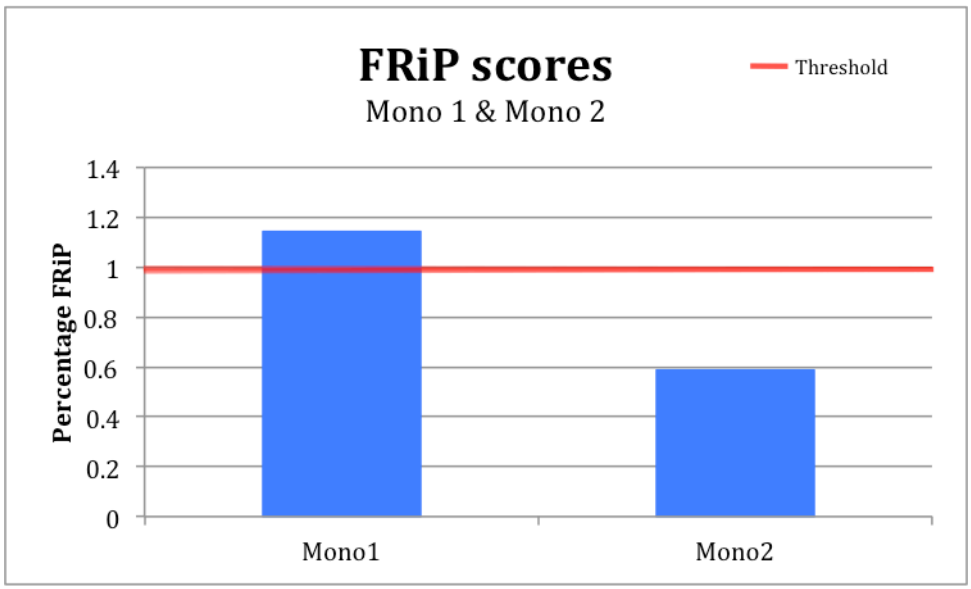


Figure 4.9 FRiP - Native data
FRiP (Fraction of reads in peaks) analysis of the CENP-A native datasets Mono1 and Mono2. Mono1 has a value of ~1.18% and therefore above the desired threshold of 1%. Mono2 is approximately 0.6% which indicates lower signal to noise ratio in immunoprecipitation.

4.5.2 Native CENP-A distribution

In order to investigate the native CENP-A binding configuration, aligned datasets were normalised using *Deeptools* and subsequently viewed using *IGV.* The CENP-A domains occupy a span in the range of 33-138 kb. Boundaries, as before, were determined by direct inspection and this information is shown below in Table 4.3. Two clear differences can be seen in the mononucleosomal data (derived from immortalised cells) compared with the cross-linked data prepared from primary cells in Chapter 3; (1) Most of the CENP-A domains occupy a narrower footprint (Figure 4.10). As mentioned, boundaries were defined by direct inspection and by comparing the span of each centromere in both datasets, the Mono1 domains exhibited on average ~25% reduction in CENP-A footprint. (2) Some of the native CENP-A domains have a slightly different distribution when compared to the cross-linked CENP-A domains, indicating the native CENP-A domains in the immortalised cells have moved somewhat (Figure 4.10, see ECA8/EAS7, ECA13/EAS14, ECA25/EAS10, ECA28/EAS4). The differences indicated could be due to (1) the differences in preparation method for both datasets and (2) the differences in cell type.

Table 4.3 CENP-A span in native data

E.ca Chr	E.as Chr	Peak 1 Start : End		Span	Peak 2 Start : End		Span
5	16	74,873,942	74,916,400	42.46	-	-	-
6	19	14,180,503	14,252,501	72.00	-	-	-
8	7	41,982,233	42,120,238	138.01	-	-	-
9	12	31,946,543	32,031,461	84.92	32,145,367	32,226,383	81.02
11	13	46,725,913	46,822,553	96.64	-	-	-
13	14	7,222,690	7,299,947	77.26	7,379,514	7,470,276	90.76
14	9	29,616,059	29,683,337	67.28	-	-	-
17	11	16,752,091	16,848,052	95.96	-	-	-
19	5	4,925,277	5,020,866	95.59	5,039,872	5,136,895	97.02
20	8	26,418,114	26,511,155	93.04	-	-	-
25	10	8,592,905	8,689,598	96.69	8,742,354	8,827,198	84.84
26	18	22,367,198	22,447,587	80.39	22,494,906	22,528,621	33.72
27	27	19,713,144	19,801,645	88.50	19,814,492	19,890,534	76.04
28	4	12,905,416	13,002,596	97.18	-	-	-
30	30	17,733,344	17,801,009	67.66	-	-	-
X	X	26,977,756	27,052,293	74.53	-	-	-

In order to resolve these possibilities, ChIP-seq from immortalised AN was performed using formaldehyde stabilised preparations (T. Masterson) This was compared to the mononucleosome data derived from immortalised cells; i.e. the same cell source. Several differences can also be seen in this comparison. While most CENP-A domains between cross-linked and native preparations from the same cell lines exhibit highly similar binding (Figure 4.11), is it evident that the boundary signal is less prominent in the native data (Figure 4.11 – see ECA8/EAS7, ECA28/EAS4, ECA30/EAS30). After applying the same boundary inspection method, an average ~15% reduction in CENP-A footprint was observed. These data suggest that a native mononucleosome preparation provide a more accurate representation of CENP-A binding.

There is a least one example of a CENP-A domain in the native data occupying a slightly different position. This can be seen in Figure 4.12– ECA19/EAS5 – right allele. Here, in the mononucleosome data, the right allele appears to have shifted leftward.

Another observation when comparing native and cross-linked data lies in the fine peak structure. The signal distribution within the domains is much more defined, and this can be seen clearly from a 25 kb view in ECAX/EASX. Here, the cross-linked data in red profiles multiple broad peaks across the domain, however, the native data in blue show multiple subpeaks resolved when compared to the corresponding broad regions (Figure 4.13). These data suggest that the native preparation gives a more accurate definition of CENP-A bound DNA.

(Mono2 and Trinuc CENP-A profiles can be seen in Appendix II-Figure 8.28-Figure 8.29)

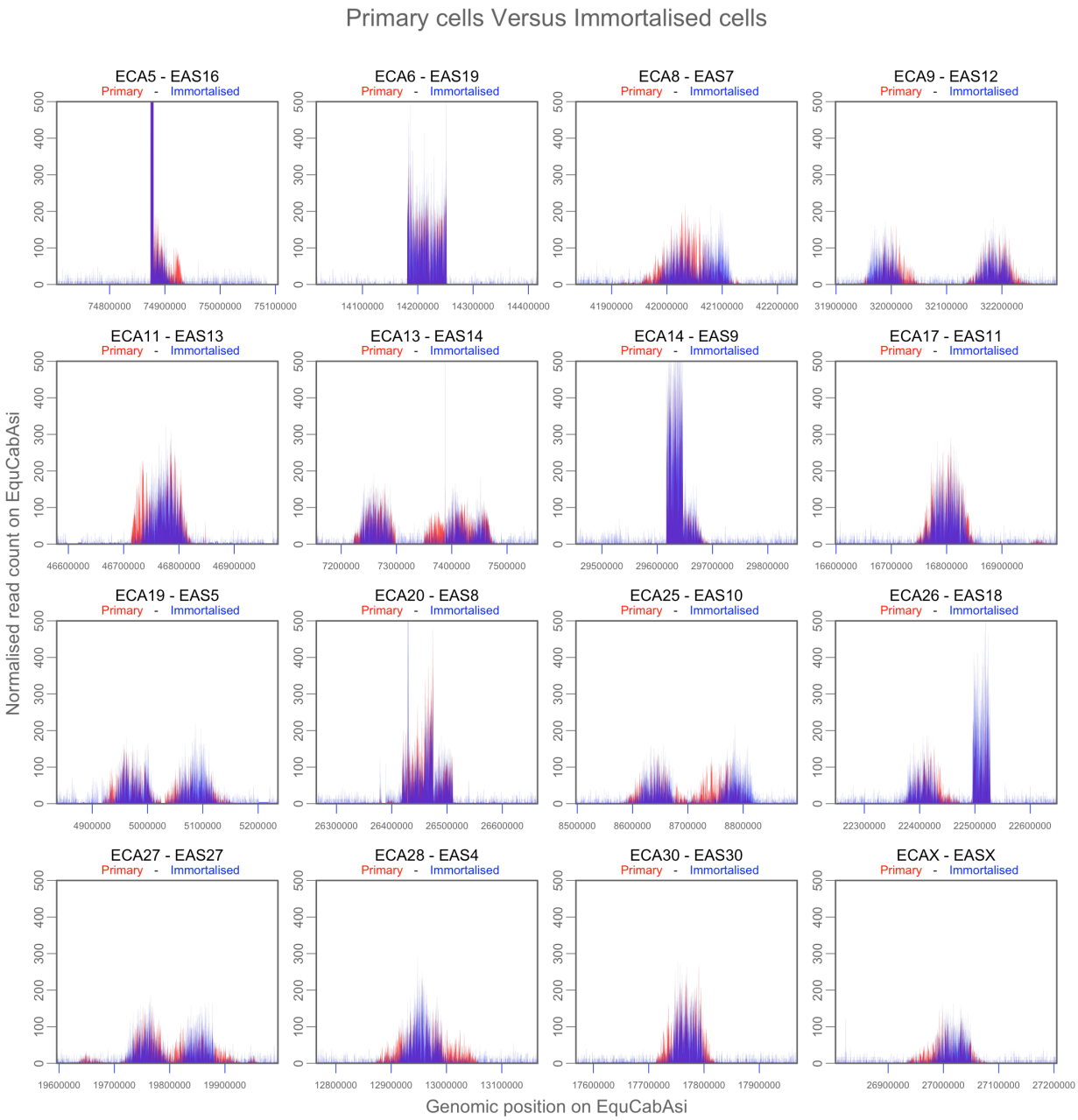


Figure 4.10 Cross-linked versus native data in primary and immortalised cells
Comparison of the native CENP-A domains versus the cross-linked CENP-A domains, prepared from two different cell sources – primary cells versus immortalised. The primary cells are cross-linked and the immortalised cells are prepared natively. Different CENP-A distributions can be seen throughout the panels indicating CENP-A positions have moved slightly after immortalisation.

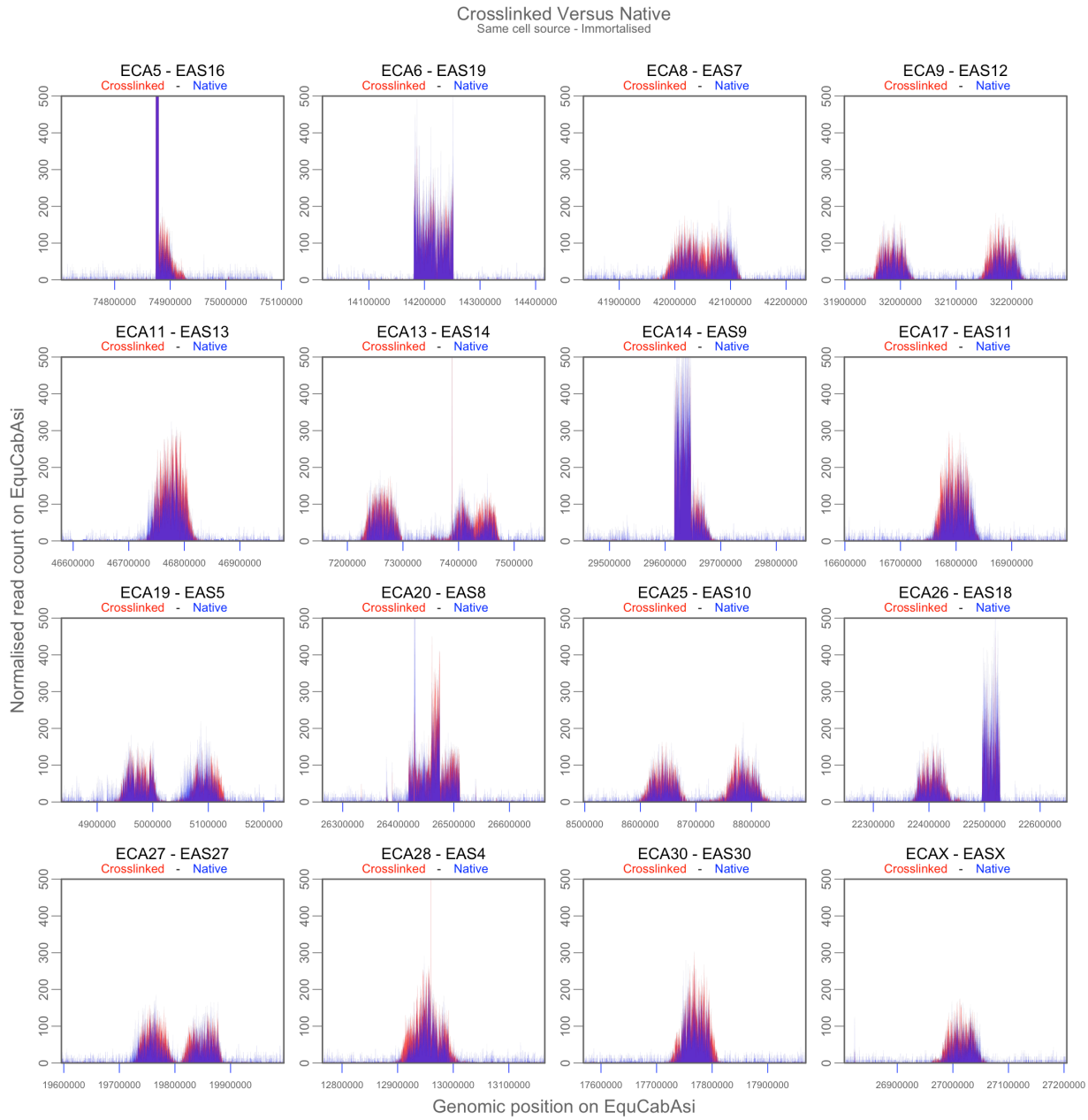


Figure 4.11 Cross-linked versus native CENP-A in immortalised cells
Comparison of the native CENP-A domains versus the cross-linked CENP-A domains, prepared from the same cell source. Cross-linked data is plotted in red and native in blue. Highly similar CENP-A distributions are observed with a few exceptions. ECA19/EAS5 shows a slightly different distribution between both datasets. The boundaries in the native data seem to be slightly narrower compared to the cross-linked data, which may be reflective of the chromatin preparation method.

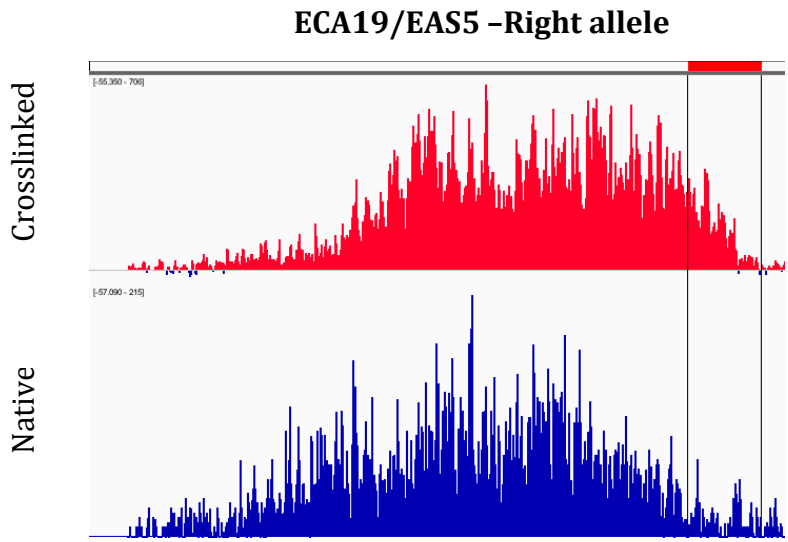


Figure 4.12 ECA19/EAS5 – Cross-linked versus Native distribution
The distribution of ECA19/EAS5-Right allele shows a slightly different distribution between cross-linked prepared data (red) and mononucleosomal ChIP-seq data (blue). It is evident that the Mono1 peak profiles slightly leftward from the cross-linked peak as indicated by the lines on the right.

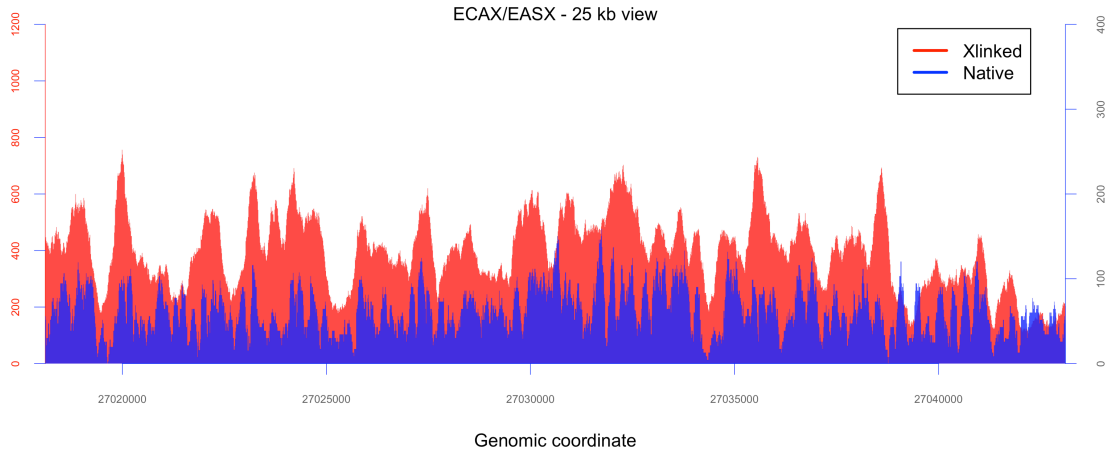


Figure 4.13 More defined subpeak structure in native data
This figure shows a 25 kb view of the mid region of ECAX/EASX. The native distribution (blue) shows subpeak finer peak structure compared to the cross-linked data (red). These data show CENP-A domains are more resolved in the native preparation.

4.5.3 Reproducibility in the ChIP

In order examine reproducibility of the ChIP data, CENP-A distribution from the independent mononucleosome datasets was compared using a Spearman rank correlation test. The *DeepTools* function *multiBamSummary* was applied on CENP-A domains that were partitioned into 200 bp bins. This function computes the read coverage per bin across genomic regions, for typically two or more alignment files (Ramírez et al., 2014). Coverage per bin for each independent CENP-A domain was then ranked and correlated using the Spearman routine in *R* (R Development Core Team, 2008). The Spearman rho values for Mono1 : Mono2 are displayed below in Table 4.4 and in most cases (nine centromeres) show a moderated positive correlation (0.5-0.7). Two of the centromeres show a low positive correlation (0.3-0.5) and five CENP-A domains show a high positive correlation (0.7-0.9). None of the comparisons show a very high positive correlation (> 0.9). While most of the correlations are in moderate-high range, concerns about the efficiency of the Mono2 ChIP are raised. The same correlative approach was applied to compare the Mono1 and Trinuc data. The Spearman values which are displayed in Table 4.4 also, show that in most cases (ten centromeres) have a high positive correlation (0.7-0.9) and four of the CENP-A domains have a very high positive correlation (0.9-1.0) with only one centromere with a Spearman value of 0.66 which is in the moderate positive correlation range. These data along with the low (0.58%) FRiP score of the Mono2 ChIP and size analysis suggest low coverage due to poor ChIP efficiency and for these reasons the Mono2 dataset was not used in further analyses. The insert size fragment analysis on the Trinuc showed a wide range of fragment sizes in the dataset, not consistent with the preservation of the original nucleosome fragment sizes. For this reason the Trinuc data were not used further in this study.

Table 4.4 Spearman values across native CENP-A data

E.ca Chr	E.as Chr	Mono1: Mono2 - ρ	Mono1: Trinuc - ρ
5	16	0.63	0.88
6	19	0.55	0.71
8	7	0.44	0.71
9	12	0.61	0.83
11	13	0.69	0.90
13	14	0.54	0.83
14	9	0.89	0.94
17	11	0.75	0.90
19	5	0.54	0.84
20	8	0.56	0.66
25	10	0.66	0.84
26	18	0.88	0.91
27	27	0.46	0.84
28	4	0.70	0.89
30	30	0.75	0.90
X	X	0.63	0.85

4.5.4 Identification of CENP-A nucleosome positions

In order to investigate centromere subunit organisation across native CENP-A datasets, nucleosome calling software was used in an effort to identify CENP-A nucleosome positions. These analyses were applied across a selection of domains; ECA9/EAS12 – right allele, ECA27/EAS27 – left allele and ECAX/EASX. These domains are single CENP-A alleles that contain DNA from a single chromosome and display a uniform Gaussian-like distribution so therefore were determined to be the most suitable candidates to identify nucleosome positions.

Numerous nucleosome calling software packages use Hidden Markov Models (HMM) (Lee et al., 2007), higher order Bayesian networks (Chen et al., 2010) and multi-layer methods to call more 'probable' nucleosome positions (Di Gesù et al., 2009). Some of these packages are only suitable for tiling arrays and remain incompatible with next generation sequencing data. Depending on the computational model, for example, in probabilistic models (HMM), nucleosome-binding locations can be ignored due to one position being less probable than the other and therefore can affect the results. The *nucleR* package uses a fast nonparametric approach to detect nucleosome dyad positions and score them based on width and height, in this way *nucleR* is able to detect multiple configurations of nucleosome binding over a given region. A noise filtering step is included using a Fast Fourier Transform (FFT) and together the approach outputs a list of nucleosome positions that can be used as a map for further quantitative studies. The stepwise process is illustrated below in Figure 4.14 and shows nucleosome positions identified across a 5 kb domain in ECAX/EASX.

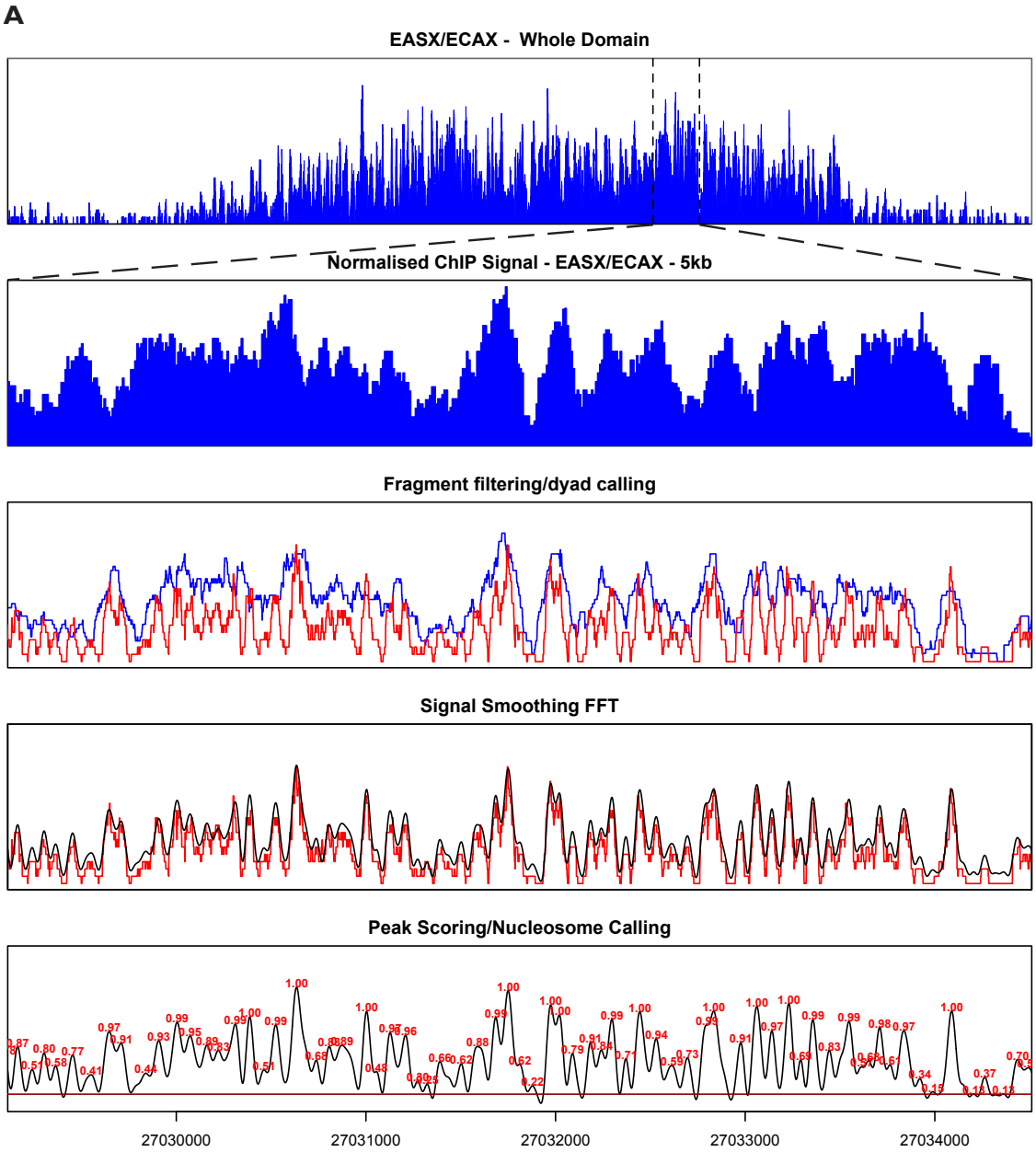


Figure 4.14 Nucleosome calling algorithm
Steps involved in calling candidate CENP-A nucleosome positions across the centromere domains. The whole ECAX/EASX domain is displayed on the top panel. The genomic interval between 27,029,311 & 27,034,311 on ECAX/EASX is shown in the next four. The ChIP signal is normalised and reads are trimmed to the dyad axis. A Fast Fourier Transform is then applied to smooth the signal and peaks are scored based on height and width revealing 'well-positioned' and 'fuzzy' nucleosome positions across the domain.

4.5.5 Quantitative analysis using *nucleR*

In order to investigate nucleosome positioning at CENP-A domains in *E. asinus* three individual centromere alleles were analysed using the Mono1 native data. The centromere alleles examined in this study spanned between 80-90 kb and the nucleosome calling function was performed on the CENP-A binding region with 10 kb either side of the boundaries to account for total CENP-A across the centromere region. The *nucleR* analysis summarised in Table 4.5 called between ~800-1000 nucleosome positions across each of the three centromere alleles cen9L, cen27R and cenX. The data observed represent a statistical distribution from a population of cells. The total number of called nucleosomes across the domains is descriptive of the population distribution by accounting for nucleosome positions that are present in subpopulations of the asynchronous cell population. We would assume that the total nucleosome numbers also contain poorly scored positions that are not representative of the preferred CENP-A positions but may be CENP-A positions in a small percentage of cells possibly corresponding to fuzzy nucleosome positions.

Table 4.5 *nucleR* summary of analysed centromeres

Centromere	Start	End	Span	#Called Nucleosomes
cen9R	32147690	32227630	79940	867
cen27L	19710800	19802220	91420	1087
cenX	26974311	27064747	90436	1003

In order to calculate a more accurate number of CENP-A nucleosomes across each domain a quantitative metric was established. Nucleosome positions identified were sorted based on *nucleR* score and integrated read counts at these positions were calculated using the corresponding alignment files. The fractional occupancy at each position was calculated by dividing the integrated read count at each nucleosome position by the total read count across the domain. It was assumed that the highest scored nucleosomes corresponded to maximally occupied CENP-A positions. Therefore if this number represented 100% occupancy, the reciprocal value would correspond to the total CENP-A nucleosomes at each centromere. To obtain this value the top three maximal occupied positions were averaged and the inverse calculated for each centromere analysed. The total numbers of CENP-A nucleosomes predicted by this method across each domain are outlined in Table

4.6. These results indicate that if the top nucleosome positions represent 100% occupancy across an 80-90 kb domain, which would correspond to ~400-500 nucleosomes, then CENP-A nucleosomes would occupy approximately 20-25% of these positions.

Table 4.6 Number of CENP-A nucleosomes at analysed centromeres

Centromere	Span	No. of Nucleosomes
cen9R	79940	104
cen27L	91420	101
cenX	90436	109

In an effort to calculate the distribution of CENP-A occupancy across the centromere domains, a threshold-filtering step was performed. Here, the nucleosome positions were sorted based on score and the top 10%, 25% and 50% of called nucleosome positions were isolated. The total signal within these thresholds was calculated by extracting read counts at the corresponding positions in the alignment data. The distribution of CENP-A occupancy was calculated by expressing total counts at each threshold as a percentage of total counts across the domain. These values are shown in Table 4.7 and a graphical representation of the thresholds is shown in Figure 4.15-Figure 4.17 The results show that the top 10% of called nucleosome positions contain ~32-42% of the total centromeric CENP-A. The top 25% of positions contain ~63-74% of CENP-A the top 50% contains up to 94% of the entire CENP-A complement. The bottom 50th percentile of called nucleosome positions only contributes a minor fraction of the total CENP-A complement. These results indicate that preferential CENP-A binding sites are present within the centromere domain.

Table 4.7 Percentage CENP-A occupancy by threshold

Centromere	Percentage of top calls	Read Count of Calls (RPKM)	Percentage CENP-A occupied
cen9R	10%	83722.3	37.03%
	25%	160710	71.09%
	50%	208123	92.06%
cen27L	10%	75176	32.46%
	25%	145800	62.96%
	50%	207541	89.62%
cenX	10%	95230.5	42.15%
	25%	167179	74.00%
	50%	212919	94.25%

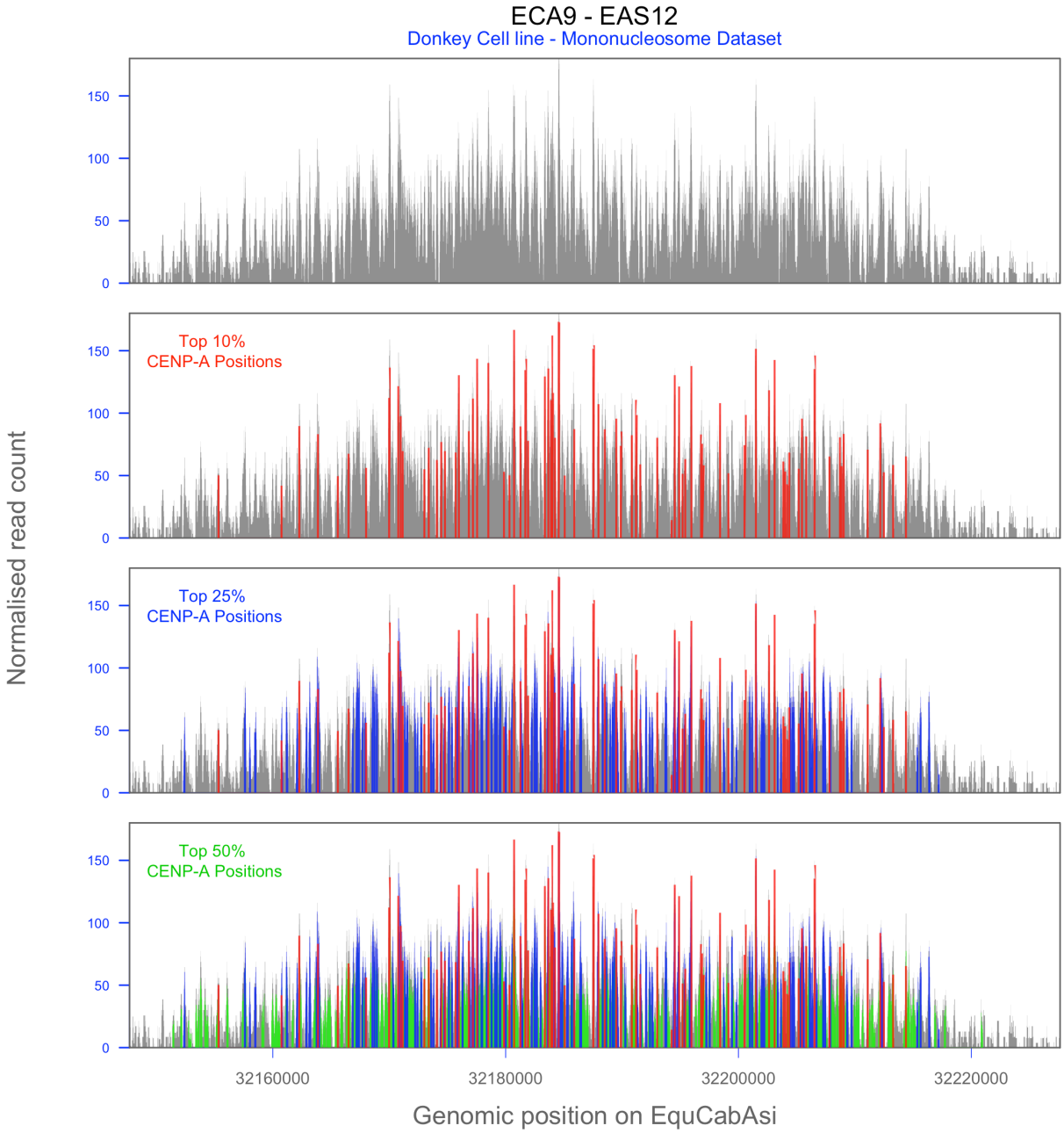


Figure 4.15 CENP-A nucleosome positions by threshold
The panel shows CENP-A nucleosome positions segmented by threshold. The top layer is the entire domain of ECA9/EAS12. The second panel displays the top 10% of called CENP-A positions. The top 25% of positions are in blue on the third panel and the last tier shows the top 50% of CENP-A positions.

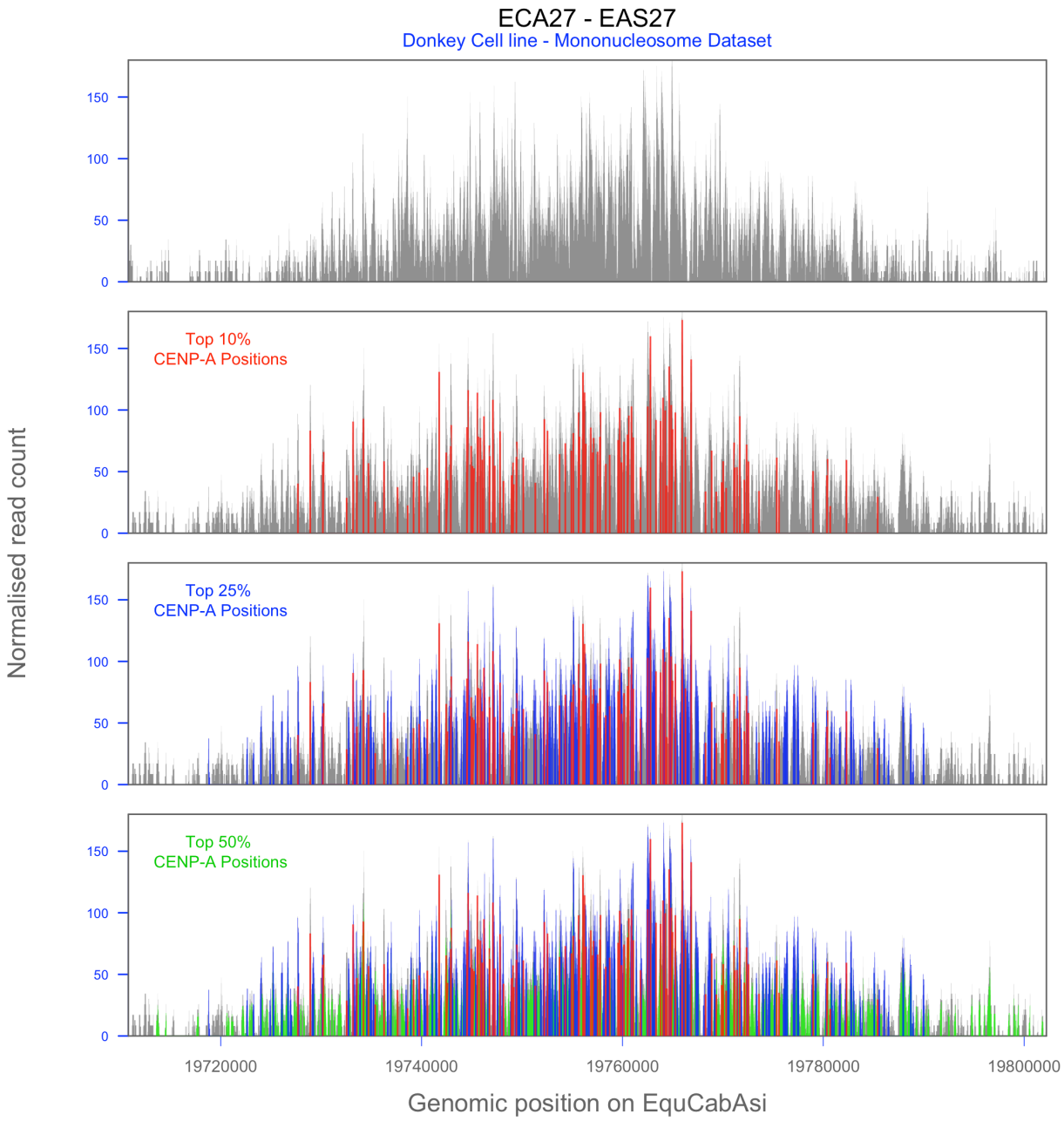


Figure 4.16 CENP-A nucleosome positions by threshold
The panel shows CENP-A nucleosome positions segmented by threshold. The top layer is the entire domain of ECA27/EAS27. The second panel displays the top 10% of called CENP-A positions. The top 25% of positions are in blue on the third panel and the last tier shows the top 50% of CENP-A positions.

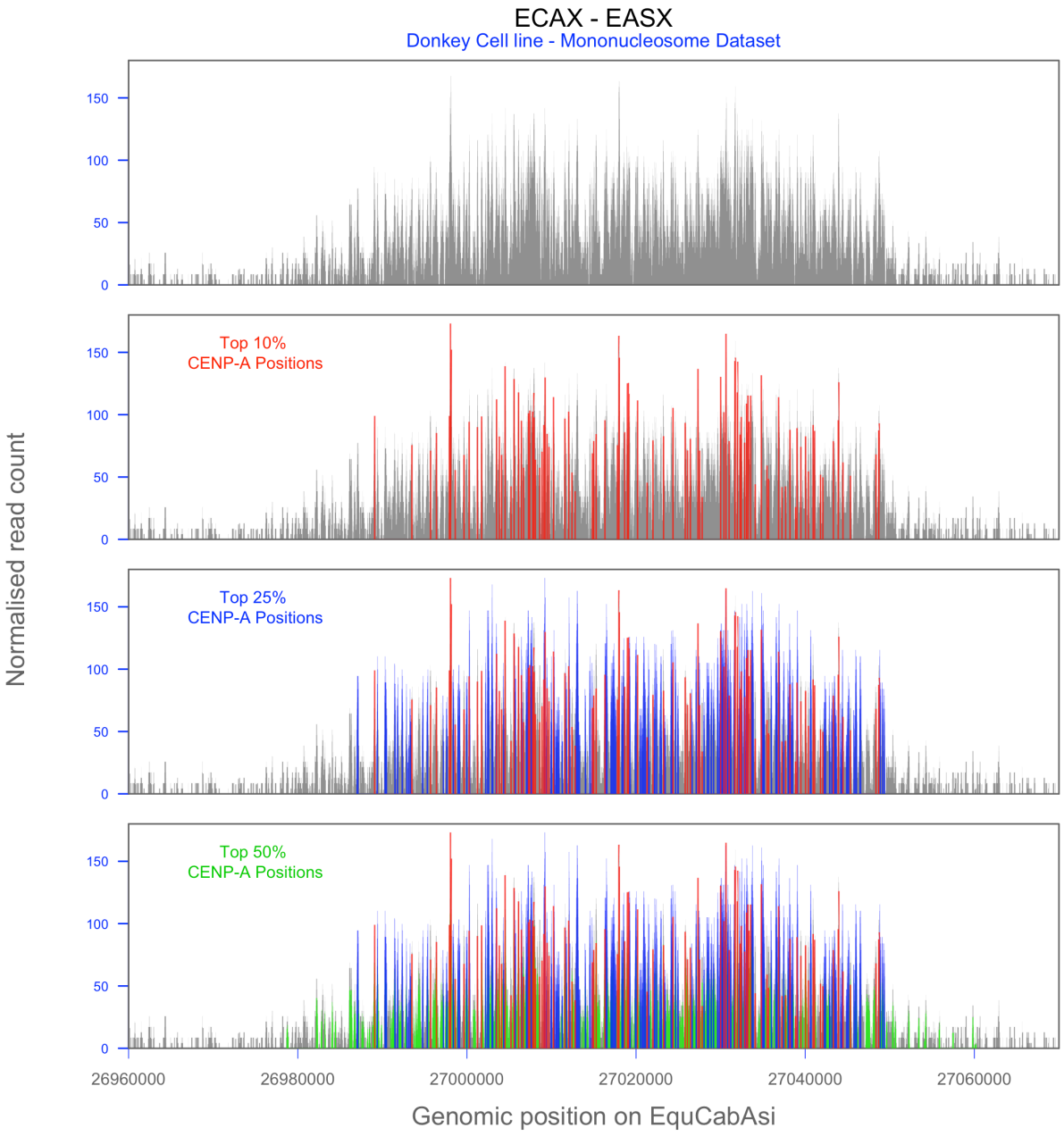


Figure 4.17 CENP-A nucleosome positions by threshold
The panel shows CENP-A nucleosome positions segmented by threshold. The top layer is the entire domain of ECAX/EASX. The second panel displays the top 10% of called CENP-A positions. The top 25% of positions are in blue on the third panel and the last tier shows the top 50% of CENP-A positions.

Although nucleosome positions can be defined from the native data, by direct inspection of the domains there was no convincing evidence of regularly space arrays. In an effort to examine periodicity further, the nucleosome positions previously identified were analysed by determining distances between called nucleosome positions. The distance between each peak summit was determined in all three previously isolated threshold fractions (Figure 4.18). Centre-centre distances are displayed using histogram distribution plots and each blue bar represents a 50 bp interval. Frequency of occurrence of nucleosome centre-centre distances in the top 10% of nucleosome positions across all three centromere domains exhibit little evidence of long range periodicity (Figure 4.18– first 3 panels). If nucleosome positioning was periodic we would expect to multiples of the nucleosome repeat, estimated at ~188-196 bp (147 bp + linker DNA) (Routh et al., 2008). In cen9R, centre-centre distances in the ranges of 50-100 bp, 150-200 bp and 350-400bp are present. Distances of 50-100 bp between nucleosomes indicate no periodicity simply because nucleosomes need a minimum of ~188 bp of DNA to position adjacent to one another (Figure 4.18 – fourth panel). These data indicate fuzzy nucleosome positions in the short range. Even in the case of CENP-A nucleosomes protecting ~120-150 bp of DNA, the distances between each centre will not accommodate for a dinucleosome array. The cen27L and cenX centromeres analysed show a peak in the 150-200 bp range, which suggests that a population of CENP-A nucleosomes are present in dinucleosome arrays for these centromeres. A broad distribution of 50-350 bp nucleosome distances can also be observed in the top 10% of called positions in cen27L and cenX. These data indicate no apparent periodicity in longer nucleosome arrays. Most of the top 25% and 50% of nucleosome distances show little evidence of regular space arrays due to an enrichment of distances falling in the 50-100 bp range (cen9-25%, 50%, cen27-25%,50% & cenX-50%). These data are shown in Figure 4.19 & Figure 4.20.

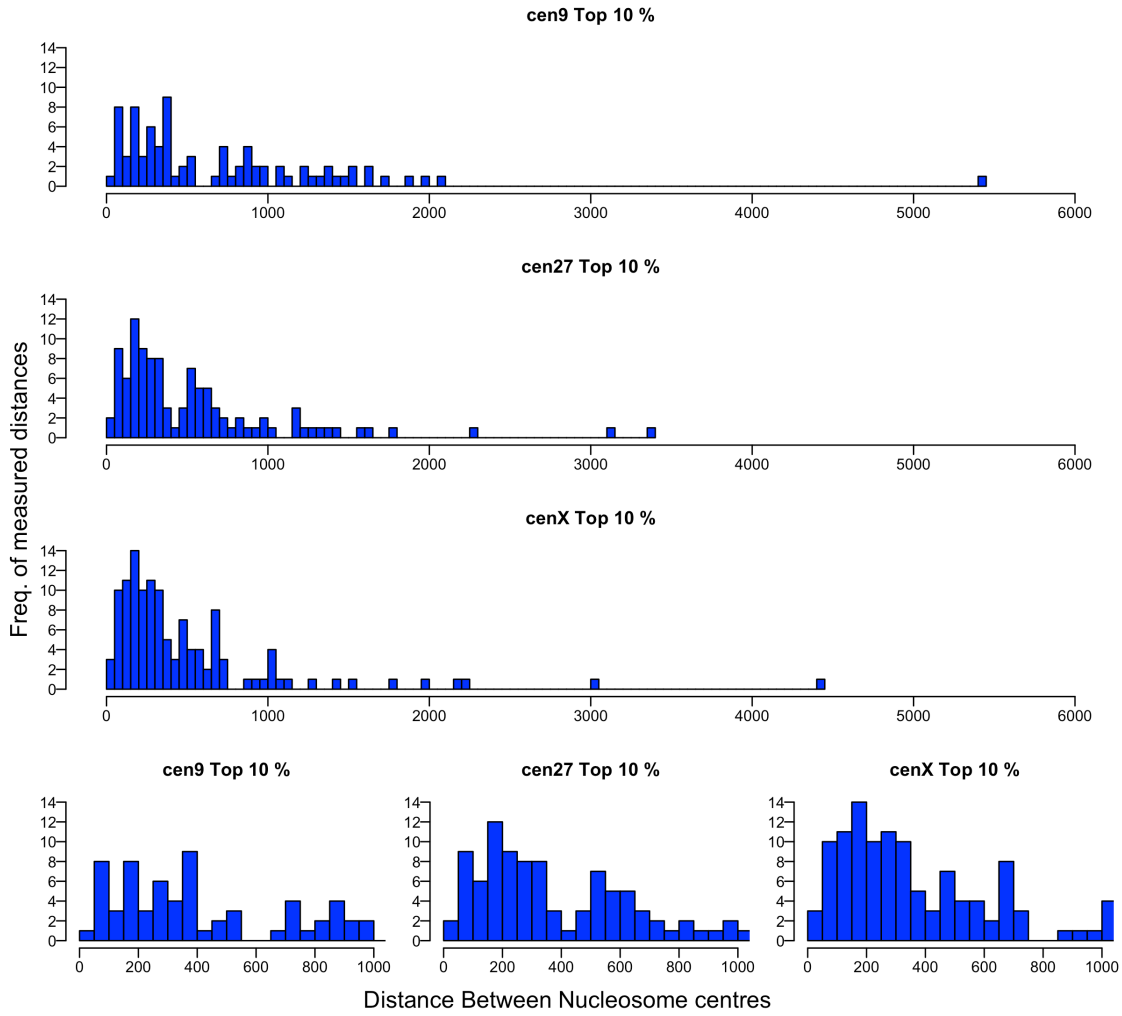


Figure 4.18 Frequency of nucleosome centre distances
Distribution of nucleosome centre distances. Centre distances defined as distance in base pairs between dyad position in called CENP-A nucleosomes, across each centromere domain. The top 10% across all three centromere domains analysed show no apparent evidence of periodicity, however, slight evidence of a spike in the 150-200 bp range in cen27L and cenX suggests a population of CENP-A nucleosomes may be present in dinucleosome arrays.

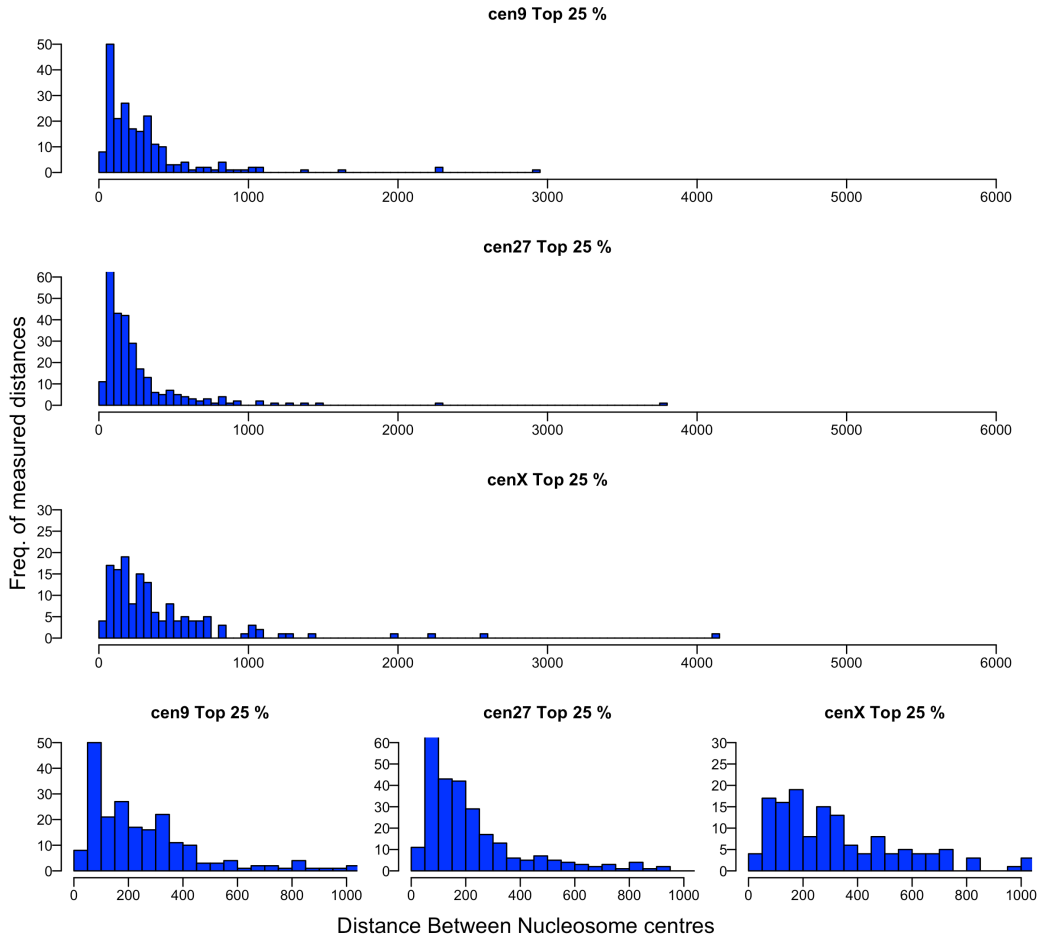


Figure 4.19 Nucleosome centre-centre distances - Top 25%

The top 25% of called CENP-A nucleosome positions across all three centromere domains analyzed show no apparent evidence of long/short range periodicity. The top 25% show that the majority of nucleosome centre-centre distances occur between 0-150 bp, with no peaks at multiples of the single nucleosome repeat, indicating "fuzzy" positioning of CENP-A.

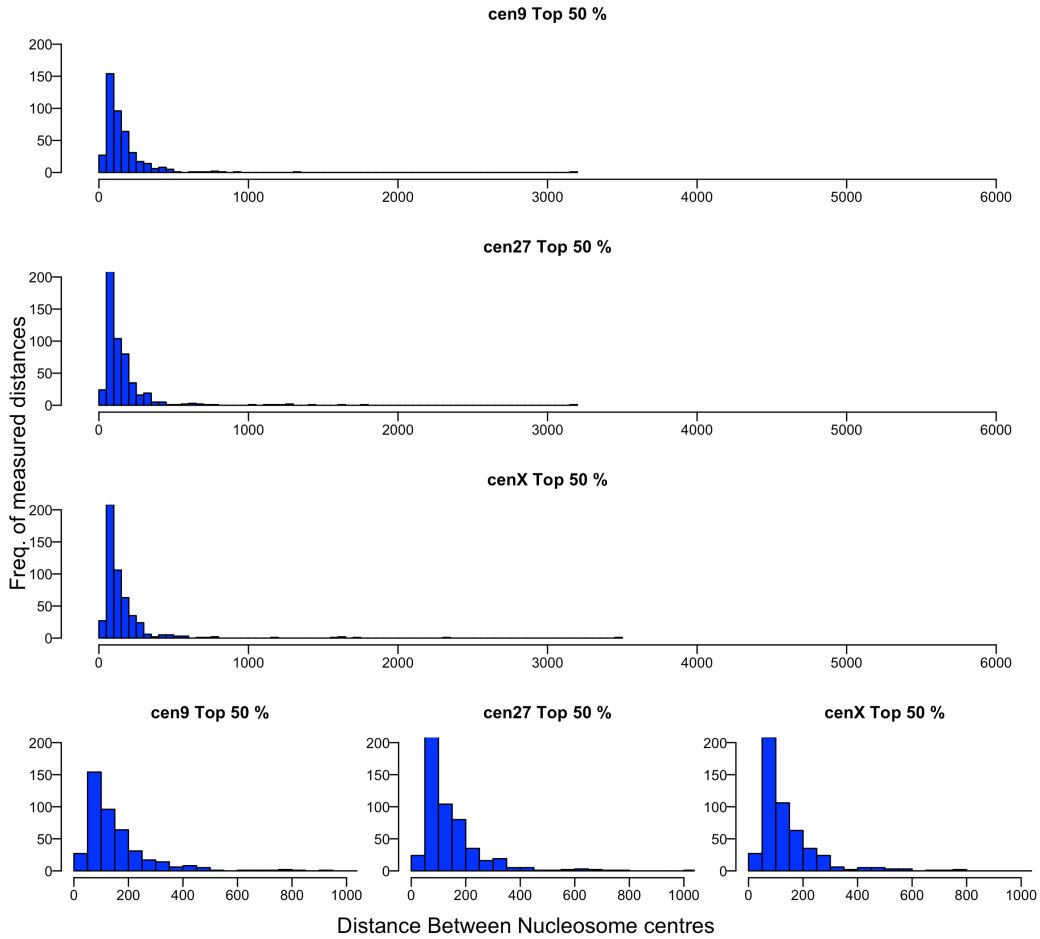


Figure 4.20 Nucleosome centre-centre distances - Top 50%
The top 50% of CENP-A nucleosome positions across all three centromere domains analysed show no apparent evidence of long/short range periodicity. The top 50% show that the majority of nucleosome centre-centre distances occur between 0-150 bp, with no peaks at multiples of the single nucleosome repeat, indicating "fuzzy" positioning of CENP-A.

These distributions represent a binned view of centre distances and therefore can only discriminate between 50 bp intervals. In order to observe the absolute distance values a rank order distribution plot was generated. Shown in Figure 4.21 are nucleosome centre distances ordered by rank. If nucleosomes were well positioned or periodically spaced on the chromatin fiber, we would expect to see signal plateaus at typical nucleosome distance spans (~180 bp, ~360 bp, ~540 bp etc.). The results in Figure 4.21 don't show any evidence of plateaued signal at these intervals but instead show a linear incremental distribution. These data suggest that nucleosome phasing does not occur at long ranges and the high abundance of 'fuzzy' positions suggest that single CENP-A nucleosomes occupy numerous DNA configurations over these genomic regions.

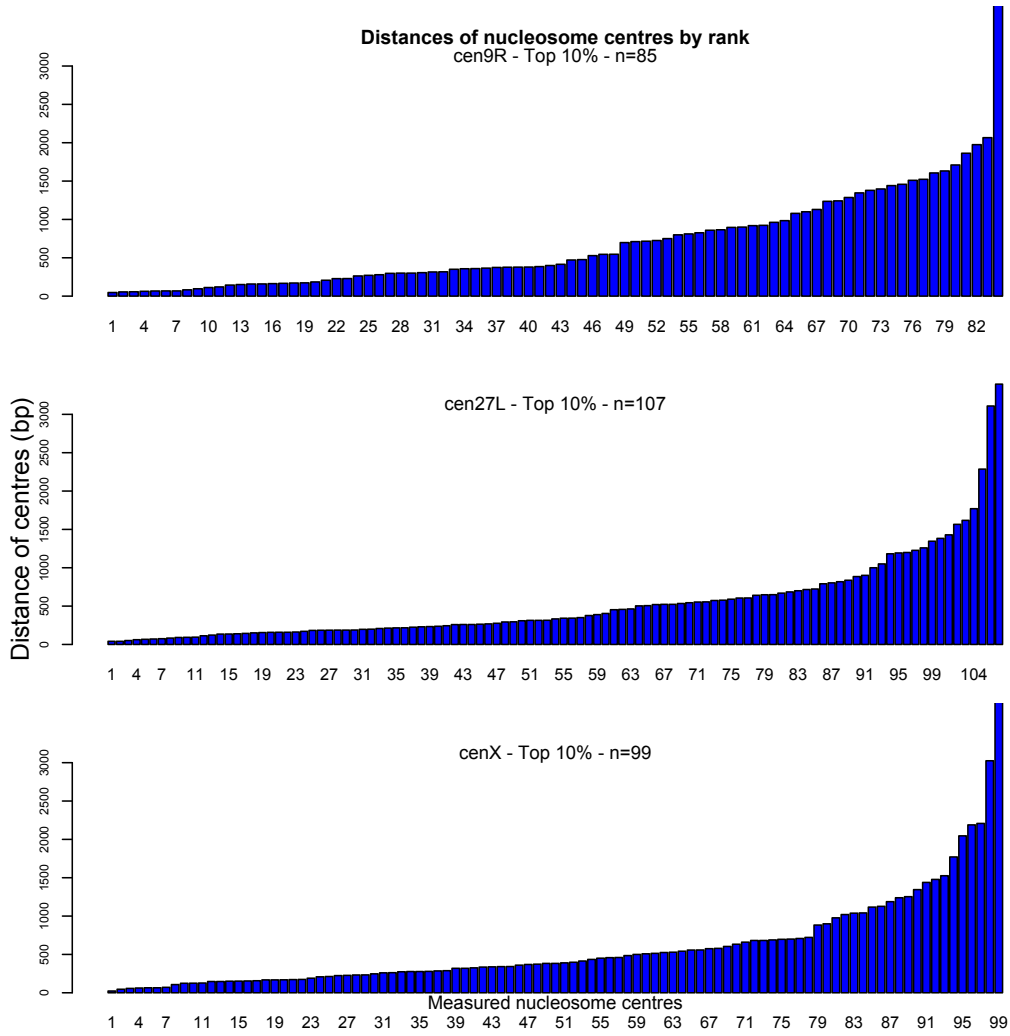


Figure 4.21 Nucleosome centre distances by rank
Nucleosome distances by rank show – x-axis shows each nucleosome centre-centre value indicated by individual bars (n=86, 108, 100). Y-axis shows the value of distance measurements (lengths between nucleosome centres.). These data show no apparent stepwise increments at expected nucleosome spacing intervals.

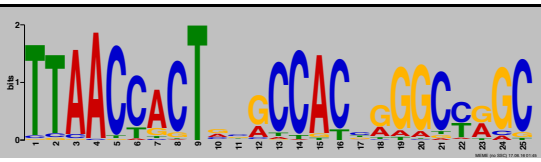
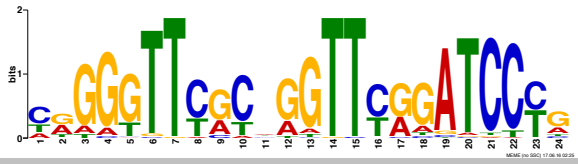
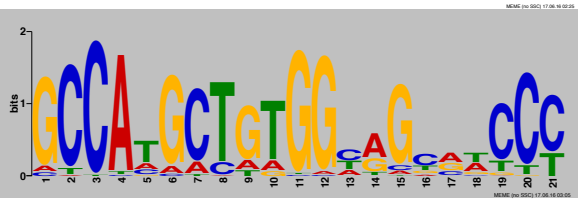
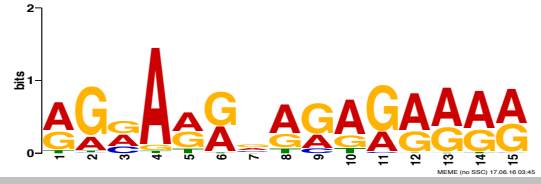
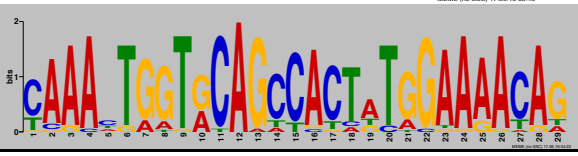
4.5.6 Motif Studies

DNA sequence has been shown to contribute to genome-wide nucleosome positioning through the binding of periodically recurring sequence motifs. These sequence motifs are typically ~10 bp periods with AA/TT/TA dinucleotides making contact with the core histones of the nucleosome and dinucleotide periods tend to configure out of phase with GC stretches (Gaffney et al., 2012).

The *E. asinus* system provides abundant unique sequence DNA in association with CENP-A. This provides an opportunity to ask if there are sequence preferences exhibited by these nucleosomes. In order to investigate DNA characteristics of CENP-A nucleosome binding a motif analyses was performed. As described in the previous section a threshold filtering step was taken to resolve the top 10%, 25% and 50% of CENP-A nucleosome positions. Here the top 10% of CENP-A positions (1716 sites) using 12 satellite-free centromere domains were isolated and analysed using the motif recognition software *MEME* (Bailey and Elkan, 1994). The 'spike-like' peaks were excluded from this experiment due to their nature of having overrepresented sequences and would such interfere with the motif searching process. The top 5 hits from the motif recognition software found sequences that were common to all tested centromere regions. No motif length limit was set in the parameters of the search. The motifs found are shown in Table 4.8 below. The E-values for the top hits are strikingly low indicating a high level of statistical significance (Bailey and Elkan, 1994). Four of the motifs found are present at all centromere domains analysed, however motif 5 is not and so is unlikely to be associated with functional or structural aspects of CENP-A nucleosomes. The most prominent motif contains a purine rich segment (motif 4). 631 of the 1716 nucleosome sites queried contain this motif which corresponds to ~37% of the total input sites. BLAST analysis showed this motif corresponded to numerous microsatellite loci specific to the genus *Equus* (data not shown). Other hits included predicted RNA transcripts (data not shown). Motifs 1,2 & 3 were only present in less than 3.8% of the total input sequences and in some instances these particular motifs were only present once at given centromere domains. Structural properties by direct inspection showed 9-10 bp spaced periods of TT (motif 2), possibly

indicating nucleosome-DNA contact periodicity. The abundance of motif 2 (65 instances in 12 centromeres) suggests it is not a major feature of centromeric chromatin. However the abundance of microsatellites suggests that CENP-A nucleosomes prefer to localise to these DNA configurations at neocentromeres.

Table 4.8 Sequence motifs associated with CENP-A nucleosomes

Hit	Motif	E-Value	Sites	Width
1		2.0e-233	58	25
2		3.1e-210	65	24
3		2.5e-169	65	21
4		3.2e-131	631	15
5		5.4e-112	24	29

4.6 Concluding Statement

This chapter describes the native ChIP-seq approach used to immunoprecipitate CENP-A associated DNA, using a CENP-A antibody, produced in-house. Western blot analysis, immunofluorescence and qPCR were used to verify that the CENP-A antibody recognises and is effective in immunoprecipitation of donkey CENP-A. The native CENP-A ChIP-seq data suggested that the native preparation gives a more accurate representation of CENP-A bound DNA. Through the use of a quantitative metric, we estimated that if the top nucleosome positions represented 100% CENP-A occupancy, across a number of centromere domains, then this would correspond to ~400-500 nucleosomes, meaning CENP-A nucleosomes would occupy approximately 20-25% of these positions. Along with this analysis we estimated that preferential CENP-A binding sites are present within the analysed centromere domains. We investigated CENP-A nucleosome phasing or periodicity in the native data by finding nucleosome dyads and calculating the distances between them. These data suggested that nucleosome phasing does not occur at long ranges and a high abundance of 'fuzzy' positions where present. The data also suggest that CENP-A nucleosomes could possibly occur in dinucleosome arrays. A motif study was also performed in this work showing four motifs that where present at all centromere domains analysed. Four of the motifs where present at all centromere regions however three of which where in very low abundance. One motif (motif 4) was found 631 times over the CENP-A input sites analysed. This motif was a purine rich segment, which corresponded to numerous microsatellite loci specific to the genus *Equus.* Collectively, these data provide more insight into CENP-A organisation at satellite-free centromeres and offer a DNA directed approach to study nucleosome organisation at centromere domains.

Chapter 5- CENP-C distribution at donkey centromeres

5.1 Introduction

Immuno-electron microscopy revealed CENP-C localisation to the inner kinetochore plate which is the interface between the centromeric chromatin and the outer kinetochore plate (Saitoh et al., 1992). As a member of the CCAN, CENP-C is expressed throughout the cell cycle (Saitoh et al., 1992), however, a progressive increasing accumulation of CENP-C transcripts is seen from S phase onwards (Knehr et al., 1996). CENP-C levels peak in G1 contemporaneously with CENP-A loading (Jansen et al., 2007; Tomkiel et al., 1994). In addition to this, CENP-C co-purifies with CENP-A nucleosomes (Foltz et al., 2006) which provides evidence for interaction of the two proteins. This interaction is mediated by the C-terminal tail of CENP-A and the acidic patch of histones H2A and H2B (Carroll et al., 2009; Kato et al., 2013). CENP-C has only been shown to be present at active centromeres and therefore is a functional centromere marker (Voullaire et al., 1993). Depletion of CENP-C causes chromosome missegregation, mitotic delay and apoptosis (Fukagawa and Brown, 1997; Fukagawa et al., 1999) and while CENP-A is the primary epigenetic marker of centromeres it should be noted that induced ectopic kinetochores by replacing DNA binding regions with CENP-C and CENP-T bypasses requirement for CENP-A nucleosomes (Gascoigne et al., 2011). Together these data enforce the importance of CENP-C in centromere-kinetochore assembly and maintenance.

CENP-C has been shown to co-localise with CENP-A through ChIP-chip and ChIP-seq approaches (Smith et al., 2011; Wade et al., 2009). A series of in-situ hybridization experiments and eventually whole genome sequencing the centromere on chromosome 11 of the domestic horse was shown to be completely devoid of satellite DNA (Carbone et al., 2006; Wade et al., 2009). ChIP-chip analysis using antibodies against CENP-A and CENP-C with a tiling array spanning the centromere region in chromosome 11, in the domestic horse, showed co-localised enrichment for centromere DNA for both proteins (Wade et al., 2009). The heterogenous repetitive nature of the centromere regions in *N. crassa* allowed for ChIP-seq studies with CENP-A and CENP-C which also revealed compete co-

localisation of the two centromere proteins (Smith et al., 2011). ChIP-seq approaches in *C. elegans* show that holocentric chromatin domains exhibit CENP-C binding at a selection of CENP-A binding regions indicating preferential kinetochore binding sites in the nematode (Steiner and Henikoff, 2014). Together these data show effectively the evidence of interdependencies between the two proteins and demonstrate the importance of elucidating the molecular organization of centromeres at the chromatin fiber level.

As described previously, the donkey genome contains 16 satellite-free centromeres, which provides a robust system to investigate centromere subunit organisation at a DNA sequence level. This chapter describes the ChIP-seq methods carried out to investigate CENP-C distribution at satellite-free centromeres in *E. asinus* and provides further insight into the spatial relationship between CENP-C and CENP-A at these domains.

5.2 Identification of CENP-C antibody suitable for Equine ChIP

In order to isolate CENP-C-associated DNA it was necessary to identify an antibody. A commercially available antibody previously used in Etemad et al., 2015; Kuijt et al., 2014; Tachiwana et al., 2015 was obtained. This antibody was first tested by western blot analysis of cellular fractions prepared from immortalised donkey skin fibroblasts (Figure 5.1-A). A number of prominent bands of were detected in whole cell extract, the cytoplasmic extract and the nuclear extract. A species just above the 98kDa marker was detected in the whole cell and nuclear fractions but this signal was not present in the cytoplasmic fraction. The same band was detected in the chromatin extract at a higher exposure (Figure 5.1-right). Human CENP-C has a molecular weight of ~107 kDa and migrates on an SDS-PAGE gel at with an apparent molecular weight of ~140 kDa (Earnshaw and Rothfield, 1985). Donkey CENP-C protein has not been previously characterised, however three isoforms of molecular weights; 112 kDa, 108 kDa & 92 kDa, are predicted to exist (Thibaud-Nissen et al., 2013). Isoform specific function of CENP-C it is still currently unknown in the equids. The prominent band seen in Figure 5.1 corresponds to the molecular weight expected of CENP-C.

Centromere specificity of the antibody was examined by immunofluorescence. Figure 5.1B shows immunofluorescence performed donkey skin fibroblasts and a mitotic cell stained with CENP-A (Cy5 - red), DNA (DAPI - blue) and CENP-C (Alexa488 - green). The double spot staining is indicative of centromere staining and clear co-localisation of CENP-A and CENP-C signals show that protein is centromere specific.

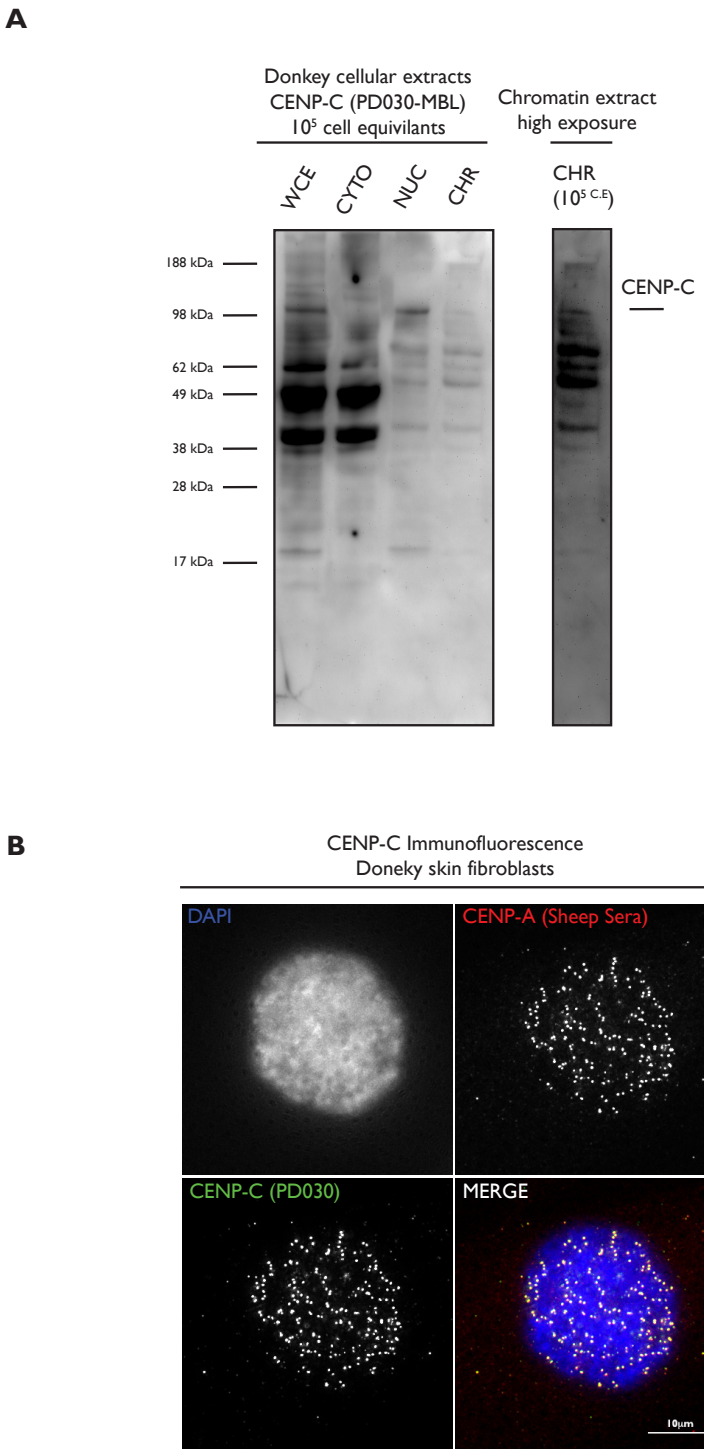


Figure 5.1 CENP-C antibody characterisation
Western blot analysis with anti-CENP-C antibody (MBL) using fractionated cellular lysates (10^5 cell equivalents per lane) show ~100kDa band present in the whole cell extract (WCE), nuclear extract (NUC) and chromatin extract (CHR)(A). Immunofluorescence on mitotic showing co-localisation of sheep anti-CENP-A and CENP-C (B).

In order to further verify that the antibody was specific to centromeres, chromatin was prepared from 10^8 donkey skin fibroblasts, by sonication, as described in section 2.2.4.8. The size distribution of fragmented DNA was examined by agarose gel electrophoresis (Figure 5.2-A). The majority of the chromatin was sheared to the desired range of 200-500 bp as indicated by the DNA size makers. (Figure 5.2-B). The ChIP experiment was set up as described in section 2.2.4.8, using 5µl of CENP-C antibody (PD030) per $12x10^6$ C.E of chromatin. Immunoprecipitation was monitored by western blot. Shown in Figure 5.2B, a band of ~100 kDa corresponding to that seen in cell extracts was observed (Figure 5.1-A).

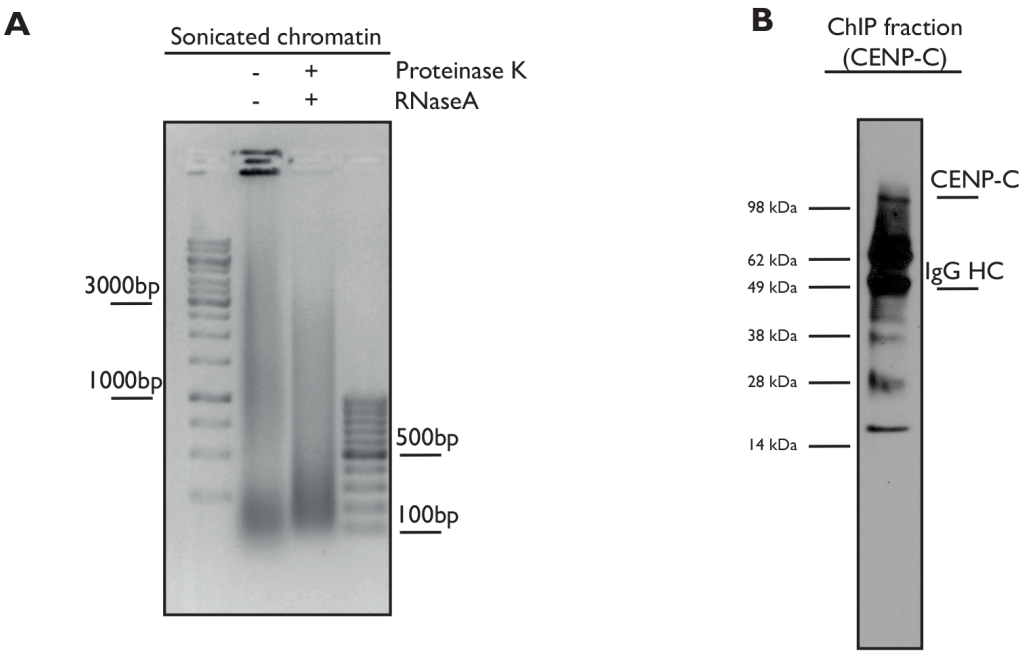


Figure 5.2 Chromatin preparation
Chromatin prepared by sonication and treated with Proteinase K and RNaseA shows DNA fragmentation in the desired size range of 200-500 bp (**A**). Western blot on ChIP fraction with CENP-C antibody (MBL) and protein-A-HRP shows signal with band corresponding to size of CENP-C (**B**).

DNA was purified from the immunoprecipitated material as described in section 2.2.4.8 and subsequently used for qPCR analysis. The CENP-C ChIP sample was probed using the three primer sets described in section 2.1.1.6. "EAS30" is a positive centromere probe that amplifies a genomic interval within the centromere of donkey chromosome 30. "ECA11" is a negative control that corresponds to a region of ECA chromosome 11 that is centromeric in the horse that but not in the donkey. "PRKC" is a region of a housekeeping gene, which is non-centromeric and therefore also negative probe for this experiment. The qPCR data show a 1.6% recovery in centromere DNA (Figure 5.3).

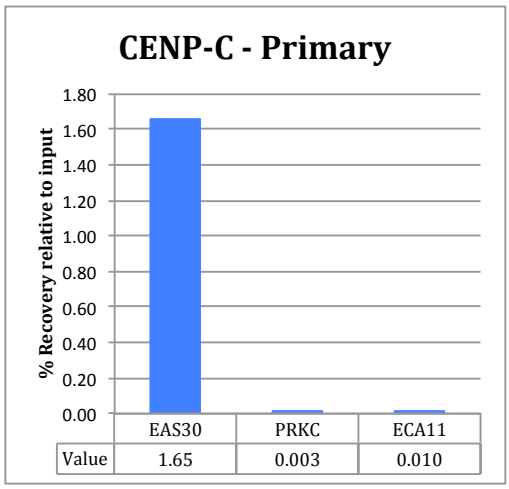


Table 5.1 Enrichment relative to negative primer probes - CENP-C

"-"Probe	ECA11	PRKC
Primary	168.90	564.18

Figure 5.3 CENP-C enriched in centromere DNA
qPCR analysis on DNA purified from CENP-C ChIP experiment shows recovery in centromere associated DNA (EAS30). Almost no recovery is seen at non-centromeric sites (PRKC, ECA11). Enrichment of centromere DNA relative to negative primer probes ECA11 and PRKC is ~169 and ~564 fold respectively (Table 5.1).

While percentage recovery of centromere associated DNA was ~1.6%, when this value is compared to the non-centromeric DNA recovery given by ECA11 and PRKC, enrichment was ~169 and ~546 fold respectively. Together these data confirm that the anti-CENP-C antibody is suitable for immunoprecipitation and highly specific for centromeres.

5.3 CENP-C ChIP-seq preparation

In order to examine CENP-C distribution across the donkey satellite-free centromere domains, ChIP-seq was performed. Three independent CENP-C ChIP experiments were carried out. One experiment was performed using chromatin prepared from primary donkey skin fibroblasts (Figure 5.2 & Figure 5.3) and the two independent replicates were performed using chromatin prepared from immortalised donkey skin fibroblasts (Figure 5.4) Due to the efficiency of the initial CENP-C ChIP, the independent replicate CENP-C ChIP experiments were processed with an 80% reduction in input chromatin. The qPCR data from the replicate experiments outlined in Figure 5.4 shows a 0.6-6% recovery in CENP-C associated DNA (Figure 5.4). (Note - The immort2 ChIP was carried out using slightly different buffer conditions as outlined in Chapter2).

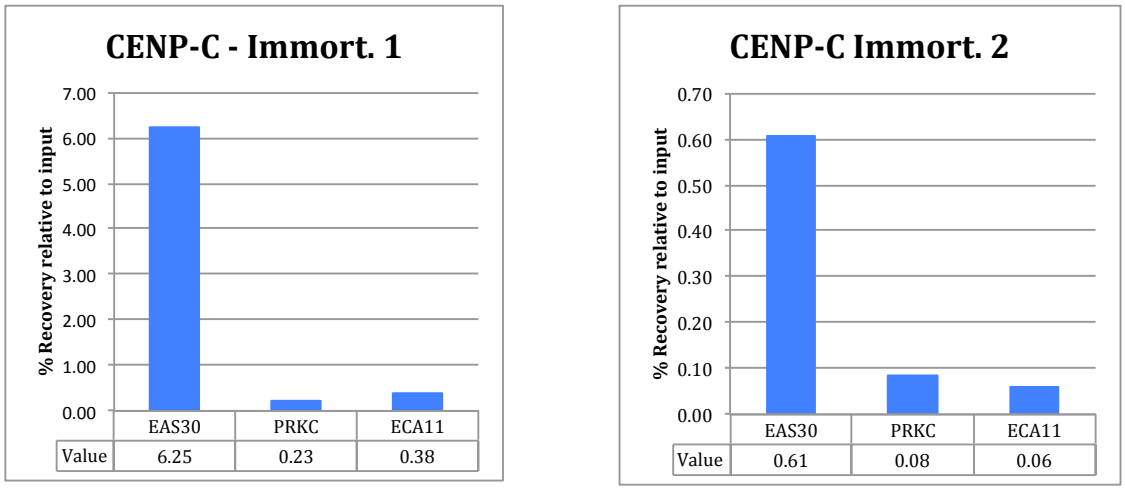


Table 5.2 Fold enrichment relative to negative primer pairs

"-"Probe	ECA11	PRKC
Immort. 1	16.53	26.60
Immort. 2	10.41	7.19

Figure 5.4 Centromere associated DNA recovered in CENP-C independent replicates.
qPCR analysis shows 6% and 0.6% recovery of centromere associated DNA in ChIP performed on immortalised cells (Immort. 1 & Immort. 2). Enrichment of centromere DNA relative to negative primer probes ECA11 and PRKC for immort.1 is ~16.5 and ~26.6 fold. Immort.2 is lower at ~10.4 and ~7.2.

The recovery of centromere associated DNA by enrichment is shown in Table 5.2. In CENP-C immort1, enrichment of centromere associated DNA is ~16.5 and ~26.6 fold for ECA11 and PRKC. Smaller values are observed in CENP-C Immort2 of ~10.4 and ~7.19 in ECA11 and PRKC.

Purified DNAs from the immunoprecipitated samples, CENP-C – Primary, Immort1 & Immort2 were sent to *IGA technologies* for library preparation and sequencing. The samples were subject to paired-end sequencing on the Illumina HiSeq 2500 V4 platform. As before, in order to saturate CENP-C coverage, we estimated that $20x10^6$ reads would be correspond up to ~500X coverage across all target domains and so we predicted this to be sufficient for the CENP-C immunoprecipitates sent. Below, Table 5.3 outlines read output from the sequencing reaction. The quantity of read data obtained was sufficient. The read length distribution ranged between 125-130 bp.

Table 5.3 CENP-C ChIP-seq data summary

Dataset	Filename	Total Sequences	Seq. length
CENP-C Input	1_AACCAG_L006_R1_001_M.fastq.gz	11184309	125
	1_AACCAG_L006_R2_001_M.fastq.gz	11184309	125
	Total	**22,368,618**	
CENP-C ChIP	2_TGGTGA_L006_R1_001_CENPC_chip.fastq.gz	11981408	125
	2_TGGTGA_L006_R2_001_CENPC_chip.fastq.gz	11981408	125
	Total	**23,962,816**	
CENP-C (rep. 1) Input	8_TCGACAAG_L003_R1_001.fastq.gz	3182948	125
	8_TCGACAAG_L003_R2_001.fastq.gz	3182948	125
	8_TCGACAAG_L005_R1_001.fastq.gz	8197190	130
	8_TCGACAAG_L005_R2_001.fastq.gz	8197190	130
	Total	**22,760,276**	
CENP-C (rep. 1) ChIP	9_AGTGCATC_L003_R1_001.fastq.gz	2216946	125
	9_AGTGCATC_L003_R2_001.fastq.gz	2216946	125
	9_AGTGCATC_L005_R1_001.fastq.gz	11146663	130
	9_AGTGCATC_L005_R2_001.fastq.gz	11146663	130
	Total	**26,727,218**	
CENP-C (rep. 2) Input	10_TGGCTACA_L003_R1_001.fastq.gz	2814934	125
	10_TGGCTACA_L003_R2_001.fastq.gz	2814934	125
	10_TGGCTACA_L005_R1_001.fastq.gz	6495155	130
	10_TGGCTACA_L005_R2_001.fastq.gz	6495155	130
	Total	**18,620,178**	
CENP-C (rep. 2) ChIP	11_GCATAGTC_L003_R1_001.fastq.gz	4573392	125
	11_GCATAGTC_L003_R2_001.fastq.gz	4573392	125
	11_GCATAGTC_L005_R1_001.fastq.gz	5303210	130
	11_GCATAGTC_L005_R2_001.fastq.gz	5303210	130
	Total	**19,753,204**	

5.4 Quality control on CENP-C ChIP-seq data

Examination of read quality by *FastQC* revealed that the input reads R1 and R2 in CENP-C - Primary had varying distributions of Phred scores (Figure 5.5). While the quality check passed inspection, the uneven read quality distributions may have resulted in some signal loss at centromeres or misalignment to another region of the genome. In order to counteract this, a filtering step was performed before alignment, using *Trimmomatic*. The filtering step was performed in paired-end mode, meaning both read libraries R1 and R2 are filtered together, in the event that if one mate in a pair is of bad quality and dropped the other mate is also dropped. In doing so, all reads were scanned over a 4 base sliding window and any window that had an average Phred score below 15 was cut. Bases at the start or end of each read with low quality were cut and any remaining reads below 75 bp in length were dropped. The trimmed input libraries shown in Figure 5.5 show an even quality distribution after filtering while still retaining read length. The ChIP reads were processed in the same fashion in order to filter out the bad quality reads (Figure 5.5). This process was applied to the Immort1 input reads for the same reason and results show the filtered data are of better quality. The Immort1 ChIP libraries failed quality control due to the low Phred score in the R2 fraction (Figure 5.6). However, after filtering with *Trimmomatic*, both the R1 and R2 libraries contain the higher quality reads only (Figure 5.6). The Immort2 libraries were also filtered (Figure 5.7). Overall, the amount of reads dropped across all datasets was in the range of ~2.03-3.18% (Table 5.4).

Table 5.4 Percentage reads dropped after filtering

Library	Reads dropped
Primary Input	2.19%
Primary ChIP	3.18%
Immort1 input	2.03%
Immort1 ChIP	2.42%
Immort2 Input	2.55%
Immort2 ChIP	2.38%

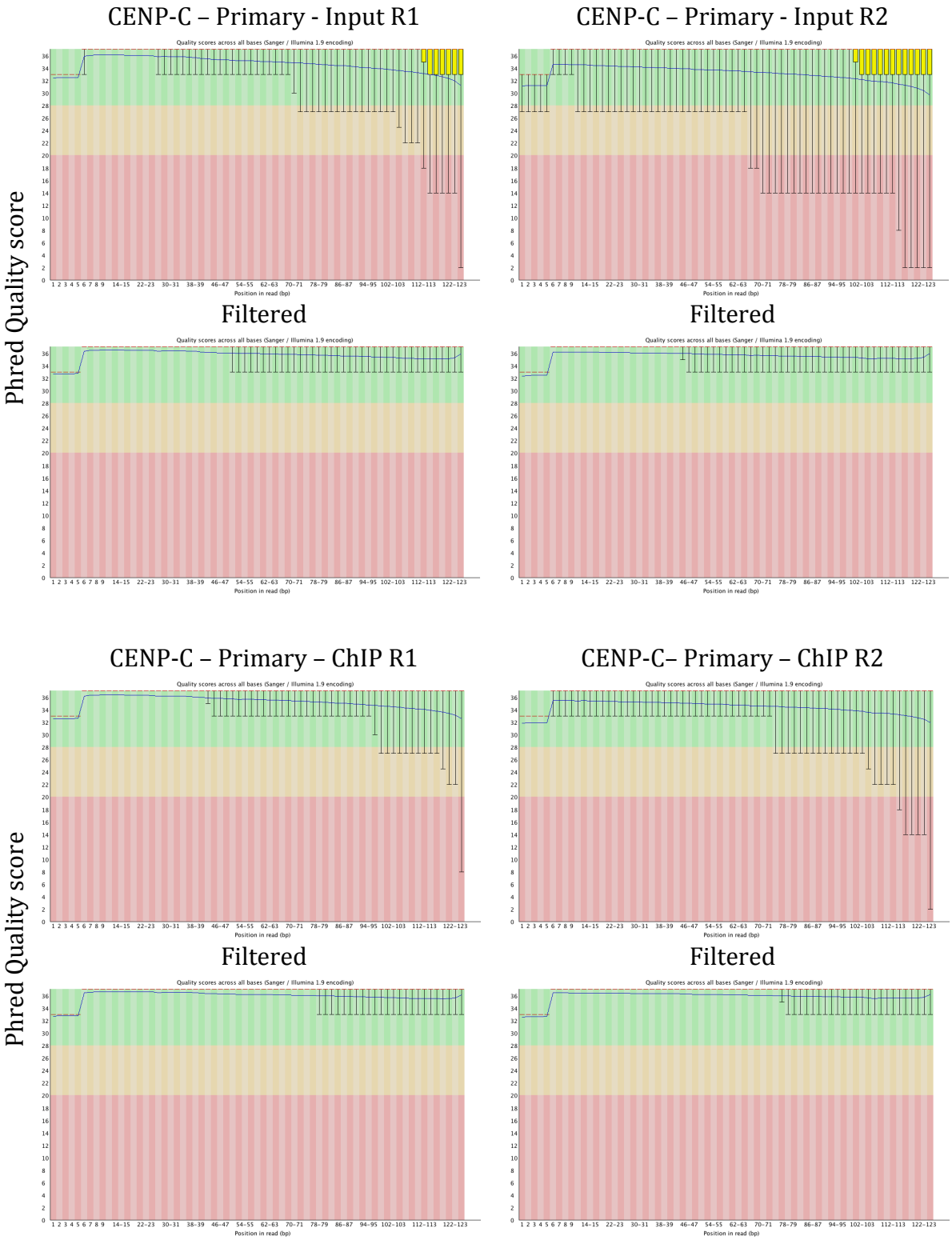


Figure 5.5 Read Quality before and after filtering - CENP-C Primary
Per base quality metrics, obtained using *FastQC*, for CENP-C (Primary cells) ChIP and input libraries, before and after filtering with *Trimmomatic*. After filtering, only the highest quality reads were retained.

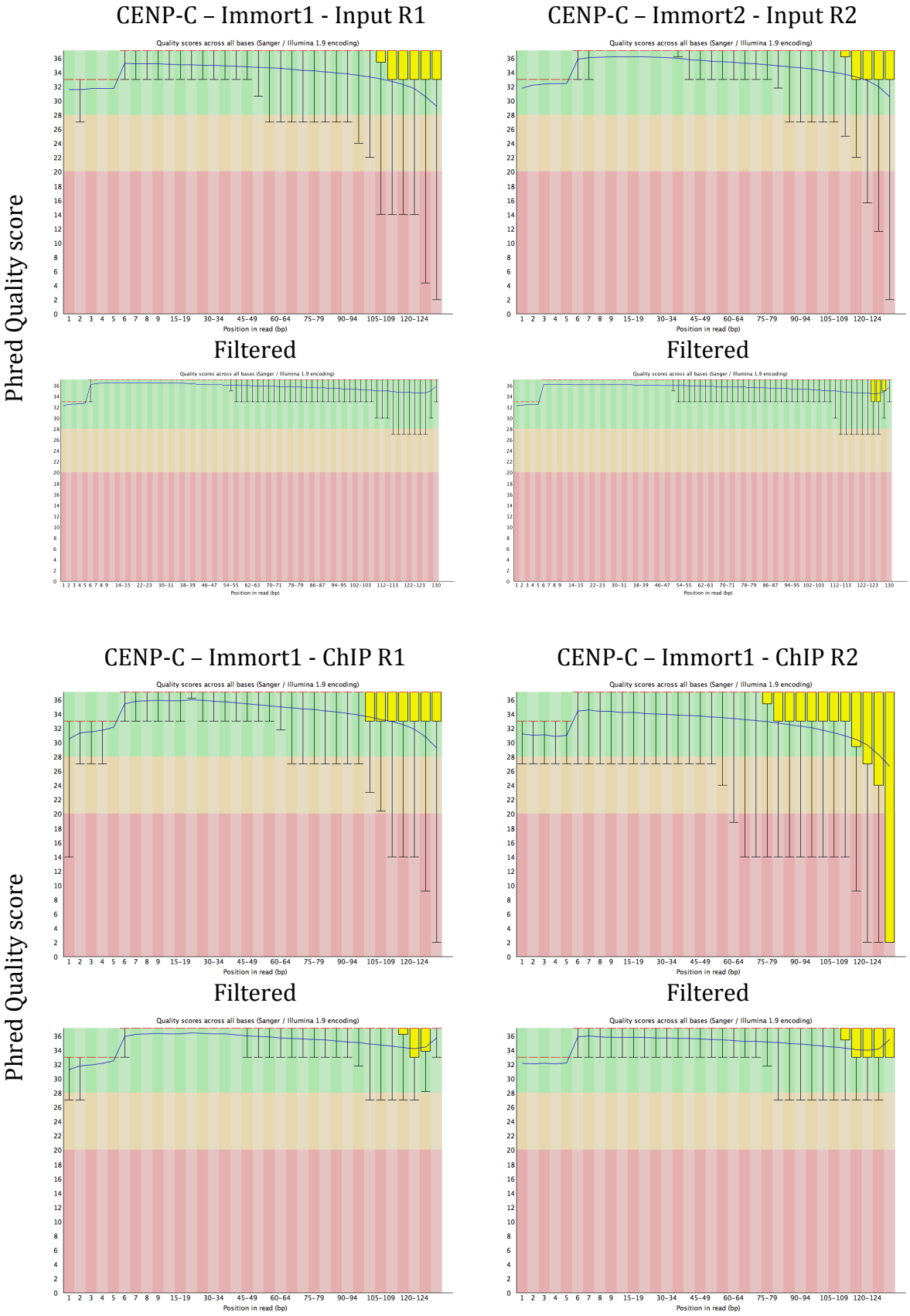


Figure 5.6 Read Quality before and after filtering - CENP-C Primary
Per base quality metrics, obtained using *FastQC*, for CENP-C (Immortalised cells) ChIP and input libraries, before and after filtering with *Trimmomatic*. CENP-C – Immort1 – ChIP R2 failed inspection and after filtering, only the highest quality reads were retained.

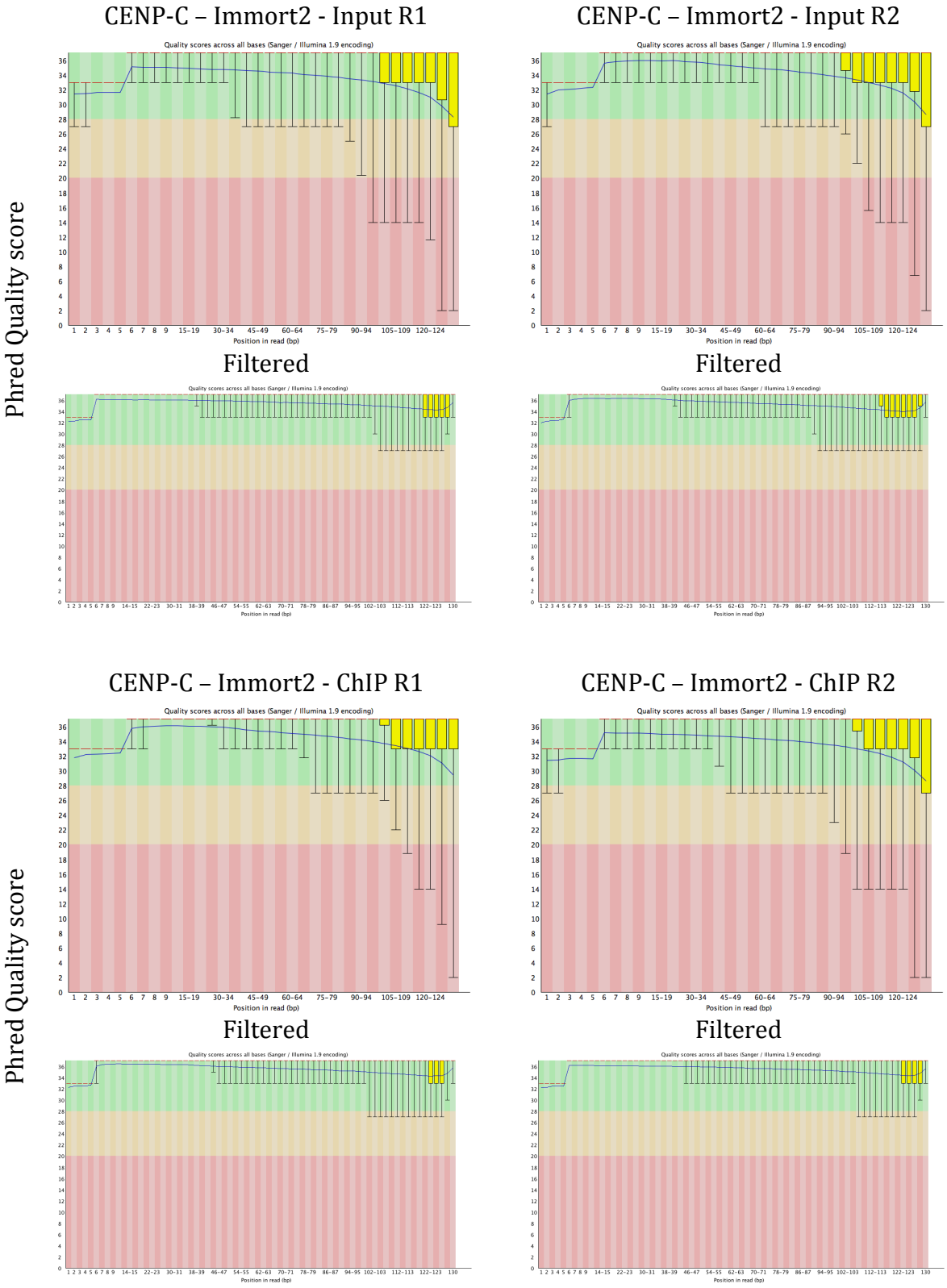
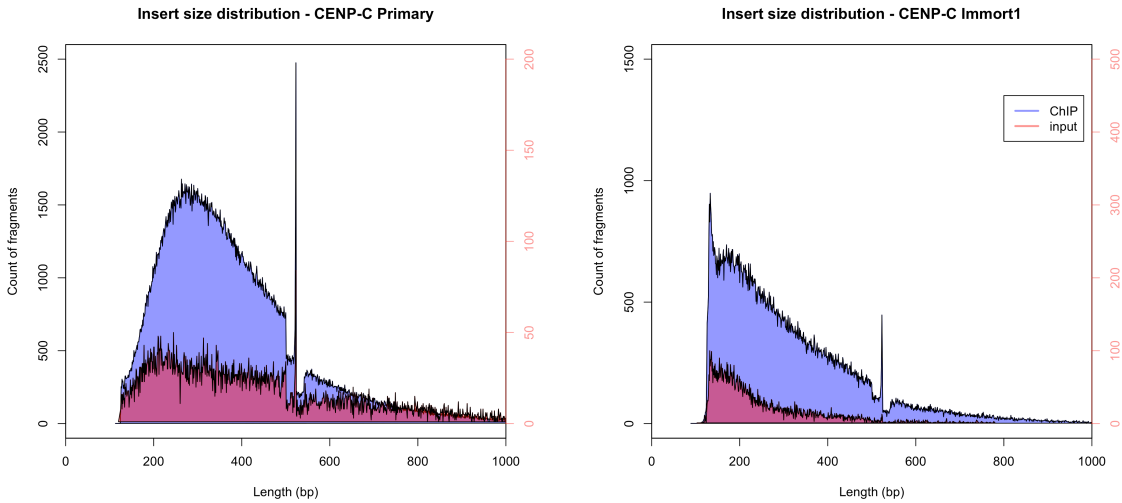


Figure 5.7 Read Quality before and after filtering - CENP-C Primary
Per base quality metrics, obtained using *FastQC*, for CENP-C (Immort2) ChIP and input libraries, before and after filtering with *Trimmomatic*. After filtering, only the highest quality reads were retained.

After the reads were filtered, all datasets were aligned to the *EquCabAsi* genome introduced in Chapter 3. In order to examine the fragment length distribution between mapped read pairs, the *CollectInsertSizeMetrics* function from *Picard tools* was performed on alignment files and restricted to the previously identified centromere regions (Chapters 3 & 4). Show in Figure 5.8(A) is the fragment lengths for the CENP-C ChIP and input datasets. The CENP-C Primary data fragment lengths are in the 200-600 bp range Figure 5.8(A). A spike at approximately 520 bp is present which is presumably an artefact of the sonication procedure. The Immort1 and Immort2 data show most fragments are in the 150-400 bp range which may indicate extensive sonication Figure 5.8(B-C).

A CENP-C - Primary **B** CENP-C – Immort1



C CENP-C Immort2

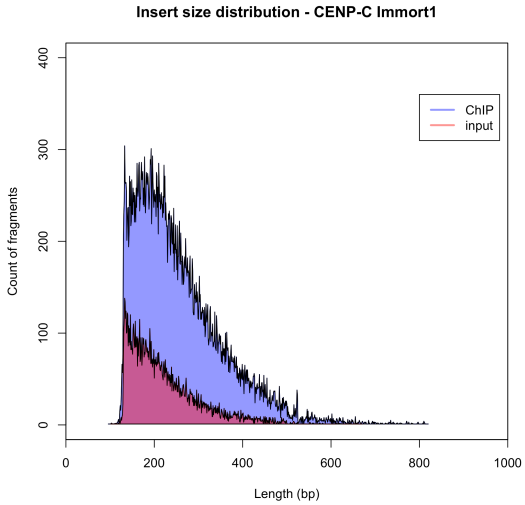


Figure 5.8 Fragment length distribution CENP-C
Fragment length function was confined to CENP-C regions and show a range of 200-600 bp for CENP-C Primary (**A**), approximately 150-500 bp for CENP-C Immort1 (**B**) and ~150-500 bp for CENP-C Immort2 (**C**).

As described previously in Chapters 3 & 4, FRiP (Fraction of reads in peaks) was used to measure enrichment of signal associated at centromeres. FRiP analysis performed on the CENP-C datasets revealed that the CENP-C Primary and Immort1 dataset are well above the desired 1% threshold indicating that these immunoprecipitations were successful (Figure 5.9). However, the CENP-C Immort2 data was just below the desired threshold at 0.72%. This value may reflect the low recovery of centromere associated DNA in Figure 5.4.

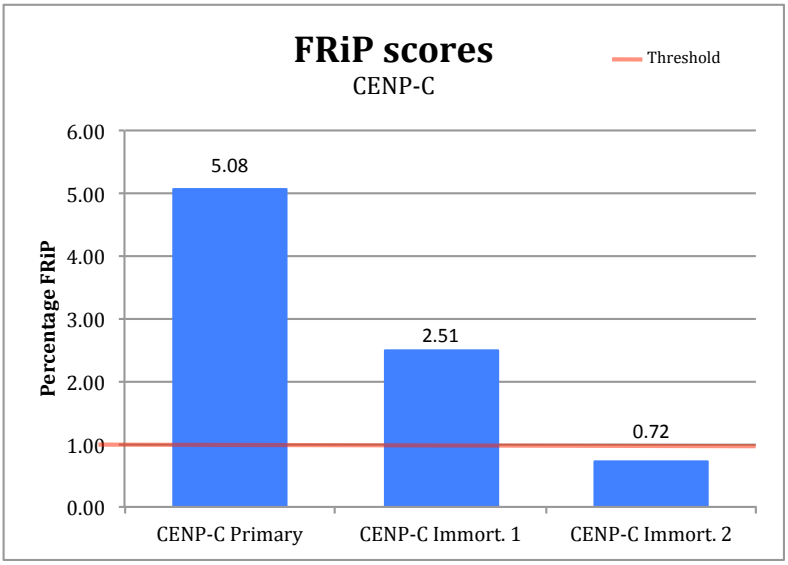


Figure 5.9 FRiP analysis of CENP-C alignment data
FRiP analysis shows 5.08%, 2.51% and 0.72% for CENP-C – Primary, Immort1 and Immort2 respectively. These analyses represent the percentage CENP-C signal that is associated with the previously identified centromere domains compared to the rest of the genome.

5.5 Visualisation of CENP-C ChIP-seq domains

In order to visualize CENP-C distribution in *E.asinus*, datasets aligned to the *EquCabAsi* genome were normalised using *Deeptools* and subsequently processed for viewing using the *R* package *Sushi*. Due to the biochemical association of the CENP-C and CENP-A (Ando et al., 2002; Carroll et al., 2010; Foltz et al., 2006; Kato et al., 2013), we expect close co-localization of the two proteins and therefore similar profile distributions. Inspection of the data revealed that CENP-C had highly similar peak structures to that of CENP-A. This can be seen very clearly in (Figure 5.10). Immort1 & Immort2 can be seen in Appendix III-Figure 8.30-Figure 8.31.

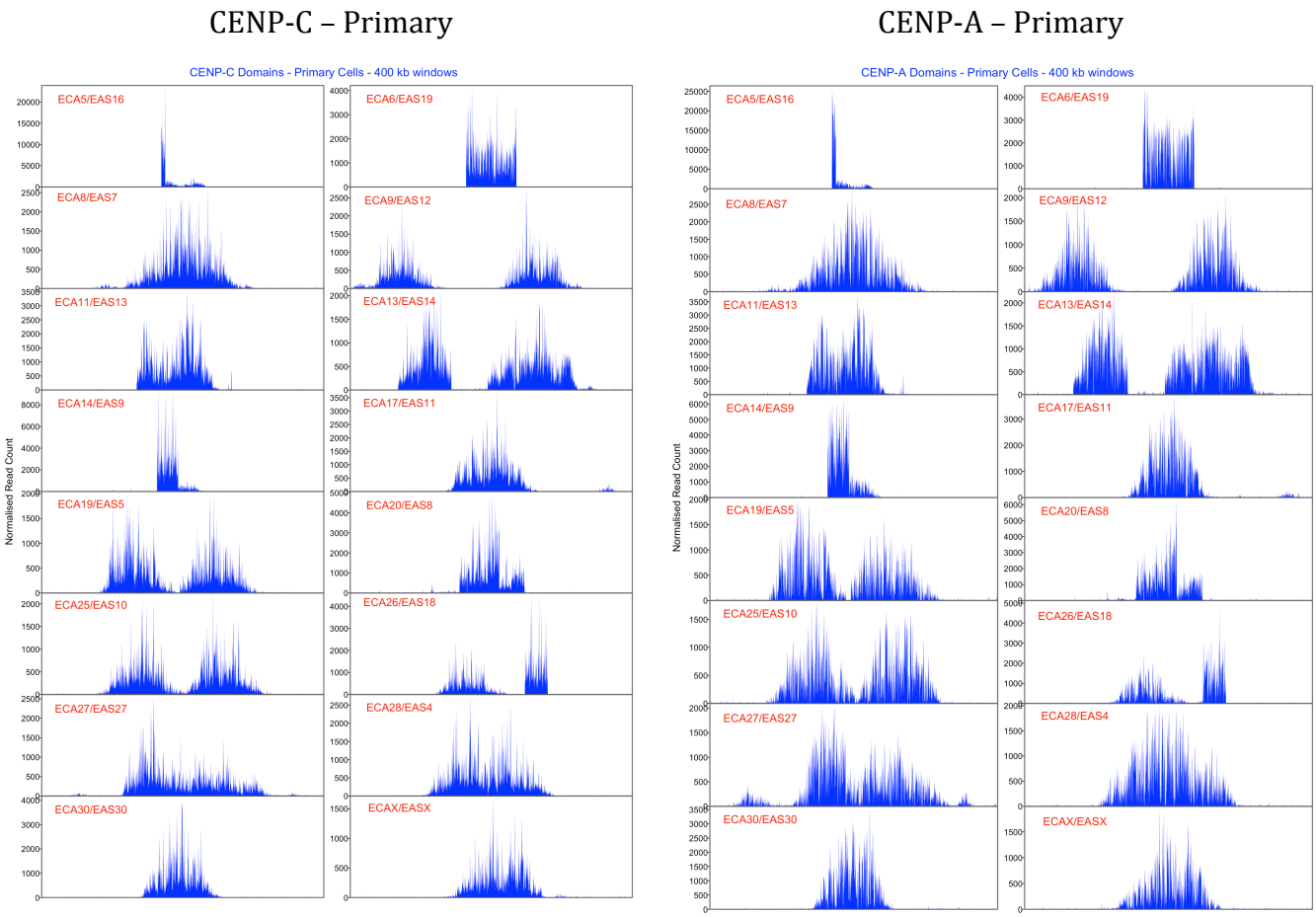


Figure 5.10 CENP-C distribution with CENP-A
ChIP-seq data from CENP-C (left) and CENP-A (right) show highly similar signal distributions at satellite-free centromeres in *E.asinus*. All centromere domains are viewed over 400 kb windows.

As shown in Figure 5.10, the CENP-C domains exhibit very comparable distributions to that of CENP-A and by direct inspection of the domains, CENP-C seems to localise to the same regions at all 16 centromeres. Through ChIP-chip, co-localisation of CENP-C and CENP-A is seen at centromeres in human and horse, (Alonso et al., 2007; Wade et al., 2009) and by ChIP-seq in *N.crassa* (Smith et al., 2011), nematodes (Kang et al., 2016; Steiner and Henikoff, 2014) and *S.pombe* (Thakur et al., 2015). The data we present here are in agreement with these findings, however, in order to examine exact localization of CENP-C relative to CENP-A we first superimposed the two datasets directly (Figure 5.11). These data clearly show that CENP-C is entirely coincident with CENP-A chromatin, within the same genomic boundaries. These findings indicate that the kinetochore-forming region is restricted to the genomic footprint of CENP-A.

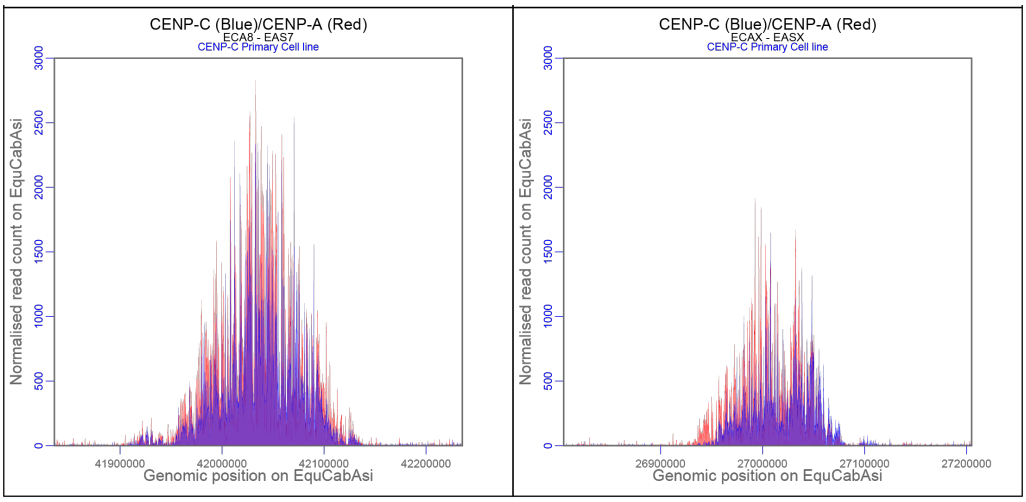


Figure 5.11 CENP-C and CENP-A profile comparison.
Superimpositions of both CENP-C (blue) and CENP-A domains (red) for ECA8/EAS7 and ECAX/EASX show clearly the distribution of signal over each domain and revealing highly similar boundaries.

In order to further examine the close relationship between CENP-C and CENP-A, a correlative approach was taken. Specifically, this approach was designed to take fine peak structure into account by comparing signal distribution within each domain. Here, using the CENP-C and CENP-A ChIP-seq data, each centromere domain was isolated and partitioned into 200 bp bins. The *DeepTools* function *multiBamSummary* was applied to the binned CENP-C and CENP-A domains. This function computes the read signal per bin across genomic regions, for typically two or more alignment files (Ramírez et al., 2014). Coverage per bin for each independent CENP-C/CENP-A domain was then ranked and correlated using the Spearman routine in *R* (R Development Core Team, 2008). Correlogram scatter plots display the results for two of the centromere domains, ECA8/EAS7, and ECAX/EASX (Figure 5.12) and show that the CENP-A and CENP-C signal is highly correlated within 200 bp windows, indicating a fine peak structure is closely related between the two domains. The Spearman values for all domains are displayed in (Table 5.5) and show that in most cases (10 centromeres) have a high positive correlation (0.7-0.89) and 5 of the CENP-A/CENPC domains have a very high positive correlation (>0.9) with only one centromere with a Spearman's rho value of 0.68 which is in the moderate positive correlation range.

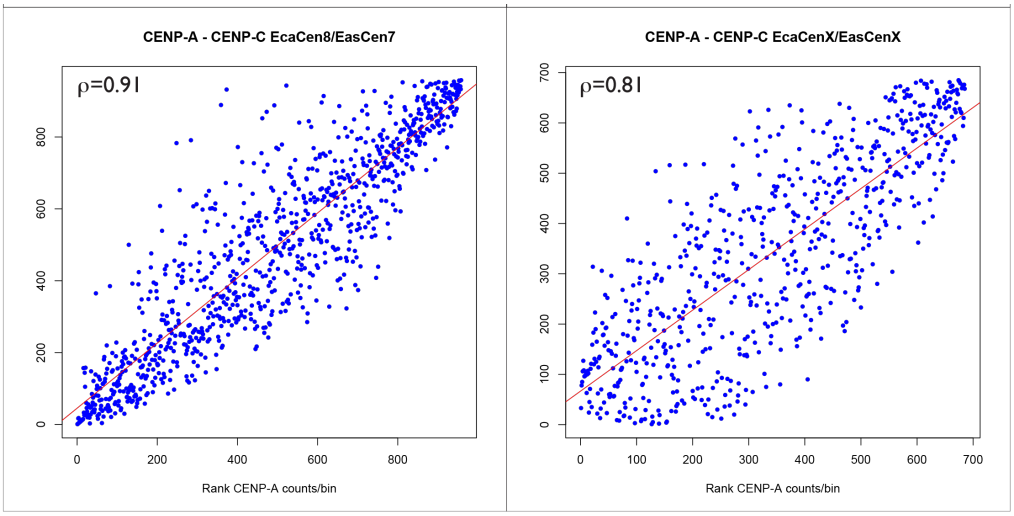


Figure 5.12 Correlative analysis of CENP-C and CENP-A domains
Rank order correlogram of CENP-A and CENP-C signal intensities by CENP-A rank. Spearman correlation test gives rho values of 0.91 and 0.81 for ECA8/EAS8 and ECAX/EASX respectively, indicating very strong association between both proteins.

Table 5.5 Spearman rho values - CENP-A : CENP-C

E.ca Chr	E.as Chr	CENP-A: CENP-C - ρ
5	16	0.68
6	19	0.85
8	7	0.91
9	12	0.89
11	13	0.91
13	14	0.83
14	9	0.96
17	11	0.86
19	5	0.86
20	8	0.90
25	10	0.81
26	18	0.80
27	27	0.85
28	4	0.84
30	30	0.91
X	X	0.81

Collectively these data show that CENP-C and CENP-A co-vary quantitatively which is consistent with co-occupancy at centromeric chromatin. CENP-A and CENP-C ChIP-seq distributions shown throughout this work represent statistical averages on a population of cells and the signal profiles could possibly indicate the presence of a unit centromere structure that occupies a slightly different position in each cell. A single cell ChIP-seq approach would be ideal to elucidate this. Within the profile distributions across the donkey centromere domains, we presume that

regions with high signal correspond to high or preferential occupancy of CENP-A or CENP-C. As mentioned previously, CENP-C is known to have DNA binding activity (Hori et al., 2008b; Sugimoto et al., 1994; Trazzi et al., 2002) and we presume the CENP-C signal observed is primarily mediated through interaction with CENP-A nucleosomes (Falk et al., 2015; Gascoigne et al., 2011; Hori et al., 2008b). In order to examine a sequence resolution view of CENP-C association and elucidate whether CENP-C demonstrates association in regions of depleted CENP-A signal, an *in silico* approach was carried out. In the previous chapter we aimed to determine CENP-A nucleosome positions within several of the centromere domains. We have used this data to investigate sequence level co-occupancy by direct inspection of the two proteins, with the same selection of centromeres. Datasets used in this analysis include the CENP-A Mono1 dataset and the CENP-C immort1 dataset, both of which were derived from the same cell source – immortalised donkey skin fibroblasts. Figure 5.13 shows the three centromere alleles analysed in Chapter 4 (Chr9R, Chr27L, ChrX) with the CENP-C immort1 dataset superimposed on the called CENP-A nucleosomes from CENP-A Mono1. Each centromere domain is displayed in blue and contains three marked regions with broken red lines. Below each centromere domain are the three marked 5 kb regions with the called CENP-A nucleosomes superimposed on the CENP-C immort1 data. We observed that many of the subpeaks across of the 5 Kb domains in the CENP-C data (blue signal) correspond with the called nucleosome positions. There are instances in which a CENP-A nucleosome is present in a region where there is a relative depletion of CENP-C signal (Triangles). There are also regions lacking clear CENP-A nucleosomes that contain high CENP-C signal (crosses). It's important to note that the both the CENP-C and CENP-A ChIP-seq data were prepared using different approaches (Crosslinked Vs Native). Taken together, these data clearly show that the majority of CENP-C co-localises closely with CENP-A nucleosomes.

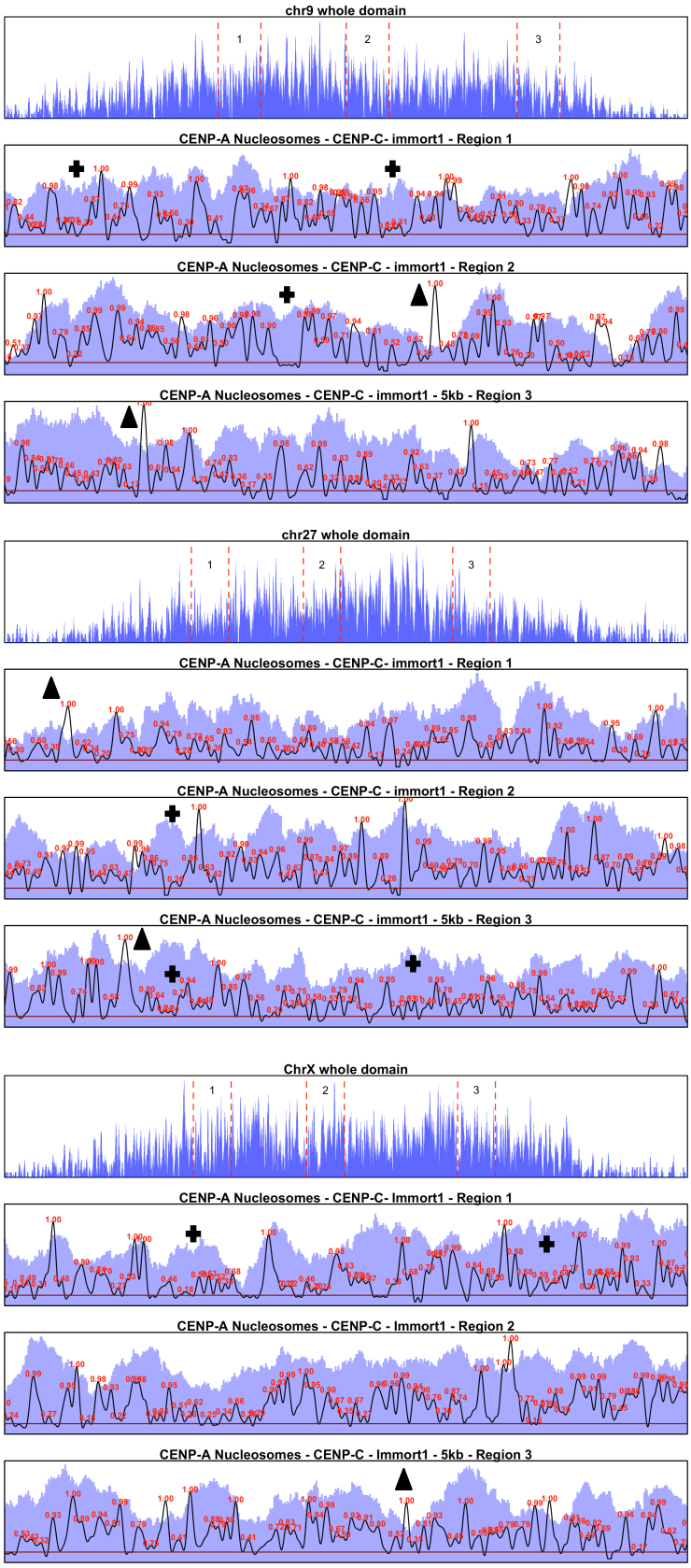


Figure 5.13 CENP-C mostly co-localises with CENP-A nucleosomes at centromeres
CENP-C ChIP-seq data and CENP-A mononucleosome data from the same cell source show close co-localisation. Triangles indicate potential regions were CENP-A nucleosomes don't bind CENP-C and crosses denote CENP-C regions that possibly don't bind CENP-A.

5.6 Relative CENP-C abundance at satellite-free centromeres

Examination of the abundance of CENP-A associated DNA at centromeres revealed a remarkable uniformity of CENP-A content at *E.asinus* satellite-free centromeres. The results suggested that CENP-A levels at satellite-free centromeres are strictly maintained. Factors that regulate CENP-A retention and replenishment have been widely studied (Jansen et al., 2007; Westhorpe and Straight, 2015) and failure to regulate CENP-A replenishment causes centromere defects and chromosome segregation (Bodor et al., 2013). CENP-B has been suggested to play a role in stabilising CENP-A nucleosomes through a dual interaction with CENP-C and the amino-terminal tail of CENP-A (Fachinetti et al., 2015). Due to the co-localisation of CENP-C with CENP-A shown in this work, we set out to determine the relative abundance of CENP-C across the satellite-free centromere domains in donkey. Using the same method in section 3.3.9, the relative abundance of CENP-C across all autosomes was determined. The results shown in Figure 5.14 indicate that like CENP-A, the levels of CENP-C are strictly maintained. The error bars denote the standard deviation of CENP-C signal at centromeres, which is a value of 0.12 and demonstrates the uniformity in the level of CENP-C across each domain. The X chromosome was treated separately as chromosome X in is haploid in these cells ("Asino Nuovo" is male) and Figure 5.14 shows that the level of CENP-C on the X chromosome contained ~42% of the signal detected in the autosomes.

The relative abundance was also measured in the replicate datasets (see Figure 5.15), ChIP-seq experiments originating from immortalized donkey skin fibroblasts. The results show uniformity in CENP-C levels, with a several outliers. The outlier in CENP-C-Immort1 shows a higher level of CENP-C at the X chromosome centromere. CENP-C-Immort2 shows a higher level in ECA5/EAS16 and ECA14/EAS9 but depletion of CENP-C in ECA13/EAS14. These results could possibly be explained by the nature of the immortalized donkey cells used in the ChIP experiment as an increased level of aneuploidy has been observed during extended culture (Teri Masterson/personal communication). These results provide both spatial and quantitative insight into the association of CENP-C with CENP-A and suggest the existence of a tightly maintained regulatory pathway for the two proteins, in the absence of DNA sequence directed binding of CENP-B.

A

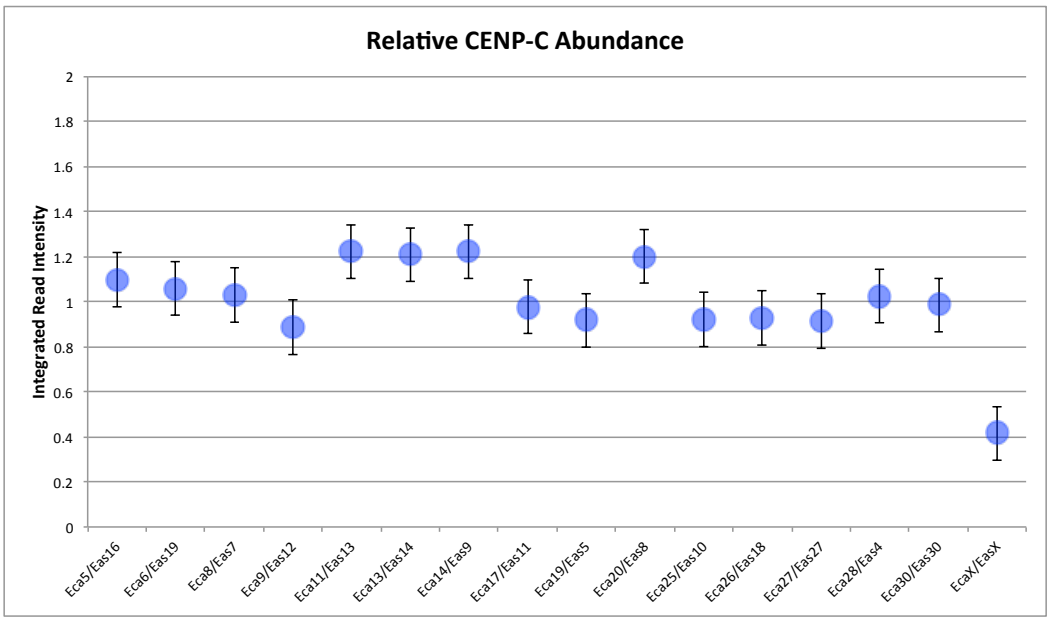


Figure 5.14 Relative abundance of CENP-C
Scatter plot showing relative abundance of CENP-C at each centromere domain. Abundance is remarkably uniform across each domain with the X chromosome 0.5X the level of CENP-C across the autosomes.

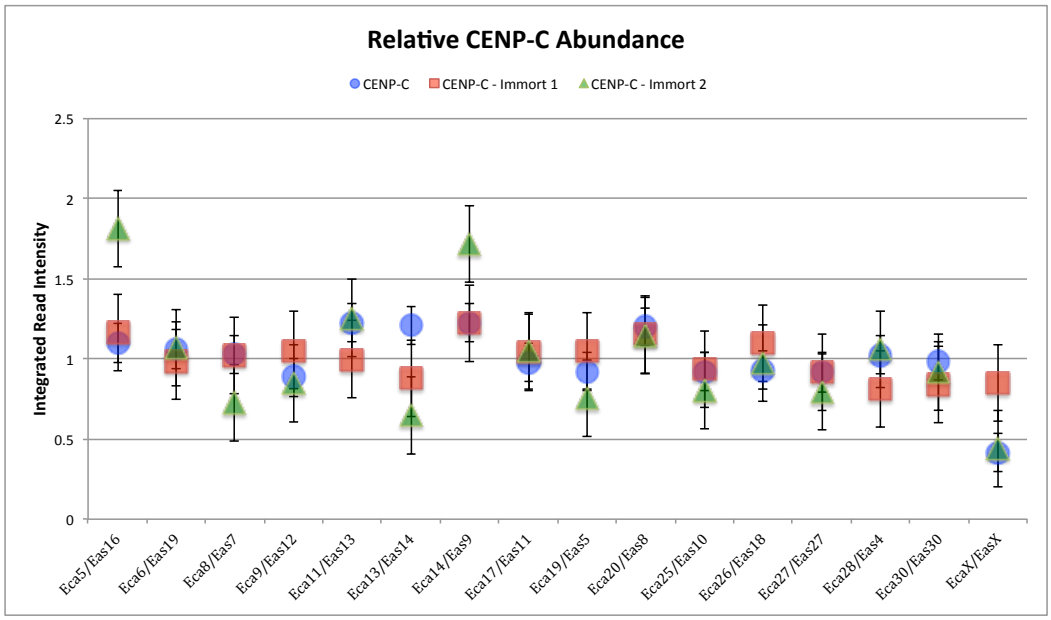


Figure 5.15 Relative CENP-C abundance all replicates
Scatter plot showing the relative abundance of CENP-C across the three CENP-C datasets. CENP-C, CENP-C – Immort1, CENP-C – Immort2. The results show uniformity across most CENP-C domains with several outliers.

5.7 Concluding Statement

This chapter describes a cross-linked ChIP-seq approach used to immunoprecipitate CENP-C associated DNA, using a commercially available CENP-C antibody. Western blot analysis, immunofluorescence and qPCR were used to verify that the CENP-C antibody recognises and is effective in immunoprecipitation of donkey CENP-C. Three independent CENP-C ChIP-seq experiments where performed which showed the successful immunoprecipitation of CENP-C associated DNA. We first presented data that clearly showed CENP-C is entirely coincident with CENP-A at centromeres and within the same genomic boundaries. In order to further examine the close relationship between CENP-C and CENP-A, a correlative approach was taken. Specifically, a Spearman's rank correlation was used to show that CENP-C and CENP-A co-vary quantitatively, which is consistent with co-occupancy at centromeric chromatin. Sequence level co-occupancy of CENP-C and CENP-A was also carried out by direct inspection. These data showed that there where instances in which a CENP-A nucleosome is present in a region of CENP-C depletion. Regions lacking CENP-A nucleosomes that contain high CENP-C signal where also observed. Taken together, these data clearly show that the majority of CENP-C co-localises closely with CENP-A nucleosomes. We then used a quantitative approach to calculate the relative abundance of CENP-C at the satellite-free centromere domains. The results showed that like CENP-A, the levels of CENP-C are also strictly maintained, providing both spatial and quantitative insight into the association of CENP-C with CENP-A and suggest the existence of a tightly maintained regulatory pathway for the two proteins, in the absence of DNA sequence directed binding of CENP-B.

Chapter 6 Discussion

6.1 Identification of satellite-free centromeres in *E. asinus*

Centromeres are epigenetic loci that direct assembly of kinetochores and are essential for the faithful transmission of life from one generation to the next. Typically defined by the presence of histone H3 variant CENP-A nucleosomes they exist in most eukaryotes associated with alpha satellite DNA (Cleveland et al., 2003). Until relatively recently with the discovery of neocentromeres and evolutionary new centromeres (ENCs) the molecular dissection of the centromeres remained quite challenging. Chromatin immunoprecipitation followed by next generation sequencing (ChIP-seq) enables the dissection of these epigenetic loci and has been demonstrated in several studies (Shang et al., 2010, 2013; Steiner and Henikoff, 2014; Thakur et al., 2015). The equids – horses, donkeys and zebras have been shown to contain satellite-free chromosomes indicating the presence of centromeres devoid of satellite DNA (Piras et al., 2010; Wade et al., 2009). Here, in collaboration with Prof. Elena Giulotto, we presented our analysis using a ChIP-seq approach that identified 16 satellite-free centromeres in the donkey.

The use of two different antibodies against CENP-A allowed two independent CENP-A immunoprecipitation experiments to be compared directly (Appendix I-Figure 8.2). Due to the lack of a donkey chromosome assembly the CENP-A ChIP-seq reads were mapped back to the to the horse reference genome. A peak calling procedure was performed using MACs (Zhang et al., 2008) followed by applying a set logical criteria allowed the isolation of 16 chromosomal locations of interest. Using the regions of interest, a set of BAC clones were identified from the horse genome library and subsequently used in FISH experiments to confirm centromere localisation (E. Giulotto). The quality control metrics of the ChIP-seq data showed the high quality of the sequence reads. A good signal to noise ratio assessed by FRiP (Fraction of reads in peaks) provide confidence in the quality of data observed. The CENP-A domains identified, spanned between 57-320 kb. This may suggest that the minimal CENP-A footprint may be in the range of 60 kb with was close to what was previously reported in (Shang et al., 2013).

Four types of CENP-A distribution were observed: “Gaussian-like”, “Spike-like”, “Multi-domain” and “Complex”. The most common type of CENP-A profile among the selection exhibited a Gaussian-like distribution (ECA8/EAS7, ECA17/EAS11, ECA28/EAS4, ECA30/EAS30, ECAX/EASX). The Gaussian-like distributions observed represent the CENP-A footprint that is consistent with the statistical distribution of CENP-A within the centromere domain over a population of cells, with higher occupancy in the centre of the peak and gradual depletion of CENP-A signal approaching the boundary regions. This type of distribution could possibly suggest that in one cell an individual CENP-A domain has a common unit structure and this distinct structure occupies a slightly different position from one cell to the next.

Several centromeres were categorised as “complex” due to their unexpected CENP-A distribution, which was much different from most of the Gaussian-like distributions we observed. ECA20/EAS8 displayed a single domain with a tripartite CENP-A signal intensity. This domain contained abrupt boundaries, which may indicate sequence differences between horse and donkey or possibly indicate a specialized heterochromatin environment in the pericentric regions at either side of the centromere. Another complex centromere was that of ECA11/EAS13. Here the signal distribution was separated by large gaps. We now know that this configuration was due to differences in the horse and donkey genome. *De novo* assembly of the donkey centromere regions and the construction of the hybrid genome *EquCabAsi* showed major DNA rearrangement events took place at the region where this centromere is located. ECA11/EAS13 in the new hybrid genome exhibited a Gaussian-like distribution.

Several centromeres (ECA5/EAS16, ECA6/EAS19, ECA14/EAS9, ECA26/EAS18) were classed as “spike-like” as they contained an intense spike in CENP-A signal up to 10 fold higher than the average peak with a Gaussian-like structure. The spike domain appears to be a region of DNA that is amplified in the donkey compared to horse and this is evident from observing a spike in signal at the same position in the corresponding input alignment files. The *de novo* assembly of donkey centromere regions using the ChIP-seq reads revealed the presence more than one copy of DNA

sequence at several spike regions. For example in the ECA14/EAS9 CENP-A domain, DNA copy number variation was indeed present over the spike-like interval. Here, approximately a 10 kb segment was tandemly repeated in 3 copies. This could possibly be a centromere in the process of obtaining new satellite-DNA.

Previous work demonstrated that the one satellite-free centromere on chromosome 11 in the horse possessed a multi-domain CENP-A distribution where CENP-A had two distinct peaks (Wade et al., 2009). In a more recent study in five unrelated horse individuals, horses that contained two distinct CENP-A peaks were shown to have positional alleles in which a centromere on each chromosome homologue occupied a different position over a ~550 kb region. A similar analysis in this study found the multi-domain peaks identified in the donkey also corresponded to positional alleles (E. Giulotto).

6.1.1 Centromere positional variation

Data presented in this study showed that centromere positional variation originally observed in the horse (Purgato et al., 2015), also occurs in the donkey. CENP-A ChIP-seq performed on another donkey individual "Blackjack" showed that each of the 16 centromere domains observed in "Asino Nuovo" were identified, but in slightly different positions. Two possibilities could explain when centromere sliding takes place; (1) the process could take place during meiosis, forming the germ cells that give rise to the offspring, or (2) during mitosis, either in the animal or in the subsequent culture of fibroblasts. In order to examine these possibilities we first presented an experiment designed to investigate positional variation in centromeres across generations. The donkey Blackjack was crossed with three unrelated horses by *in vitro* fertilisation and gave rise to three mule offspring. The mule offspring provide a haploid view of centromere domains and so we were able to follow transmission of CENP-A through generations. ChIP-seq analysis performed on Blackjack and the concepti showed two clear observations.

The first observation showed that the centromere epialleles in the paternal donkey Blackjack were independently assorted during transmission therefore suggesting that transmission events are random, proving that centromere peaks present in the parent where on different homologues. This observation begs the question is the independent assortment of the centromere alleles is entirely random? The meiotic drive or "centromere-drive" model was proposed previously. This model proposes that centromeres undergo competitive selection to be included in the oocyte (Henikoff et al., 2001). While we can't provide evidence for "meiotic drive induced positive selection" of assorted alleles throughout the 15 transmission events, one centromere showed transmission of the same allele in all three offspring (ECA26/EAS18). This observation may hint the potential occurrence meiotic drive induced assortment, however, in order strengthen this observation many more concepti would have to be analysed to increase the statistical significance of this result.

The second observation we presented here was that CENP-A domains displayed positional variation during transmission of the epialleles. In this experiment centromere positional variation was examined quantitatively by applying a peak centre of gravity based displacement analysis. We showed an absolute displacement range between 1 and 78 kb, across the 11 centromeres examined. In doing so we identified a set of high displacement centromeres (cen8, cen9, cen27, cen28, cen30). These centromeres are candidates for which exhibit movement between generations and so should be examined further. A thorough analysis of the underlying DNA sequence at these domains might provide insight into why these particular centromeres exhibit high displacement or perhaps a detailed chromatin profile looking at localisation of heterochromatin marks.

We then presented a set of analyses that examined the mitotic stability of CENP-A domains by measuring centromere displacement on populations of single cell clones, derived from an immortalised mule cell line (c1009). The absolute displacement observed was between 0-18 kb, a much narrower displacement value than what was reported in the family experiment. Its important to note that cen9 showed the highest displacement in this experiment (~18 kb), as this centromere

was categorised as a high displacement centromere in the family experiment. An ANOVA analysis to test the range of variance of displacement in clonal cell lines showed there was no significant variation in centromere behaviour supporting the observation that centromeres only subject to successive rounds of mitosis display negligible displacement.

The sliding analysis on single cell clones implies that centromere sliding occurs more frequently in cells subject to the meiosis and process of embryonic development than in mitosis. It is well established that neocentromeres lack CENP-B binding (Saffery et al., 2000) due to the absence of the "CENP-B box" motif found in repetitive satellite arrays (Masumoto et al., 2004; Muro et al., 1992; Ohzeki et al., 2002). Recently is has been shown that CENP-B helps stabilise CENP-A nucleosomes through its DNA binding capability along with direct interaction with the CENP-A N-terminal tail and CENP-C (Fachinetti et al., 2015; Fujita et al., 2015). The absence of CENP-B at equid neocentromeres could potentially play a role in centromere plasticity, however its important to note that an instance of positional variation in CENP-A domains has been demonstrated before in the form of positional alleles on alpha satellite containing centromeres (Maloney et al., 2012), where CENP-B is present. It is suggested that the evolutionary new centromeres in equids are in a young state (Piras et al., 2010) and therefore, in time, are proposed to recruit satellite repeats to their domains. The recruitment of these satellites may in turn provide a structural foundation for reducing plasticity of CENP-A domains. Fachinetti et al. (2015) also reported that the absence of CENP-B at neocentromeres resulted in an increased level of chromosome missegregation. The findings in our study suggest that the equids may have developed another regulatory mechanism to counteract the increased level of chromosome segregation, in the absence of DNA sequence directed binding of CENP-B.

6.1.2 Relative abundance of CENP-A at centromeres

A recent study (Bodor et al., 2014) carried out a series of rigorously designed experiments using fluorescently tagged proteins and quantitative fluorescent-microscopy to calculate CENP-A abundance at centromere domains. They estimated the average CENP-A abundance at centromeres is contained within a ~2.5X fold range. Our calculations using the CENP-A ChIP-seq data, estimate that the relative CENP-A abundance is well within that range, coming out at ~0.8-1.3X fold. These calculations were based on a read count yield per centromere plotted as a function of averaged integrated intensity of 15 autosomes. In doing so, the data also showed that the abundance of CENP-A on the X chromosome represented 50% of what was calculated for the autosomes, which is appropriate considering the sex of the donkey characterised is a male and therefore the X chromosome is only present in one copy. The uniformity observed may indicate that a distinct unit structure of CENP-A at these satellite-free centromeres may exist and also clearly show that centromeres are maintained without the requirement of repetitive DNA sequences. As discussed previously CENP-B has been proposed to stabilise CENP-A nucleosomes through direct association satellite-DNA and direct interaction with CENP-A and CENP-C (Fachinetti et al., 2015). Here we presented a uniform abundance of CENP-A across each centromere indicating the stability of CENP-A in the donkey satellite-free centromeres is maintained in a CENP-B independent manner.

6.2 CENP-A nucleosome distribution

In Chapter 4 we presented our analysis of CENP-A nucleosome distribution using a native ChIP-seq approach. We first optimised the native chromatin preparation procedure by titrating concentrations of S7 nuclease over a fixed time in order to identify conditions for effective chromatin digestion. We then used sucrose gradient fractionation to isolate different size classes of nucleosome arrays. Two sucrose gradients were performed. The Mono1 and trinuc samples originated from the same gradient and Mono2 was derived from another (Appendix II-Figure 8.24).

Immunoprecipitations were performed using a CENP-A antibody produced in the lab. Specificity of the antibody examined by immunofluorescence on mitotic cells and metaphase chromosomes showed discrete localisation to centromeres. qPCR analysis using previously identified primer pairs confirmed centromere association. Paired-end sequencing was performed on the mononucleosome data in order to obtain complete coverage of the DNA protecting the CENP-A core nucleosomes.

6.2.1 Quality control

The quality control metrics we presented showed two important aspects of the ChIP-seq data. **(1)** The reads for Mono1 were of very high quality providing us with confidence in the data. Another observation was presented using the FRiP analysis which showed the Mono1 dataset was above the desired 1% threshold. **(2)** We also determined the distribution of fragment lengths for all datasets and the Mono1 data showed significant enrichment of fragments in the range of 130-135 bp, consistent with the shorter DNA protection previously predicted from the crystal structure of CENP-A nucleosomes (Tachiwana et al., 2011) and also observed directly through extensive micrococcal nuclease digestion of CENP-A nucleosomes (Hasson et al., 2013). Fragment length analysis on the Mono2 data showed an enrichment in both 130-135 bp and larger than 150 bp suggesting higher background from canonical nucleosomes in the sample.

6.2.2 Reproducibility of the chip

We presented the quality control analysis on the Mono2 data showing that the low recovery and low FRiP score indicated a poor quality chip experiment. The fragment size distribution analysis strongly agreed with this due to the enrichment of longer fragment lengths, suggesting extensive recovery in non-CENP-A nucleosomes. We then carried out a correlative analysis using the three native datasets. Ranking CENP-A signal in 200 bp bins and performing a Spearman routine allowed us to compare fine peak structure at all centromeres between Mono1 and Mono2. This analysis clearly showed much lower correlation in Mono1 and Mono2 compared to Mono1 and trinuc. For this reason Mono2 was not used in further analysis.

While the trinuc chip displayed a varying distribution of fragment lengths CENP-A was shown to be highly enriched at centromere domains as indicated centromere associated DNA recovery by the qPCR. A high FRiP score of 9.8% was observed (Appendices), indicating high CENP-A enrichment. Unfortunately, the varying distributions of fragment lengths were not consistent with the preservation of the original nucleosome fragment sizes and therefore suggested possible degradation of DNA in the sample or possible PCR artefacts derived from downstream processing for sequencing (library preparation or sequencing reaction itself). For this reason the sample was not brought forward for further analysis. It may be possible to resuscitate the sample by isolating the reads within the desired range however it seems that the ratio of native trinucleosomal fragments to non-trinucleosomal fragments is too low to obtain good quality signal.

6.2.3 Cross-linked versus native data

Here we briefly presented a comparison in CENP-A distribution between native and cross-linked data, using CENP-A data derived from primary and immortalised cell lines. Mononucleosomal CENP-A data from immortalised cells was compared with cross-linked data in primary cells. Here we noticed that CENP-A occupies a slightly different footprint. The average span of CENP-A was reduced by approximately 25% in the native domains compared to the cross-linked domains. Two factors could contribute to this observation; (1) the cell lines are different and (2) the preparation method is different. We anticipated that using a cross-linked preparation method could possibly lead to some false CENP-A signal at the boundary regions. If CENP-A

occupying the boundary region is immunoprecipitated, then due to association with proximal regions through crosslinking it is possible that non-CENP-A associated DNA could be captured. In order to investigate this further we presented analysis on mononucleosomal CENP-A data compared with cross-linked CENP-A data derived from the same cell source i.e. immortalised cells. Here we observed a 15% reduction in CENP-A span in the mononucleosomal data compared with the cross-linked data. Based on our previous observation these data suggest that mononucleosomal preparations represent a more accurate CENP-A distribution. This is in agreement with the 25 kb view of ECAX/EASX when fine peak structure was compared between the two datasets. Here, we observed higher resolution in subpeaks within the domain. These findings could be strengthened by carrying out replicate CENP-A ChIP experiments and by developing an algorithm that estimates centromere boundaries in quantitative manner. The sliding metric developed in Chapter 3 could possibly be used to do this.

6.2.4 Nucleosome calling and analysis

In this chapter, the nucleosome positioning software *nucleR* detected 867, 1087 and 1003 nucleosome positions across each of the centromere alleles analysed, cen9R, cen27L and cenX respectively. Given that each one of the centromere domains analysed had a span between ~80-92 kb, the number of nucleosome positions reported would correspond to ~88 bp between each nucleosome dyad. This is not an accurate representation of real nucleosome spacing, however, it does represent the population distribution of CENP-A positions previously discussed and suggests the presence of fuzzy nucleosome positions. We then presented another method which aimed to calculate the number of CENP-A nucleosomes at a centromere, based on the assumption that the highest *nucleR* scored positions corresponded to the maximally occupied CENP-A sites. We estimated ~100-109 CENP-A nucleosomes to present at any of the domains analysed. These data suggested that if the top nucleosome positions called represented 100% CENP-A occupancy at the 80-90 kb domains, then CENP-A nucleosomes would occupy 20-25% of these positions. These numbers are comparable with recent literature (Bodor et al., 2014) who estimated ~200 CENP-A nucleosomes to present in at human centromeres in interphase which is split into 100 nucleosomes on mitotic chromosomes. Two

aspects of our calculation should be taken into account when comparing nucleosome numbers estimated both in this study and Bodor et al (2014). The figure we reported denotes nucleosome numbers only if the top positions represent 100% occupancy meaning 100% of cells contain CENP-A at these positions. If only 80% or 50% of cells contained CENP-A at these positions then the total number of CENP-A nucleosomes present would be 1.25X or 2X the reported ~100 nucleosomes per domain. The second aspect to take into account is that Bodor et al (2014) estimated the number of CENP-A nucleosomes in human RPE cells and therefore their estimate is based on satellite-containing centromeres. Neocentromeres are known to contain less CENP-A than satellite containing centromeres (Irvine et al., 2005; Marshall et al., 2008b) which was measured to be in the range of ~75% less CENP-A. Together our findings suggest that the numbers of CENP-A nucleosomes calculated for each of the centromeres are in a reasonable range to support kinetochore function.

We also presented an experiment aimed to calculate the distribution of CENP-A occupancy across centromere domains by filtering the top 10%, 25% and 50% of CENP-A nucleosomes based on *nucleR* score. We found that the top 10% of positions contained ~32-42% of total centromeric CENP-A suggesting preferential CENP-A binding in these regions. The top 25% contained ~63-74% and the top 50% contained up to 94% of the entire CENP-A complement. The bottom 50th percentile only seems to contribute a minor fraction of CENP-A. Our findings are very comparable to what was reported in Bodor et al (2014). Using CENP-A ChIP-seq data on the PDNC-4 neocentromere cell line (Amor et al., 2004) they found that the top 10% of potential CENP-A positions contained just over 30% of centromeric CENP-A.

Next we set out to investigate periodicity in nucleosome positions by calculating nucleosome centre-centre distances. Using the top 10%, 25% and 50% of CENP-A nucleosome positions little evidence of periodicity was observed. The centre-centre distances in the top 10% of positions exhibited broad distributions in the histogram plots (100-400 bp) indicating no apparent periodicity. In cen27L and cenX (top 25%) there was slight evidence of multiples of the nucleosome repeat, however this

was not observed in the cen9R centromere domain. The top 50% of CENP-A positions across the three centromere domains showed that the majority of nucleosome centre-centre distances occur between 0-150 bp indicating fuzzy nucleosome positioning of CENP-A at all domains. We then presented an alternative view of the absolute centre-centre distances across each centromere. Here, we postulated if nucleosomes were well positioned or periodically spaced on the chromatin fiber, we would expect to see signal plateaus at the typical nucleosome repeat (~180 bp, ~360 bp, ~540 bp). These results did not show any intervals but instead showed a linear incremental distribution. We concluded from these findings that nucleosome phasing does not occur at long ranges and that fuzzy positions suggest that singe CENP-A nucleosomes occupy numerous DNA configurations over these centromere regions.

6.2.5 Motif analysis

As previously described we used a nucleosome positioning software to call CENP-A nucleosome positions across several centromeres. Using *nucleR* we called the top 10% of CENP-A nucleosome positions across 12 satellite-free centromere domains which corresponded to 1716 sites. Using *MEME*, a motif recognition software (Bailey and Elkan, 1994) we searched for sequence motifs that associated with CENP-A nucleosome positions. *MEME* found five motifs, one of which was a purine rich segment. BLAST analysis of the motif segment showed numerous hits to microsatellite loci along with predicted RNA transcripts specific to the genus *Equus*. Three of the motifs were in very low abundance (~3.8% of total input sequences) and therefore were discounted as the low abundance suggested they weren't major features of centromeric chromatin. The remaining motif was not present at all centromere domains also indicating its unlikeliness to be associated with structural or functional aspects of CENP-A. The findings we report are partly comparable with a previous report on motifs analysis performed on a neocentromere domain (Barry et al., 1999). Barry et al. (1999) reported the presence of the classical microsatellite repeat (Satellite III) on a mardel(10) human neocentromere. They also reported the presence of architectural and regulatory protein HMGI and topoII DNA binding motifs, which have a normal abundance and random distribution genome wide, however, the motif search we performed did not report either of these. This may be

due to the fact that our motif search was specific to the top 10% of CENP-A nucleosomes. Perhaps investigating the location of the motifs with respect to the nucleosome dyad will reveal more about the DNA binding of the CENP-A nucleosome. Current efforts are now focused on characterising the abundance of DNA elements across the domains, which will provide more insight into the DNA configuration across these satellite-free centromere domains.

6.3 CENP-C

In the previous chapter we presented our analysis of CENP-C distribution at satellite-free centromeres in *E.asinus*. We tested a commercially available antibody against CENP-C for suitability in equine chip. The western blot characterisation showed multiple bands across cellular fractions and with one band corresponding to an expected size of donkey CENP-C. This band was present in the chromatin extract at high exposure. We confirmed centromere localisation of the antibody by immunofluorescence by showing clear co-localisation with CENP-A. Then to verify centromere DNA specificity we performed a ChIP-qPCR experiment showed recovery of centromere DNA.

Three CENP-C chip experiments were performed in this study and qPCR analysis showed a varying range of centromere DNA recovery. The first CENP-C chip experiment showed a 1.6% recovery. The low percentage recovery can be explained by the total amount of chromatin used in the experiment which was in the order of 10^8 cell equivalents (C.E), however, this experiment showed a very low background when percentage recovery in the centromere probe was compared to the non-centromeric probes. The qPCR analysis for the second CENP-C chip experiment (Immort1) showed a 6% recovery and this chip assay was performed on ~$20x10^6$ C.E therefore comparable to the recovery in the CENP-C-Primary chip. The CENP-C-Immort2 had a very low recovery of 0.6%, which could be explained by the increased concentration of SDS in the chip buffer.

The quality control analysis we documented in this chapter showed that all CENP-C ChIP-seq datasets had to be filtered in order to obtain maximal read mapping. Part of the problem with bad quality reads that have to be filtered is the consequence of loosing a high percentage of the reads. In our case while that *FastQC* analysis

showed some of the datasets had uneven quality distribution, most of the whisker plots showed that only a minority of the reads were of bad quality and this was clearly represented in the Trimmomatic summary (Table 5.4). The fragment length distributions reported the extent of sonication in all three ChIP-seq samples however a spike anomaly was present in CENP-C-primary and CENP-C-immort1. We presumed this spike was an artefact of the sonication procedure or library preparation. The FRiP scores for two of the datasets (CENP-C-primary, CENP-C-immort1) showed that the immunoprecipitations were successful however the CENP-C-immort2 did not reach the required threshold, which is reflective of the low recovery and higher background observed in the qPCR analysis.

Visualisation of the CENP-C ChIP-seq data revealed remarkably similar peak structures to that of CENP-A, which appeared to localise to the same regions. These data were in agreement with previous ChIP-chip studies in horse (Wade et al., 2009), ChIP-seq studies in *N. Crassa* (Smith et al., 2011), nematodes (Kang et al., 2016; Steiner and Henikoff, 2014) and *S. Pombe* (Thakur et al., 2015).

We then presented a set of boundary comparisons and showed very clearly that the CENP-C and CENP-A boundaries are entirely coincident suggesting the kinetochore forming region is restricted to the genomic footprint occupied by CENP-A, in equids. In order to investigate fine peak structure a correlative approach was performed and the Spearman values indicated a high positive correlation of the two proteins at centromere domains, in strong agreement with our previous findings. Direct inspection of CENP-C distribution with highly scored CENP-A nucleosomes showed that co-distribution of the two protein domains is clearly evident, however there are instances where the two proteins do not show complete correspondence.

We finally reported the relative abundance of CENP-C across the 16 satellite-free centromere domains and showed that like CENP-A, CENP-C levels are strictly maintained. This observation was not entirely consistent in the quantitative analysis of CENP-C-immort1 and CENP-C-immort2, which showed several outliers at some of the centromeres. These inconsistencies could be explained by the nature of the immortalised donkey cells used in the chip experiments as an increased level of aneuploidy has been observed during extended culture. This analysis strongly

indicated that CENP-C and CENP-A covary spatially and quantitatively which is consistent with co-occupancy at centromeric chromatin.

6.4 Concluding Remarks

This thesis has introduced a unique model system in Equids that provides us with a new and powerful tool to investigate centromere structure and regulation. The 16 satellite-free centromere domains in *E. asinus* are stably present and therefore are advantageous with respect to engineered neocentromeres. In this regard, they can be utilised to study the maturation, transmission and regulation of centromeres in a natural environment. The satellite-free centromeres in *E. asinus* have shown us a number of features or properties regarding centromere structure and function. We showed that independent of the genomic footprint occupied by CENP-A, the level of CENP-A protein at centromeres remains remarkably similar, further suggesting the presence of a tight regulatory network in place to retain CENP-A at the locus. This finding is in strong agreement with CENP-C levels at the centromere (also shown in this work), which are remarkably constant across the cohort of satellite-free centromeres in the donkey. These findings tell us that CENP-A maintenance and stability at satellite-free centromeres may occur in a CENP-B independent manner. We showed centromere positional variation as a property of centromeres at some stage during transmission from parents to offspring, therefore the donkey system also provides us with a way of investigating centromere inheritance at the DNA sequence level. The donkey system also allows us to investigate directly the DNA sequences surrounding CENP-A nucleosomes thereby possibly providing more insight into DNA selection by centromere components.

The non-repetitive nature of these centromeres provides a means to further examine centromere structure and nucleosome organisation, at a molecular level and this remains even more significant as single-cell technologies mature. Technologies such as HiC could also be introduced to examine the centromeric chromatin architecture providing more insight into how centromere structure defines function. The quantitative profile presented in this work can be used as a means to report centromere behaviour after introducing methods like siRNA knockdowns or introducing mutations with genome editing technologies like

CRISPR, to target regulators of CENP-A. Assembly pathway components are one set of targets, however, any aspect of chromosome regulation such as transcription, DNA methylation or CTCF binding could be examined.

Chapter 7 References

Allshire, R.C., and Karpen, G.H. (2008). Epigenetic regulation of centromeric chromatin: old dogs, new tricks? Nature Reviews. Genetics *9*, 923–937.

Alonso, A., Mahmood, R., Li, S., Cheung, F., Yoda, K., and Warburton, P.E. (2003). Genomic microarray analysis reveals distinct locations for the CENP-A binding domains in three human chromosome 13q32 neocentromeres. Human Molecular Genetics *12*, 2711–2721.

Alonso, A., Fritz, B., Hasson, D., Abrusan, G., Cheung, F., Yoda, K., Radlwimmer, B., Ladurner, A.G., and Warburton, P.E. (2007). Co-localization of CENP-C and CENP-H to discontinuous domains of CENP-A chromatin at human neocentromeres. Genome Biology *8*, R148.

Amano, M., Suzuki, A., Hori, T., Backer, C., Okawa, K., Cheeseman, I.M., and Fukagawa, T. (2009). The CENP-S complex is essential for the stable assembly of outer kinetochore structure. Journal of Cell Biology *186*, 173–182.

Amaro, A.C., Samora, C.P., Holtackers, R., Wang, E., Kingston, I.J., Alonso, M., Lampson, M., McAinsh, A.D., and Meraldi, P. (2010). Molecular control of kinetochore-microtubule dynamics and chromosome oscillations. Nature Cell Biology *12*, 319–329.

Amor, D.J., Bentley, K., Ryan, J., Perry, J., Wong, L., Slater, H., and Choo, K.H.A. (2004). Human centromere repositioning “in progress”. Proceedings of the National Academy of Sciences of the United States of America *101*, 6542–6547.

Ando, S., Yang, H., Nozaki, N., Okazaki, T., and Yoda, K. (2002). CENP-A, -B, and -C chromatin complex that contains the I-type alpha-satellite array constitutes the prekinetochore in HeLa cells. Molecular And Cellular Biology *22*, 2229–2241.

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data.

Bailey, T.L., and Elkan, C. (1994). Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Bipolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology 28–36.

Barry, A.E., Howman, E. V., Cancilla, M.R., Saffery, R., and Choo, K.H.A. (1999). Sequence analysis of an 80 kb human neocentromere. Human Molecular Genetics *8*, 217–227.

Basilico, F., Maffini, S., Weir, J.R., Prumbaum, D., Rojas, A.M., Zimniak, T., De Antoni, A., Jeganathan, S., Voss, B., Van Gerwen, S., et al. (2014). The pseudo GTPase CENP-M drives human kinetochore assembly. eLife *2014*.

Bassett, E.A., DeNizio, J., Barnhart-Dailey, M.C., Panchenko, T., Sekulic, N., Rogers, D.J., Foltz, D.R., and Black, B.E. (2012). HJURP Uses Distinct CENP-A Surfaces to Recognize and to Stabilize CENP-A/Histone H4 for Centromere Assembly. Developmental Cell *22*, 749–762.

Bergmann, J.H., guez, M.G. oacute mez R. iacute, Martins, N.M.C., Kimura, H., Kelly, D.A., Masumoto, H., Larionov, V., Jansen, L.E.T., and Earnshaw, W.C. (2011). Epigenetic engineering shows H3K4me2 is required for HJURP targeting and CENP-A assembly on a synthetic human kinetochore. The EMBO Journal *30*, 328–

340.

Black, B.E., Foltz, D.R., Chakravarthy, S., Luger, K., Woods, V.L., and Cleveland, D.W. (2004). Structural determinants for generating centromeric chromatin. Nature *430*, 578–582.

Black, B.E., Brock, M.A., Bédard, S., Woods, V.L., and Cleveland, D.W. (2007a). An epigenetic mark generated by the incorporation of CENP-A into centromeric nucleosomes. Proceedings of the National Academy of Sciences of the United States of America *104*, 5008–5013.

Black, B.E., Jansen, L.E.T., Maddox, P.S., Foltz, D.R., Desai, A.B., Shah, J. V., and Cleveland, D.W. (2007b). Centromere Identity Maintained by Nucleosomes Assembled with Histone H3 Containing the CENP-A Targeting Domain. Molecular Cell *25*, 309–322.

Blower, M.D., Sullivan, B.A., and Karpen, G.H. (2002). Conserved organization of centromeric chromatin in flies and humans. Developmental Cell *2*, 319–330.

Bodor, D.L., Valente, L.P., Mata, J.F., Black, B.E., and Jansen, L.E.T. (2013). Assembly in G1 phase and long-term stability are unique intrinsic features of CENP-A nucleosomes. Molecular Biology of the Cell *24*, 923–932.

Bodor, D.L., Mata, J.F.J.F., Sergeev, M., David, A.F., Salimian, K.J., Panchenko, T., Cleveland, D.W., Black, B.E., Shah, J. V., and Jansen, L.E.T. (2014). The quantitative architecture of centromeric chromatin. eLife *2014*, 1–26.

Capozzi, O., Purgato, S., Verdun di Cantogno, L., Grosso, E., Ciccone, R., Zuffardi, O., Della Valle, G., and 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., and Rocchi, M. (2009). Evolutionary descent of a human chromosome 6 neocentromere: A jump back to 17 million years ago. Genome Research *19*, 778–784.

Carbone, L., Nergadze, S.G., 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, M.F., Alonso, A., Pazienza, M., Ventura, M., Montemurro, G., Carbone, L., de Jong, P.J., 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 Biology *7*, R91.

Carroll, C.W., Silva, M.C.C., Godek, K.M., Jansen, L.E.T., and Straight, A.F. (2009). Centromere assembly requires the direct recognition of CENP-A nucleosomes by CENP-N. Nature Cell Biology *11*, 896–902.

Carroll, C.W., Milks, K.J., and Straight, A.F. (2010). Dual recognition of CENP-A nucleosomes is required for centromere assembly. Journal of Cell Biology *189*, 1143–1155.

Castillo, A.G., Mellonea, B.G., Partridgeb, J.F., Richardsonc, W., Hamilton, G.L., Allshire, R.C., and Pidoux, A.L. (2007). Plasticity of fission yeast CENP-A chromatin driven by relative levels of histone H3 and H4. PLoS Genetics *3*, 1264–1274.

Cheeseman, I.M., and Desai, A. (2008). Molecular architecture of the kinetochore-

microtubule interface. Nat. Rev Mol. Cell Biol. *9*, 33–46.

Chen, X., Hoffman, M.M., Bilmes, J.A., Hesselberth, J.R., and Noble, W.S. (2010). A dynamic Bayesian network for identifying protein-binding footprints from single molecule-based sequencing data. Bioinformatics *26*.

Chueh, A.C., Wong, L.H., Wong, N., and Choo, K.H.A. (2005). Variable and hierarchical size distribution of L1-retroelement-enriched CENP-A clusters within a functional human neocentromere. Human Molecular Genetics *14*, 85–93.

Clarke, L. (1998). Centromeres: Proteins, protein complexes, and repeated domains at centromeres of simple eukaryotes. Current Opinion in Genetics and Development *8*, 212–218.

Clemente, I.C., Ponsà, M., García, M., and Egozcue, J. (1990). Evolution of the Simiiformes and the phylogeny of human chromosomes. Human Genetics *84*, 493–506.

Cleveland, D., Mao, Y., and Sullivan, K. (2003). Centromeres and Kinetochores-From Epigenetics to Mitotic Checkpoint Signaling. Cell *112*, 407–422.

Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L., and Rice, P.M. (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research *38*, 1767–1771.

Coghlan, A., Eichler, E.E., Oliver, S.G., Paterson, A.H., and Stein, L. (2005). Chromosome evolution in eukaryotes: A multi-kingdom perspective. Trends in Genetics *21*, 673–682.

Dambacher, S., Deng, W., Hahn, M., Sadic, D., Fröhlich, J., Nuber, A., Hoischen, C., Diekmann, S., Leonhardt, H., and Schotta, G. (2012). CENP-C facilitates the recruitment of M18BP1 to centromeric chromatin. Nucleus (Austin, Tex.) *3*, 101–110.

Dornblut, C., Quinn, N., Monajambashi, S., Prendergast, L., van Vuuren, C., Münch, S., Deng, W., Leonhardt, H., Cardoso, M.C., Hoischen, C., et al. (2014). A CENP-S/X complex assembles at the centromere in S and G2 phases of the human cell cycle. Open Biology *4*, 130229.

Dunleavy, E.M., Roche, D., Tagami, H., Lacoste, N., Ray-Gallet, D., Nakamura, Y., Daigo, Y., Nakatani, Y., and Almouzni-Pettinotti, G. (2009). HJURP Is a Cell-Cycle-Dependent Maintenance and Deposition Factor of CENP-A at Centromeres. Cell *137*, 485–497.

Dunleavy, E.M., Almouzni, G., and Karpen, G.H. (2011). H3.3 is deposited at centromeres in S phase as a placeholder for newly assembled CENP-A in G_1 phase. Nucleus (Austin, Tex.) *2*, 146–157.

Dutrillaux, B. (1979). Chromosomal evolution in Primates: Tentative phylogeny from Microcebus murinus (Prosimian) to man. Human Genetics *48*, 251–314.

Earnshaw, W.C., and Migeon, B.R. (1985). Three related centromere proteins are absent from the inactive centromere of a stable isodicentric chromosome. Chromosoma *92*, 290–296.

Earnshaw, W.C., and Rothfield, N. (1985). Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma. Chromosoma *91*, 313–321.

Earnshaw, W.C., Sullivan, K.F., Machlin, P.S., Cooke, C.A., Kaiser, D.A., Pollard, T.D., Rothfield, N.F., and Cleveland, D.W. (1987). Molecular cloning of cDNA for CENP-B, the major human centromere autoantigen. Journal of Cell Biology *104*, 817–829.

Eskat, A., Deng, W., Hofmeister, A., Rudolphi, S., Emmerth, S., Hellwig, D., Ulbricht, T., Döring, V., Bancroft, J.M., McAinsh, A.D., et al. (2012). Step-Wise Assembly, Maturation and Dynamic Behavior of the Human CENP-P/O/R/Q/U Kinetochore Sub-Complex. PLoS ONE *7*.

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research *8*, 186–194.

Ewing, B., Hillier, L., Wendl, M., and Green, P. (1998). Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment. Genome Research 175–185.

Fachinetti, D., Folco, H.D., Nechemia-Arbely, Y., Valente, L.P., Nguyen, K., Wong, A.J., Zhu, Q., Holland, A.J., Desai, A., Jansen, L.E.T., et al. (2013). A two-step mechanism for epigenetic specification of centromere identity and function. Nature Cell Biology *15*, 1056–1066.

Fachinetti, D., Han, J.S., McMahon, M.A., Ly, P., Abdullah, A., Wong, A.J., and Cleveland, D.W. (2015). DNA Sequence-Specific Binding of CENP-B Enhances the Fidelity of Human Centromere Function. Developmental Cell *33*, 314–327.

Falk, S.J., Guo, L.Y., Sekulic, N., Smoak, E.M., Mani, T., Logsdon, G.A., Gupta, K., Jansen, L.E.T., Van Duyne, G.D., Vinogradov, S.A., et al. (2015). CENP-C reshapes and stabilizes CENP-A nucleosomes at the centromere. Science (New York, NY) *348*, 699–703.

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

Flores, O., and Orozco, M. (2011). nucleR: A package for non-parametric nucleosome positioning. Bioinformatics *27*, 2149–2150.

Foltz, D.R., Jansen, L.E.T., Black, B.E., Bailey, A.O., Yates, J.R., and Cleveland, D.W. (2006). The human CENP-A centromeric nucleosome-associated complex. Nature Cell Biology *8*, 458–469.

Foltz, D.R., Jansen, L.E.T., Bailey, A.O., Yates, J.R., Bassett, E.A., Wood, S., Black, B.E., and Cleveland, D.W. (2009). Centromere-Specific Assembly of CENP-A Nucleosomes Is Mediated by HJURP. Cell *137*, 472–484.

Forus, A., Bjerkehagen, B., Sirvent, N., Meza-Zepeda, L.A., Coindre, J.M., Berner, J.M., Myklebost, O., and Pedeutour, F. (2001). A well-differentiated liposarcoma with a new type of chromosome 12-derived markers. Cancer Genetics and Cytogenetics *131*, 13–18.

Fujita, R., Otake, K., Arimura, Y., Horikoshi, N., Miya, Y., Shiga, T., Osakabe, A., Tachiwana, H., Ohzeki, J.I., Larionov, V., et al. (2015). Stable complex formation of CENP-B with the CENP-A nucleosome. Nucleic Acids Research *43*, 4909–4922.

Fujita, Y., Hayashi, T., Kiyomitsu, T., Toyoda, Y., Kokubu, A., Obuse, C., and Yanagida, M. (2007). Priming of Centromere for CENP-A Recruitment by Human hMis18α, hMis18β, and M18BP1. Developmental Cell *12*, 17–30.

Fukagawa, T., and Brown, W.R. (1997). Efficient conditional mutation of the vertebrate CENP-C gene. Human Molecular Genetics *6*, 2301–2308.

Fukagawa, T., and Earnshaw, W.C. (2014). The centromere: Chromatin foundation for the kinetochore machinery. Developmental Cell *30*, 496–508.

Fukagawa, T., Pendon, C., Morris, J., and Brown, W. (1999). CENP-C is necessary but not sufficient to induce formation of a functional centromere. EMBO Journal *18*, 4196–4209.

Furuyama, S., and Biggins, S. (2007). Centromere identity is specified by a single centromeric nucleosome in budding yeast. Proceedings of the National Academy of Sciences of the United States of America *104*, 14706–14711.

Gaffney, D.J., McVicker, G., Pai, A.A., Fondufe-Mittendorf, Y.N., Lewellen, N., Michelini, K., Widom, J., Gilad, Y., and Pritchard, J.K. (2012). Controls of Nucleosome Positioning in the Human Genome. PLoS Genetics *8*.

Gascoigne, K.E., Takeuchi, K., Suzuki, A., Hori, T., Fukagawa, T., and Cheeseman, I.M. (2011). Induced ectopic kinetochore assembly bypasses the requirement for CENP-A nucleosomes. Cell *145*, 410–422.

Gassmann, R., Rechtsteiner, A., Yuen, K.W., Muroyama, A., Egelhofer, T., Gaydos, L., Barron, F., Maddox, P., Essex, A., Monen, J., et al. (2012). An inverse relationship to germline transcription defines centromeric chromatin in C. elegans. Nature *484*, 534–537.

Di Gesù, V., Lo Bosco, G., Pinello, L., Yuan, G.C., and Corona, D.F. V (2009). A multi-layer method to study genome-scale positions of nucleosomes. Genomics *93*, 140–145.

Goshima, G., Kiyomitsu, T., Yoda, K., and Yanagida, M. (2003). Human centromere chromatin protein hMis12, essential for equal segregation, is independent of CENP-A loading pathway. Journal of Cell Biology *160*, 25–39.

Greil, F., Van Der Kraan, I., Delrow, J., Smothers, J.F., De Wit, E., Bussemaker, H.J., Van Driel, R., Henikoff, S., and Van Steensel, B. (2003). Distinct HP1 and Su(var)3-9 complexes bind to sets of developmentally coexpressed genes depending on chromosomal location. Genes and Development *17*, 2825–2838.

Guse, A., Carroll, C.W., Moree, B., Fuller, C.J., and Straight, A.F. (2011). In vitro centromere and kinetochore assembly on defined chromatin templates. Nature *477*, 354–358.

Hasson, D., Panchenko, T., Salimian, K.J., Salman, M.U., Sekulic, N., Alonso, A., Warburton, P.E., and Black, B.E. (2013). The octamer is the major form of CENP-A nucleosomes at human centromeres. Nature Structural & Molecular Biology *20*, 687–695.

Hayashi, T., Fujita, Y., Iwasaki, O., Adachi, Y., Takahashi, K., and Yanagida, M. (2004). Mis16 and Mis18 are required for CENP-A loading and histone deacetylation at centromeres. Cell *118*, 715–729.

Henikoff, S., Ahmad, K., and Malik, H.S. (2001). The centromere paradox: stable inheritance with rapidly evolving DNA. Science (New York, N.Y.) *293*, 1098–1102.

Heun, P., Erhardt, S., Blower, M.D., Weiss, S., Skora, A.D., and Karpen, G.H. (2006). Mislocalization of the drosophila centromere-specific histone CID promotes formation of functional ectopic kinetochores. Developmental Cell *10*, 303–315.

Hooser, A.A. Van, Ouspenski, I.I., Gregson, H.C., Starr, D.A., Yen, T.J., Goldberg, M.L.,

Yokomori, K., Earnshaw, W.C., Sullivan, K.F., and Brinkley, B.R. (2001). Specification of kinetochore-forming chromatin by the histone H3 variant CENP-A. Journal of Cell Science *114*, 3529–3542.

Hori, T., Okada, M., Maenaka, K., and Fukagawa, T. (2008a). CENP-O class proteins form a stable complex and are required for proper kinetochore function. Molecular Biology of the Cell *19*, 843–854.

Hori, T., Amano, M., Suzuki, A., Backer, C.B., Welburn, J.P., Dong, Y., McEwen, B.F., Shang, W.-H.H., Suzuki, E., Okawa, K., et al. (2008b). CCAN Makes Multiple Contacts with Centromeric DNA to Provide Distinct Pathways to the Outer Kinetochore. Cell *135*, 1039–1052.

Hu, H., Liu, Y., Wang, M., Fang, J., Huang, H., Yang, N., Li, Y., Wang, J., Yao, X., Shi, Y., et al. (2011). Structure of a CENP-A-histone H4 heterodimer in complex with chaperone HJURP. Genes and Development *25*, 901–906.

Hudson, D.F., Fowler, K.J., Earle, E., Saffery, R., Kalitsis, P., Trowell, H., Hill, J., Wreford, N.G., De Kretser, D.M., Cancilla, M.R., et al. (1998). Centromere protein B null mice are mitotically and meiotically normal but have lower body and testis weights. Journal of Cell Biology *141*, 309–319.

Hughes-Schrader, S., and Ris, H. (1941). The diffuse spindle attachment of coccids, verified by the mitotic behavior of induced chromosome fragments. Journal of Experimental Zoology *87*, 429–456.

Irvine, D. V., Amor, D.J., Perry, J., Sirvent, N., Pedeutour, F., Choo, K.H.A., and Saffery, R. (2005). Chromosome size and origin as determinants of the level of CENP-A incorporation into human centromeres. Chromosome Research *12*, 805–815.

Ishii, K., Ogiyama, Y., Chikashige, Y., Soejima, S., Masuda, F., Kakuma, T., Hiraoka, Y., and Takahashi, K. (2008). Heterochromatin integrity affects chromosome reorganization after centromere dysfunction. Science (New York, N.Y.) *321*, 1088–1091.

Izuta, H., Ikeno, M., Suzuki, N., Tomonaga, T., Nozaki, N., Obuse, C., Kisu, Y., Goshima, N., Nomura, F., Nomura, N., et al. (2006). Comprehensive analysis of the ICEN (Interphase Centromere Complex) components enriched in the CENP-A chromatin of human cells. Genes to Cells *11*, 673–684.

Jansen, L.E.T., Black, B.E., Foltz, D.R., and Cleveland, D.W. (2007). Propagation of centromeric chromatin requires exit from mitosis. Journal of Cell Biology *176*, 795–805.

Jiang, C., and Pugh, B.F. (2009). Nucleosome positioning and gene regulation: advances through genomics. Nature Reviews. Genetics *10*, 161–172.

Kalitsis, P., and Choo, K.H.A. (2012). The evolutionary life cycle of the resilient centromere. Chromosoma *121*, 327–340.

Kang, Y., Wang, J., Neff, A., Kratzer, S., Kimura, H., and Davis, R.E. (2016). Differential Chromosomal Localization of Centromeric Histone CENP-A Contributes to Nematode Programmed DNA Elimination.

Kapoor, M., Montes De Oca Luna, R., Liu, G., Lozano, G., Cummings, C., Mancini, M., Ouspenski, I., Brinkley, B.R., and May, G.S. (1998). The cenpB gene is not essential in mice. Chromosoma *107*, 570–576.

Kato, H., Jiang, J., Zhou, B.-R.R., Rozendaal, M., Feng, H., Ghirlando, R., Xiao, T.S., Straight, A.F., and Bai, Y. (2013). A conserved mechanism for centromeric nucleosome recognition by centromere protein CENP-C. Science (New York, N.Y.) *340*, 1110–1113.

Klare, K., Weir, J.R., Basilico, F., Zimniak, T., Massimiliano, L., Ludwigs, N., Herzog, F., and Musacchio, A. (2015). CENP-C is a blueprint for constitutive centromere-associated network assembly within human kinetochores. The Journal of Cell Biology *210*, 11–22.

Knehr, M., Poppe, M., Schroeter, D., Eickelbaum, W., Finze, E.M., Kiesewetter, U.L., Enulescu, M., Arand, M., and Paweletz, N. (1996). 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 *93*, 10234–9.

Lagana, A., Dorn, J.F., De Rop, V., Ladouceur, A.-M., Maddox, A.S., and Maddox, P.S. (2010). A small GTPase molecular switch regulates epigenetic centromere maintenance by stabilizing newly incorporated CENP-A. Nature Cell Biology *12*, 1186–1193.

Lander, E.S., and Waterman, M.S. (1988). Genomic mapping by fingerprinting random clones: A mathematical analysis. Genomics *2*, 231–239.

Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research *22*, 1813–1831.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357–359.

Lee, W., Tillo, D., Bray, N., Morse, R.H., Davis, R.W., Hughes, T.R., and Nislow, C. (2007). A high-resolution atlas of nucleosome occupancy in yeast. Nature Genetics *39*, 1235–1244.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Liu, S.-T., Hittle, J.C., Jablonski, S.A., Campbell, M.S., Yoda, K., and Yen, T.J. (2003). Human CENP-I specifies localization of CENP-F, MAD1 and MAD2 to kinetochores and is essential for mitosis. Nature Cell Biology *5*, 341–345.

Lo, A.W.I., Magliano, D.J., Sibson, M.C., Kalitsis, P., Craig, J.M., and Choo, K.H.A. (2001a). A novel chromatin immunoprecipitation and array (CIA) analysis identifies a 460-kb CENP-A-binding neocentromere DNA. Genome Research *11*, 448–457.

Lo, A.W.I., Craig, J.M., Saffery, R., Kalitsis, P., Irvine, D. V., Earle, E., Magliano, D.J., and Choo, K.H.A. (2001b). A 330 kb CENP-A binding domain and altered replication timing at a human neocentromere. EMBO Journal *20*, 2087–2096.

Locke, D.P., Hillier, L.W., Warren, W.C., Worley, K.C., Nazareth, L. V, Muzny, D.M., Yang, S.-P., Wang, Z., Chinwalla, A.T., Minx, P., et al. (2011). Comparative and demographic analysis of orang-utan genomes. Nature *469*, 529–533.

Logsdon, G.A., Barrey, E.J., Bassett, E.A., DeNizio, J.E., Guo, L.Y., Panchenko, T.,

Dawicki-McKenna, J.M., Heun, P., and Black, B.E. (2015). Both tails and the centromere targeting domain of CENP-A are required for centromere establishment. Journal of Cell Biology *208*, 521–531.

Luger, K., Mäder, a W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature *389*, 251–260.

Malik, H.S., and Henikoff, S. (2003). Phylogenomics of the nucleosome. Nature Structural Biology *10*, 882–891.

Maloney, K. a., Sullivan, L.L., Matheny, J.E., Strome, E.D., Merrett, S.L., Ferris, A., and Sullivan, B. a. (2012). Functional epialleles at an endogenous human centromere. Proceedings of the National Academy of Sciences *109*, 13704–13709.

Marshall, O.J., Chueh, A.C., Wong, L.H., and Choo, K.H.A. (2008a). Neocentromeres: New Insights into Centromere Structure, Disease Development, and Karyotype Evolution. American Journal of Human Genetics *82*, 261–282.

Marshall, O.J., Marshall, A.T., and Choo, K.H.A. (2008b). Three-dimensional localization of CENP-A suggests a complex higher order structure of centromeric chromatin. Journal of Cell Biology *183*, 1193–1202.

Masumoto, H., Nakano, M., and Ohzeki, J.I. (2004). The role of CENP-B and ??-satellite DNA: De novo assembly and epigenetic maintenance of human centromeres. Chromosome Research *12*, 543–556.

McKinley, K.L., and Cheeseman, I.M. (2014). Polo-like kinase 1 licenses CENP-a deposition at centromeres. Cell *158*, 397–411.

McKinley, K.L., and Cheeseman, I.M. (2016). The molecular basis for centromere identity and function. Nat Rev Mol Cell Biol *17*, 16–29.

Mello, J. a, and Almouzni, G. (2001). The ins and outs of nucleosome assembly. Current Opinion in Genetics & Development *11*, 136–141.

Meluh, P.B., Yang, P., Glowczewski, L., Koshland, D., and Smith, M.M. (1998). Cse4p is a component of the core centromere of Saccharomyces cerevisiae. Cell *94*, 607–613.

Di Meo, G.P., Perucatti, A., Peretti, V., Incarnato, D., Ciotola, F., Liotta, L., Raudsepp, T., Di Berardino, D., Chowdhary, B., and Iannuzzi, L. (2009). The 450-band resolution G- and R-banded standard karyotype of the donkey (Equus asinus, 2n = 62). Cytogenetic and Genome Research *125*, 266–271.

Montefalcone, G., Tempesta, S., Rocchi, M., and Archidiacono, N. (1999). Centromere repositioning. Genome Research *9*, 1184–1188.

Moree, B., Meyer, C.B., Fuller, C.J., and Straight, A.F. (2011). CENP-C recruits M18BP1 to centromeres to promote CENP-A chromatin assembly. Journal of Cell Biology *194*, 855–871.

Moroi, Y., Peebles, C., Fritzler, M.J., Steigerwald, J., and Tan, E.M. (1980). Autoantibody to centromere (kinetochore) in scleroderma sera. Proceedings of the National Academy of Sciences of the United States of America *77*, 1627–1631.

Muro, Y., Masumoto, H., Yoda, K., Nozaki, N., Ohashi, M., and Okazaki, T. (1992). Centromere protein B assembles human centromeric ??-satellite DNA at the 17-bp sequence, CENP-B box. Journal of Cell Biology *116*, 585–596.

Murphy, T.D., and Karpen, G.H. (1995). Localization of centromere function in a drosophila minichromosome. Cell *82*, 599–609.

Nagaki, K., Cheng, Z., Ouyang, S., Talbert, P.B., Kim, M., Jones, K.M., Henikoff, S., Buell, C.R., and Jiang, J. (2004). Sequencing of a rice centromere uncovers active genes. Nature Genetics *36*, 138–145.

Ngan, V.K., and Clarke, L. (1997). The centromere enhancer mediates centromere activation in Schizosaccharomyces pombe. Molecular and Cellular Biology *17*, 3305–3314.

Nishihashi, A., Haraguchi, T., Hiraoka, Y., Ikemura, T., Regnier, V., Dodson, H., Earnshaw, W.C., and Fukagawa, T. (2002). CENP-I is essential for centromere function in vertebrate cells. Developmental Cell *2*, 463–476.

Nishino, T., Takeuchi, K., Gascoigne, K.E., Suzuki, A., Hori, T., Oyama, T., Morikawa, K., Cheeseman, I.M., and Fukagawa, T. (2012). CENP-T-W-S-X forms a unique centromeric chromatin structure with a histone-like fold. Cell *148*, 487–501.

Nishino, T., Rago, F., Hori, T., Tomii, K., Cheeseman, I.M., and Fukagawa, T. (2013). CENP-T provides a structural platform for outer kinetochore assembly. The EMBO Journal *32*, 424–436.

Obuse, C., Yang, H., Nozaki, N., Goto, S., Okazaki, T., and Yoda, K. (2004). Proteomics analysis of the centromere complex from HeLa interphase cells: UV-damaged DNA binding protein 1 (DDB-1) is a component of the CEN-complex, while BMI-1 is transiently co-localized with the centromeric region in interphase. Genes to Cells *9*, 105–120.

Ohzeki, J. ichirou, Nakano, M., Okada, T., and Masumoto, H. (2002). CENP-B box is required for de novo centromere chromatin assembly on human alphoid DNA. Journal of Cell Biology *159*, 765–775.

Okada, M., Cheeseman, I.M., Hori, T., Okawa, K., McLeod, I.X., Yates, J.R., Desai, A., and Fukagawa, T. (2006). The CENP-H-I complex is required for the efficient incorporation of newly synthesized CENP-A into centromeres. Nature Cell Biology *8*, 446–457.

Okada, T., Ohzeki, J. ichirou, Nakano, M., Yoda, K., Brinkley, W.R., Larionov, V., and Masumoto, H. (2007). CENP-B Controls Centromere Formation Depending on the Chromatin Context. Cell *131*, 1287–1300.

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, D.K., and Margolis, R.L. (1985). Kinetochore components recognized by human autoantibodies are present on mononucleosomes. Molecular and Cellular Biology *5*, 173–186.

Palmer, D.K., O'Day, K., Wener, M.H., Andrews, B.S., and Margolis, R.L. (1987). A 17-kD centromere protein (CENP-A) copurifies with nucleosome core particles and with histones. Journal of Cell Biology *104*, 805–815.

Partridge, J.F., Borgstrøm, B., and Allshire, R.C. (2000). Distinct protein interaction domains and protein spreading in a complex centromere. Genes and Development

14, 783–791.

Pedeutour, F., Forus, A., Coindre, J.M., Berner, J.M., Nicolo, G., Michiels, J.F., Terrier, P., Ranchere-Vince, D., Collin, F., Myklebost, O., et al. (1999). Structure of the supernumerary ring and giant rod chromosomes in adipose tissue tumors. Genes Chromosomes and Cancer *24*, 30–41.

Perez-Castro, a V, Shamanski, F.L., Meneses, J.J., Lovato, T.L., Vogel, K.G., Moyzis, R.K., and Pedersen, R. (1998). Centromeric protein B null mice are viable with no apparent abnormalities. Developmental Biology *201*, 135–143.

Perpelescu, M., Nozaki, N., Obuse, C., Yang, H., and Yoda, K. (2009). Active establishment of centromeric cenp-a chromatin by rsf complex. Journal of Cell Biology *185*, 397–407.

Piras, F.M., Nergadze, S.G., Magnani, E., Bertoni, L., Attolini, C., Khoriauli, L., Raimondi, E., and Giulotto, E. (2010). Uncoupling of satellite DNA and centromeric function in the genus Equus. PLoS Genetics *6*.

Prendergast, L., van Vuuren, C., Kaczmarczyk, A., Doering, V., Hellwig, D., Quinn, N., Hoischen, C., Diekmann, S., and Sullivan, K.F. (2011). Premitotic assembly of human CENPs -T and -W switches centromeric chromatin to a mitotic state. PLoS Biology *9*, e1001082.

Prendergast, L., Müller, S., Liu, Y., Huang, H., Dingli, F., Loew, D., Vassias, I., Patel, D.J., Sullivan, K.F., and Almouzni, G. (2016). The CENP-T/-W complex is a binding partner of the histone chaperone FACT. Genes & Development *30*, 1313–1326.

Przewloka, M.R., Venkei, Z., Bolanos-Garcia, V.M., Debski, J., Dadlez, M., and Glover, D.M. (2011). CENP-C is a structural platform for kinetochore assembly. Current Biology *21*, 399–405.

Purgato, S., Belloni, E., Piras, F.M., Zoli, M., Badiale, C., Cerutti, F., Mazzagatti, A., Perini, G., Della Valle, G., Nergadze, S.G., et al. (2015). Centromere sliding on a mammalian chromosome. Chromosoma *124*, 277–287.

R Development Core Team, R. (2008). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing *1*, 2673.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). DeepTools: A flexible platform for exploring deep-sequencing data. Nucleic Acids Research *42*.

Ribeiro, S.A., Vagnarelli, P., Dong, Y., Hori, T., McEwen, B.F., Fukagawa, T., Flors, C., and Earnshaw, W.C. (2010). A super-resolution map of the vertebrate kinetochore. Proceedings of the National Academy of Sciences of the United States of America *107*, 10484–10489.

Rocchi, M., Archidiacono, N., Schempp, W., Capozzi, O., and Stanyon, R. (2012). Centromere repositioning in mammals. Heredity *108*, 59–67.

Ross, J.E., Woodlief, K.S., and Sullivan, B.A. (2016). Inheritance of the CENP-A chromatin domain is spatially and temporally constrained at human centromeres. Epigenetics & Chromatin *9*, 20.

Routh, A., Sandin, S., and Rhodes, D. (2008). Nucleosome repeat length and linker histone stoichiometry determine chromatin fiber structure. Proceedings of the National Academy of Sciences of the United States of America *105*, 8872–8877.

Ryder, O.A., Epel, N.C., and Benirschke, K. (1978). Chromosome banding studies of

the equidae. Cytogenetic and Genome Research *20*, 323–350.

Saffery, R., Irvine, D. V, Griffiths, B., Kalitsis, P., Wordeman, L., and Choo, K.H. (2000). Human centromeres and neocentromeres show identical distribution patterns of >20 functionally important kinetochore-associated proteins. Human Molecular Genetics *9*, 175–185.

Saitoh, H., Tomkiel, J., Cooke, C.A., Ratrie, H., Maurer, M., Rothfield, N.F., and Earnshaw, W.C. (1992). CENP-C, an autoantigen in scleroderma, is a component of the human inner kinetochore plate. Cell *70*, 115–125.

Schleiffer, A., Maier, M., Litos, G., Lampert, F., Hornung, P., Mechtler, K., and Westermann, S. (2012). CENP-T proteins are conserved centromere receptors of the Ndc80 complex. Nature Cell Biology *14*, 1–12.

Shang, W.H., Hori, T., Toyoda, A., Kato, J., Popendorf, K., Sakakibara, Y., Fujiyama, A., and Fukagawa, T. (2010). Chickens possess centromeres with both extended tandem repeats and short non-tandem-repetitive sequences. Genome Research *20*, 1219–1228.

Shang, W.H., Hori, T., Martins, N.M.C., Toyoda, A., Misu, S., Monma, N., Hiratani, I., Maeshima, K., Ikeo, K., Fujiyama, A., et al. (2013). Chromosome Engineering Allows the Efficient Isolation of Vertebrate Neocentromeres. Developmental Cell *24*, 1–14.

Shelby, R.D., Vafa, O., and Sullivan, K.F. (1997). Assembly of CENP-A into centromeric chromatin requires a cooperative array of nucleosomal DNA contact sites. Journal of Cell Biology *136*, 501–513.

Shelby, R.D., Monier, K., and Sullivan, K.F. (2000). Chromatin assembly at kinetochores is uncoupled from DNA replication. Journal of Cell Biology *151*, 1113–1118.

Shuaib, M., Ouararhni, K., Dimitrov, S., and Hamiche, A. (2010). HJURP binds CENP-A via a highly conserved N-terminal domain and mediates its deposition at centromeres. Proceedings of the National Academy of Sciences of the United States of America *107*, 1349–1354.

Silva, M.C.C., Bodor, D.L., Stellfox, M.E., Martins, N.M.C., Hochegger, H., Foltz, D.R., and Jansen, L.E.T. (2012). Cdk Activity Couples Epigenetic Centromere Inheritance to Cell Cycle Progression. Developmental Cell *22*, 52–63.

Sirvent, N., Forus, A., Lescaut, W., Burel, F., Benzaken, S., Chazal, M., Bourgeon, A., Vermeesch, J.R., Myklebost, O., Turc-Carel, C., et al. (2000). Characterization of centromere alterations in liposarcomas. Genes Chromosomes and Cancer *29*, 117–129.

Smith, K.M., Phatale, P. a, Sullivan, C.M., Pomraning, K.R., and Freitag, M. (2011). Heterochromatin is required for normal distribution of Neurospora crassa CenH3. Molecular and Cellular Biology *31*, 2528–2542.

Song, J.S., Liu, X., Liu, X.S., and He, X. (2008). A high-resolution map of nucleosome positioning on a fission yeast centromere. Genome Research *18*, 1064–1072.

Steiner, F.A., and Henikoff, S. (2014). Holocentromeres are dispersed point centromeres localized at transcription factor hotspots. eLife *2014*.

Stellfox, M.E., Bailey, A.O., and Foltz, D.R. (2013). Putting CENP-A in its place. Cellular and Molecular Life Sciences *70*, 387–406.

Sugata, N., Munekata, E., and Todokoro, K. (1999). Characterization of a novel kinetochore protein, CENP-H. Journal of Biological Chemistry *274*, 27343–27346.

Sugata, N., Li, S., Earnshaw, W.C., Yen, T.J., Yoda, K., Masumoto, H., Munekata, E., Warburton, P.E., and Todokoro, K. (2000). Human CENP-H multimers colocalize with CENP-A and CENP-C at active centromere--kinetochore complexes. Human Molecular Genetics *9*, 2919–2926.

Sugimoto, K., Yata, H., Muro, Y., and Himeno, M. (1994). Human centromere protein C (CENP-C) is a DNA-binding protein which possesses a novel DNA-binding motif. Journal of Biochemistry *116*, 877–881.

Sullivan, B.A., and Karpen, G.H. (2004). Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. Nature Structural & Molecular Biology *11*, 1076–1083.

Sullivan, B., and Karpen, G. (2001). Centromere identity in Drosophila is not determined in vivo by replication timing. Journal of Cell Biology *154*, 683–690.

Sullivan, K.F., Hechenberger, M., and Masri, K. (1994). Human CENP-A contains a histone H3 related histone fold domain that is required for targeting to the centromere. The Journal of Cell Biology *3*, 581–592.

Sun, X., Wahlstrom, J., and Karpen, G. (1997). Molecular structure of a functional Drosophila centromere. Cell *91*, 1007–1019.

Tachiwana, H., Kagawa, W., Shiga, T., Osakabe, A., Miya, Y., Saito, K., Hayashi-Takanaka, Y., Oda, T., Sato, M., Park, S.-Y., et al. (2011). Crystal structure of the human centromeric nucleosome containing CENP-A. Nature *476*, 232–235.

Takahashi, K., Chen, E.S., and Yanagida, M. (2000). Requirement of Mis6 centromere connector for localizing a CENP-A-like protein in fission yeast. Science (New York, N.Y.) *288*, 2215–2219.

Takeuchi, K., Nishino, T., Mayanagi, K., Horikoshi, N., Osakabe, A., Tachiwana, H., Hori, T., Kurumizaka, H., and Fukagawa, T. (2014). The centromeric nucleosome-like CENP-T-W-S-X complex induces positive supercoils into DNA. Nucleic Acids Research *42*, 1644–1655.

Thakur, J., Talbert, P.B., and Henikoff, S. (2015). Inner kinetochore protein interactions with regional centromeres of fission yeast. Genetics *201*, 543–561.

Thibaud-Nissen, F., Souvorov, A., Murphy, T., DiCuccio, M., and Kitts, P. (2013). Eukaryotic genome annotation pipeline.

Tomkiel, J., Cooke, C.A., Saitoh, H., Bernat, R.L., and Earnshaw, W.C. (1994). CENP-C is required for maintaining proper kinetochore size and for a timely transition to anaphase. Journal of Cell Biology *125*, 531–545.

Trazzi, S., Bernardoni, R., Diolaiti, D., Politi, V., Earnshaw, W.C., Perini, G., and Della Valle, G. (2002). In vivo functional dissection of human inner kinetochore protein CENP-C. In Journal of Structural Biology, pp. 39–48.

Trazzi, S., Perini, G., Bernardoni, R., Zoli, M., Reese, J.C., Musacchio, A., and Della Valle, G. (2009). The C-terminal domain of CENP-C displays multiple and critical functions for mammalian centromere formation. PLoS ONE *4*.

Trifonov, V.A., Stanyon, R., Nesterenko, A.I., Fu, B., Perelman, P.L., O'Brien, P.C.M., Stone, G., Rubtsova, N. V., Houck, M.L., Robinson, T.J., et al. (2008). Multidirectional

cross-species painting illuminates the history of karyotypic evolution in Perissodactyla. Chromosome Research *16*, 89–107.

Tsunaka, Y., Kajimura, N., Tate, S.I., and Morikawa, K. (2005). Alteration of the nucleosomal DNA path in the crystal structure of a human nucleosome core particle. Nucleic Acids Research *33*, 3424–3434.

Vafa, O., and Sullivan, K.F. (1997). Chromatin containing CENP-A and alpha-satellite DNA is a major component of the inner kinetochore plate. Current Biology : CB *7*, 897–900.

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

Ventura, M., Antonacci, F., Cardone, M.F., Stanyon, R., D'Addabbo, P., Cellamare, A., Sprague, L.J., Eichler, E.E., Archidiacono, N., and Rocchi, M. (2007). Evolutionary formation of new centromeres in macaque. Science (New York, N.Y.) *316*, 243–246.

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

Voullaire, L.E., Slater, H.R., Petrovic, V., and Choo, K.H. (1993). A functional marker centromere with no detectable alpha-satellite, satellite III, or CENP-B protein: activation of a latent centromere? American Journal of Human Genetics *52*, 1153–1163.

Wade, C.M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., Lear, T.L., Adelson, D.L., Bailey, E., Bellone, R.R., et al. (2009). Genome sequence, comparative analysis, and population genetics of the domestic horse. Science (New York, N.Y.) *326*, 865–867.

Warburton, P.E. (2004). Chromosomal dynamics of human neocentromere formation. Chromosome Research *12*, 617–626.

Waye, J.S., and Willard, H.F. (1985). Chromosome-specific alpha satellite DNA: Nucleotide sequence analysis of the 2.0 kilobasepair repeat from the human X chromosome. Nucleic Acids Research *13*, 2731–2743.

Weiler, K.S., and Wakimoto, B.T. (1995). Heterochromatin and gene expression in Drosophila. Annual Review of Genetics *29*, 577–605.

Westhorpe, F.G., and Straight, A.F. (2015). The centromere: Epigenetic control of chromosome segregation during mitosis. Cold Spring Harbor Perspectives in Biology *7*.

Westhorpe, F.G., Fuller, C.J., and Straight, A.F. (2015). A cell-free CENP-A assembly system defines the chromatin requirements for centromere maintenance. The Journal of Cell Biology *209*, 789–801.

Willard, H.F. (1985). Chromosome-specific organization of human alpha satellite DNA. American Journal of Human Genetics *37*, 524–532.

Yan, H., Talbert, P.B., Lee, H.R., Jett, J., Henikoff, S., Chen, F., and Jiang, J. (2008). Intergenic locations of rice centromeric chromatin. PLoS Biology *6*, 2563–2575.

Yoda, K., Ando, S., Morishita, S., Houmura, K., Hashimoto, K., Takeyasu, K., and Okazaki, T. (2000). Human centromere protein A (CENP-A) can replace histone H3 in nucleosome reconstitution in vitro. Proc. Natl. Acad. Sci. USA *97*, 7266–7271.

Zasadzińska, E., Barnhart-Dailey, M.C., Kuich, P.H.J.L., and Foltz, D.R. (2013). Dimerization of the CENP-A assembly factor HJURP is required for centromeric nucleosome deposition. The EMBO Journal *32*, 2113–2124.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J.J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biology *9*, R137.

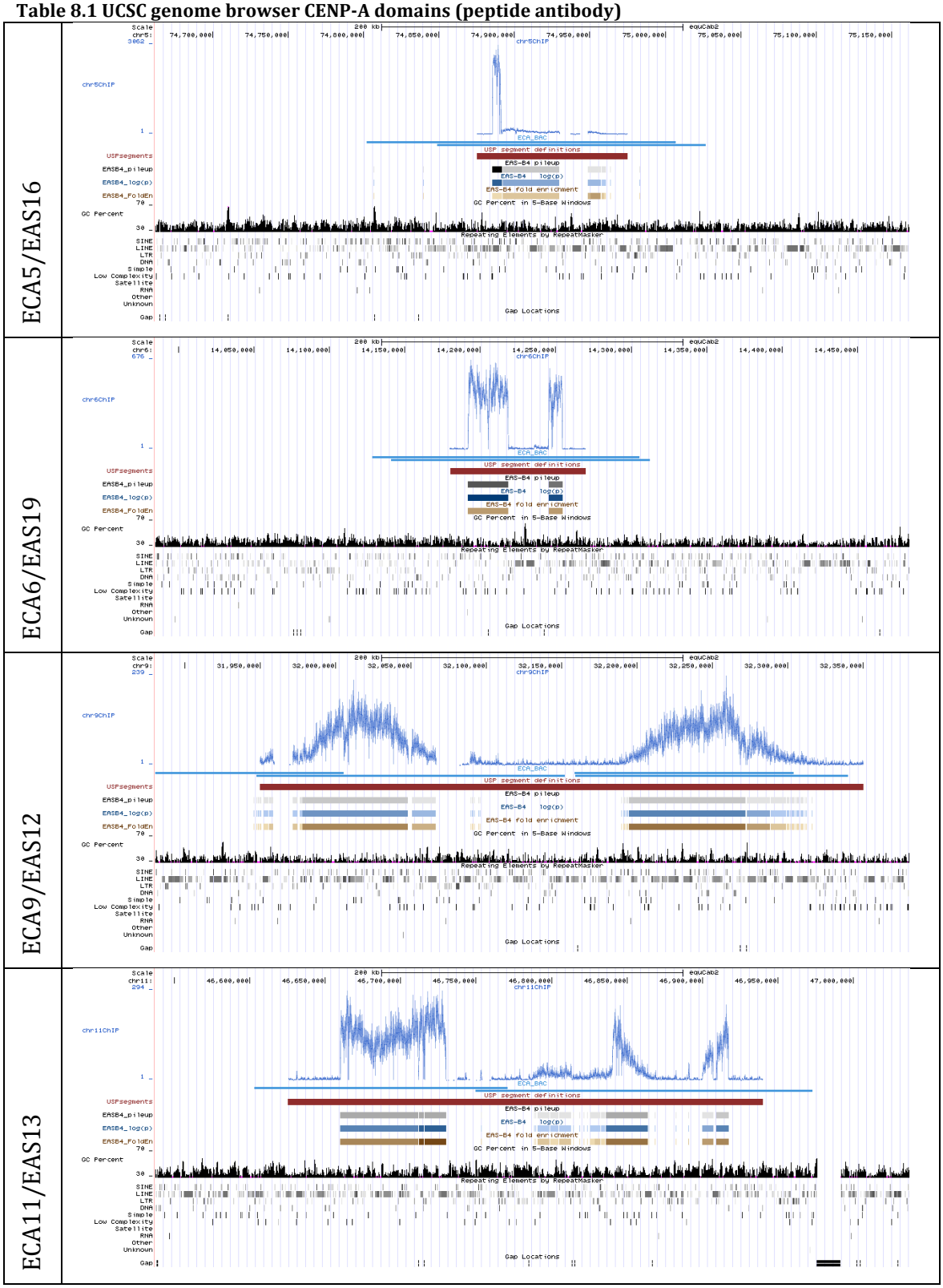
Zhou, Z., Feng, H., Zhou, B.-R., Ghirlando, R., Hu, K., Zwolak, A., Miller Jenkins, L.M., Xiao, H., Tjandra, N., Wu, C., et al. (2011). Structural basis for recognition of centromere histone variant CenH3 by the chaperone Scm3. Nature *472*, 234–237.

Zinkowski, R.P., Meyne, J., and Brinkley, B.R. (1991). The centromere-kinetochore complex: A repeat subunit model. Journal of Cell Biology *113*, 1091–1110.

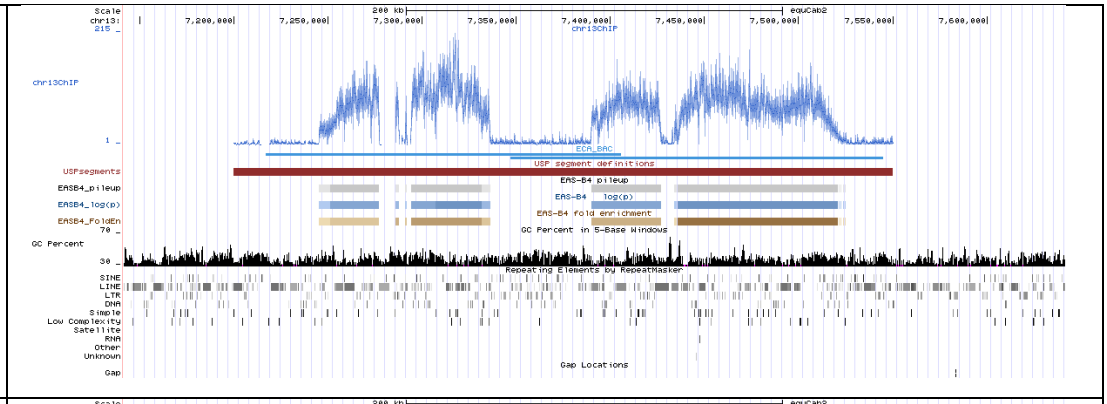
Chapter 8 Appendices

8.1 Appendix I

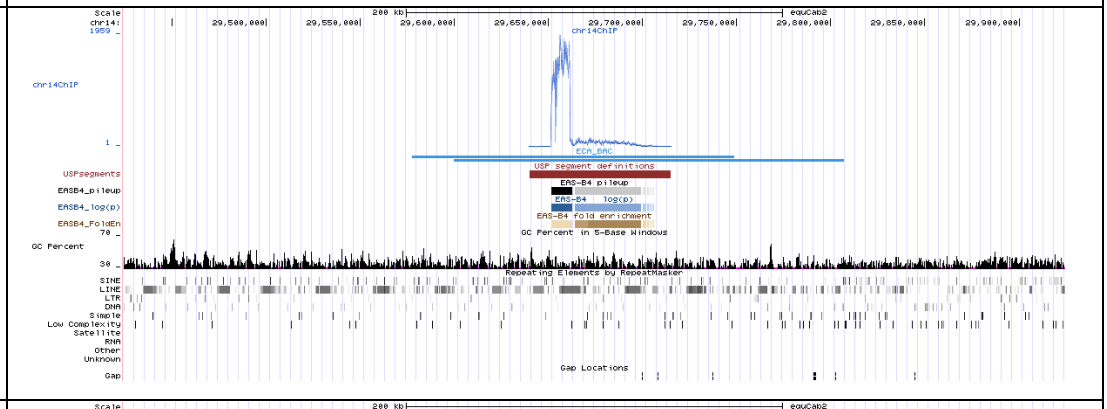
Table 8.1 UCSC genome browser CENP-A domains (peptide antibody)



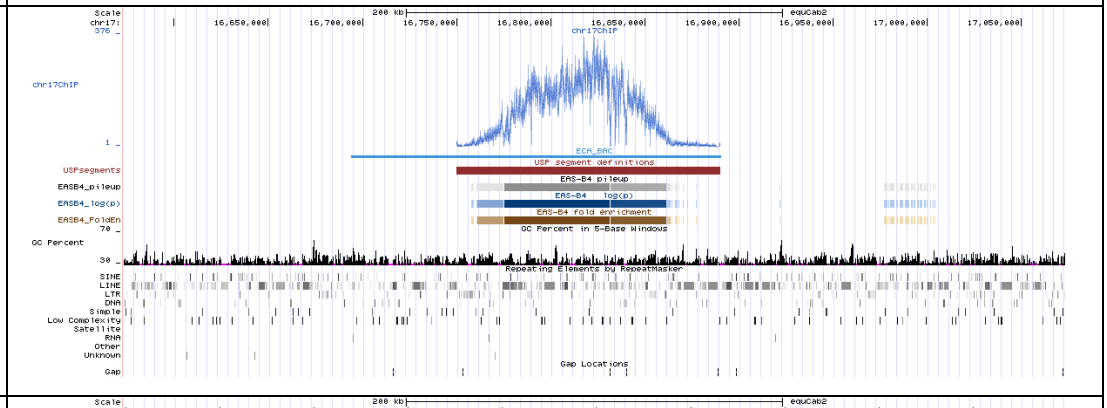
ECA113/EAS14



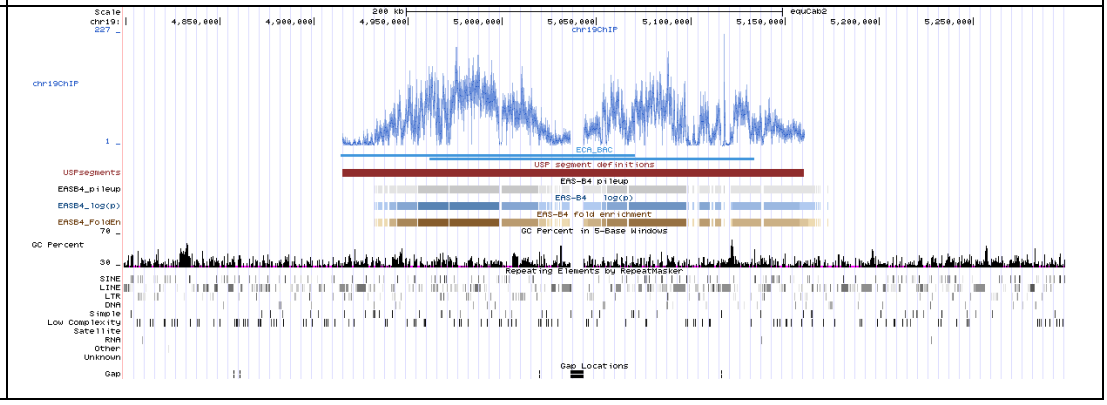
ECA14/EAS9



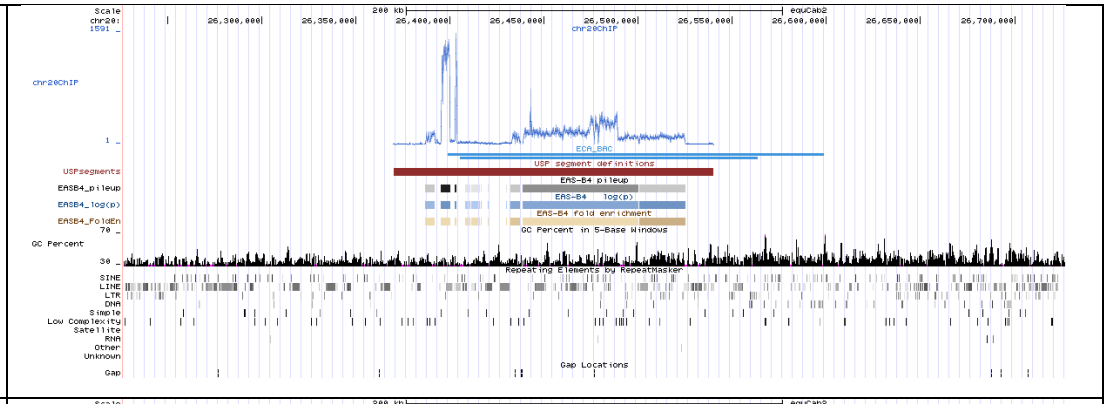
EAC17/EAS11



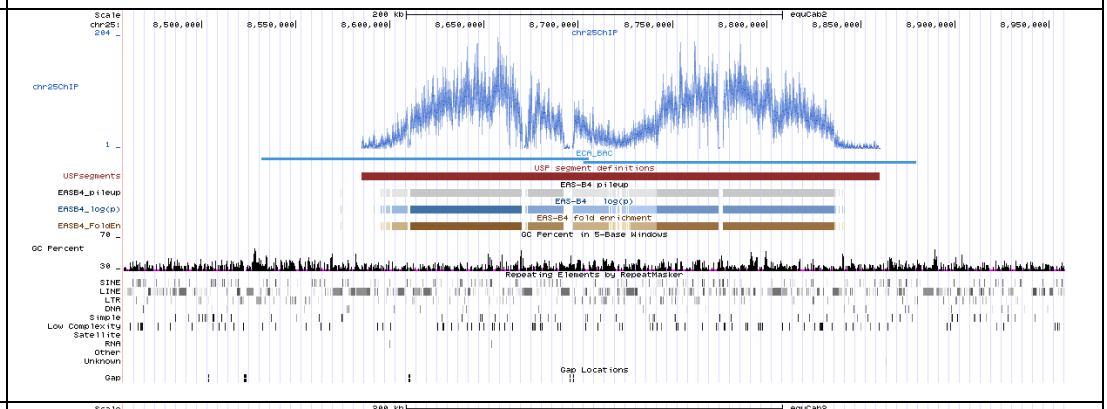
ECA19/EAS5



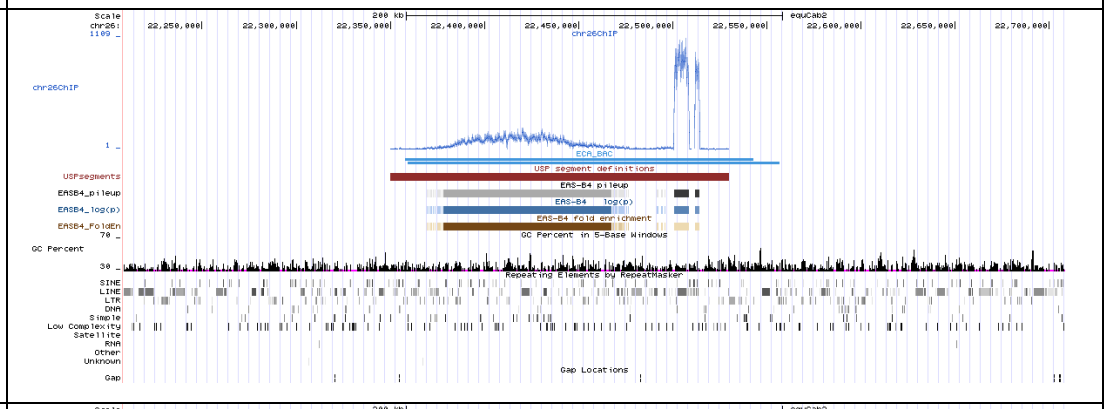
EAC20/EAS8



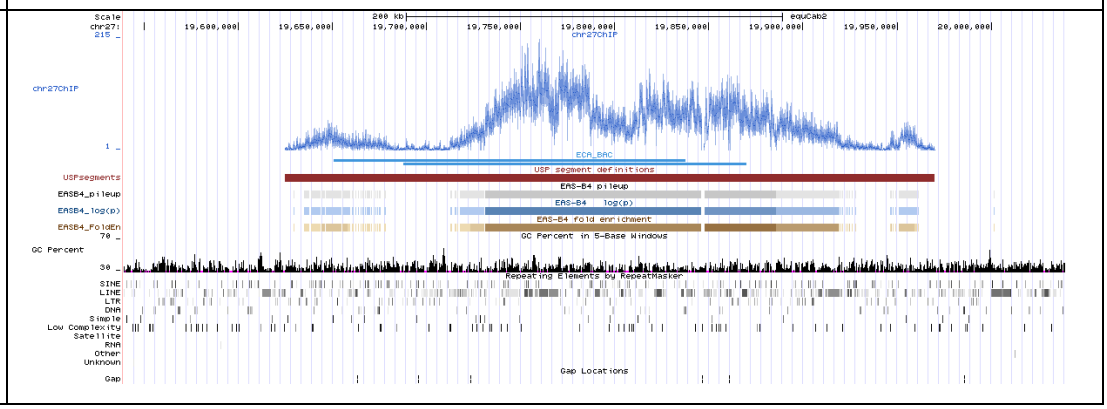
ECA25/EAS10



ECA26/EAS18



ECA27/EAS27



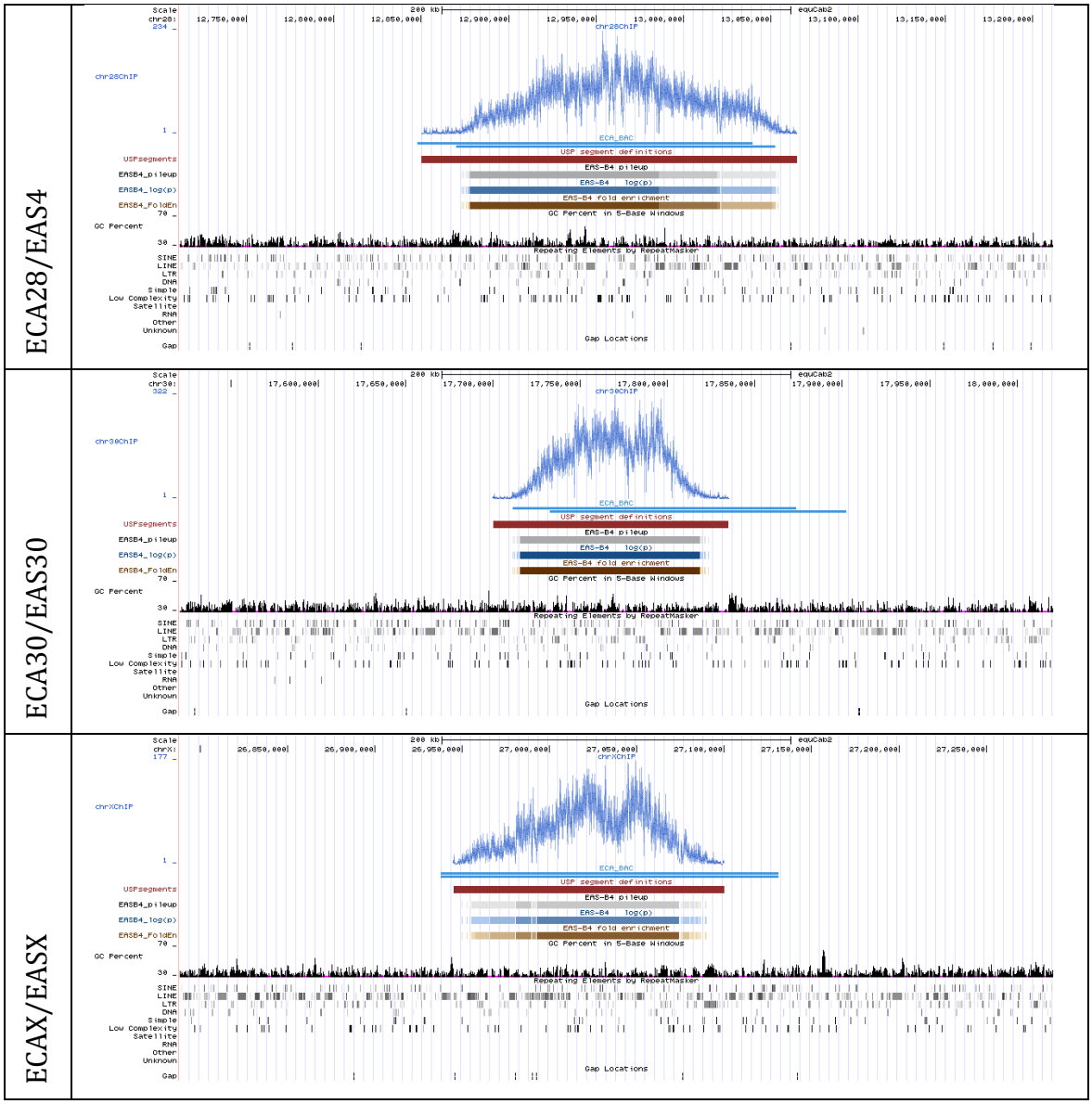


Table 8.1 displays the CENP-A data track files visualised on the UCSC genome browser. The UCSC genome browser served as a resource for collating the ChIP-seq data originally mapped back to the horse genome *EquCab2.0* before the construction of the hybrid genomes.

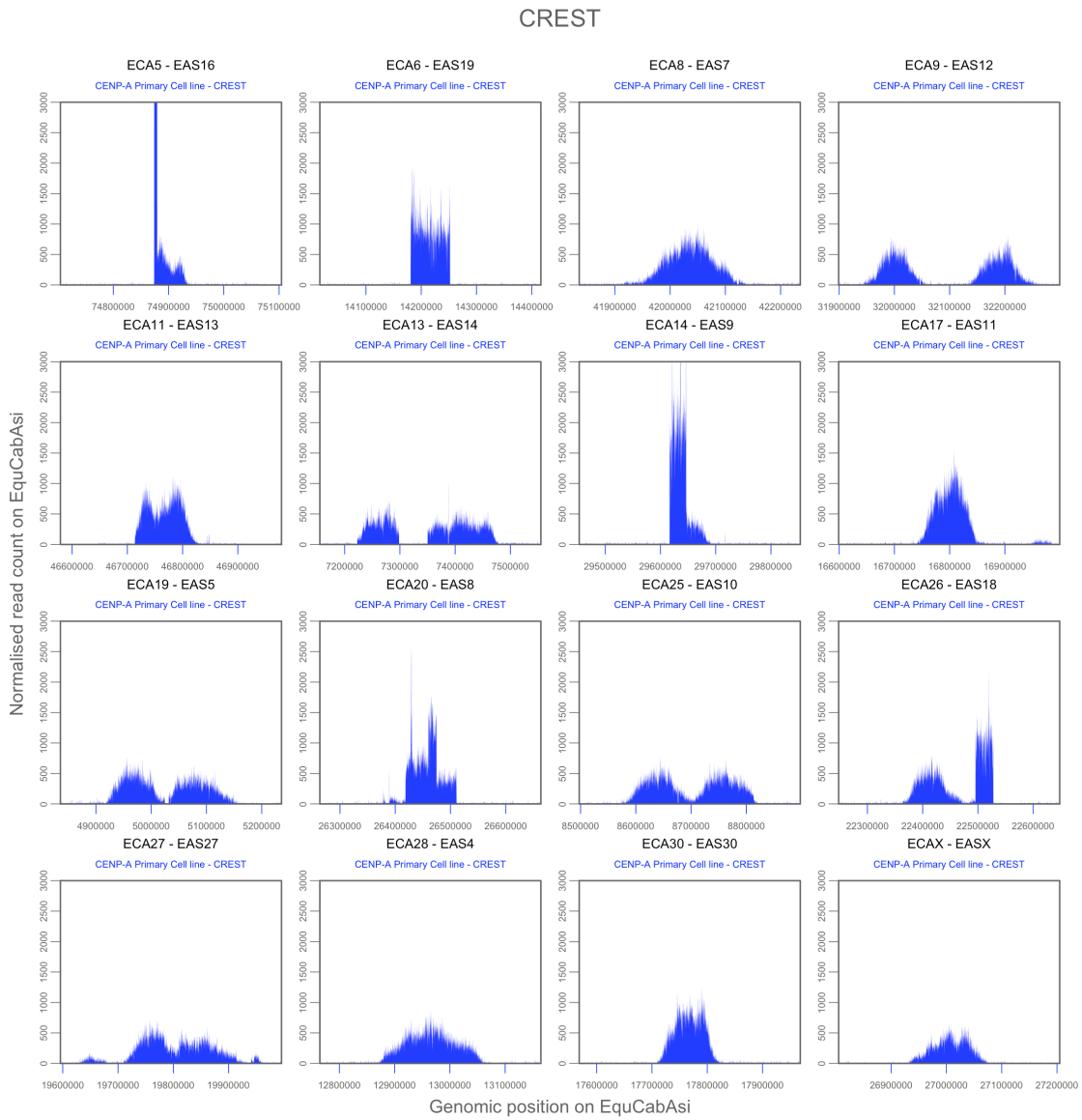


Figure 8.1 CREST sera CENP-A domains
16 CENP-A distributions obtained from the CREST CENP-A ChIP-seq data mapped to the hybrid genome *EquCabAsi*. Horse chromosomes and corresponding donkey chromosomes above each profile. The CENP-A domains (400 kb window views) display four types of distributions, "Gaussian-like", "Spike-like", "Complex" and "Multi-domain".

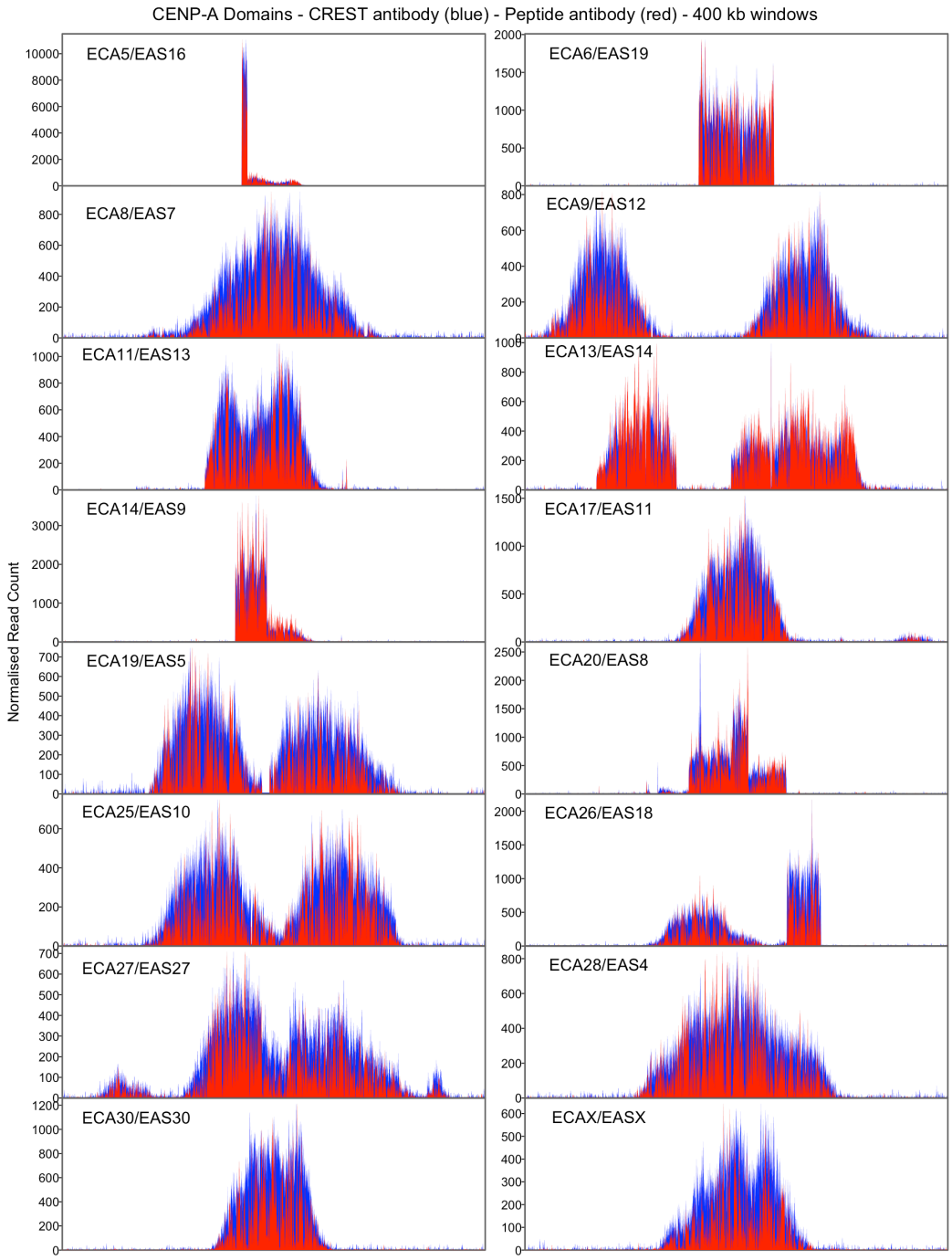
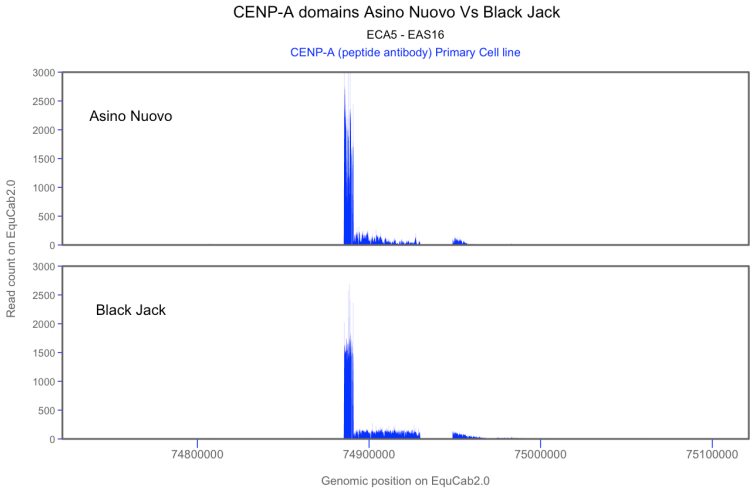


Figure 8.2 Co-localisation of Peptide antibody data and CREST sera CENP-A domains
16 CENP-A distributions obtained from the CENP-A peptide ChIP-seq compared with the CREST CENP-A ChIP-seq data, mapped to the hybrid genome *EquCabAsi*. Horse chromosomes and corresponding donkey chromosomes above and left of each profile. The CENP-A domains (400 kb window views) show identical co-localization of all 16 domains.

A



B

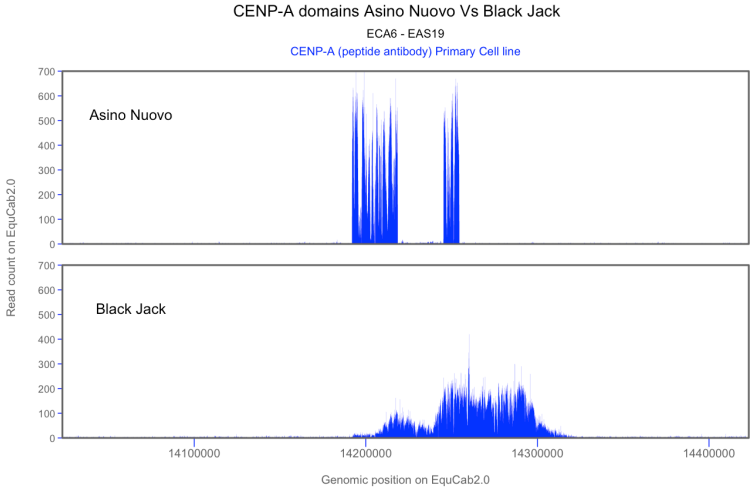
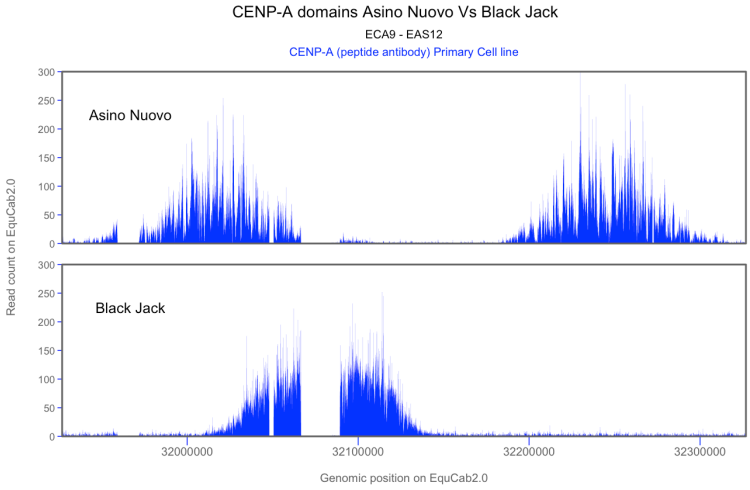


Figure 8.3 Positional variation in unrelated individuals – ECA5/EAS16, ECA6/EAS19
CENP-A domains are localised in different positions in two unrelated donkey individuals. ChIP-seq data from AN (top) and BJ (bottom) were aligned to *EquCab2.0*. ECA5/EAS16 (A) shows very little positional variation between two individuals. (B) ECA6/EAS19 shows a rightward shift in BJ which exhibits a Gaussian-like domain compared to a spike-like peak in AN.

A



B

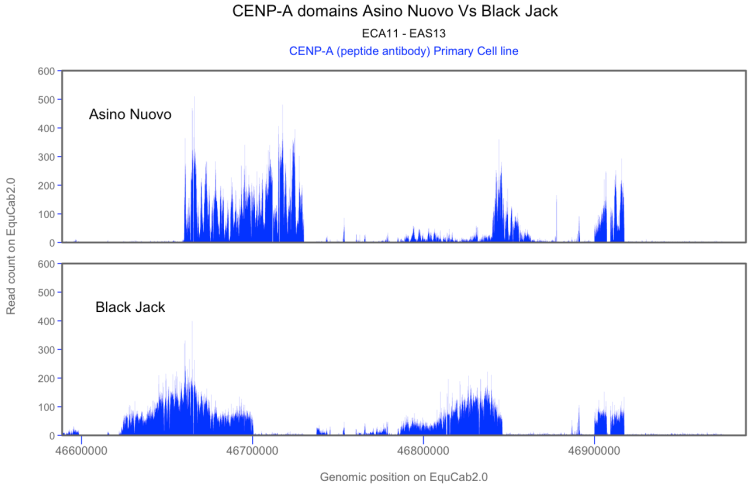
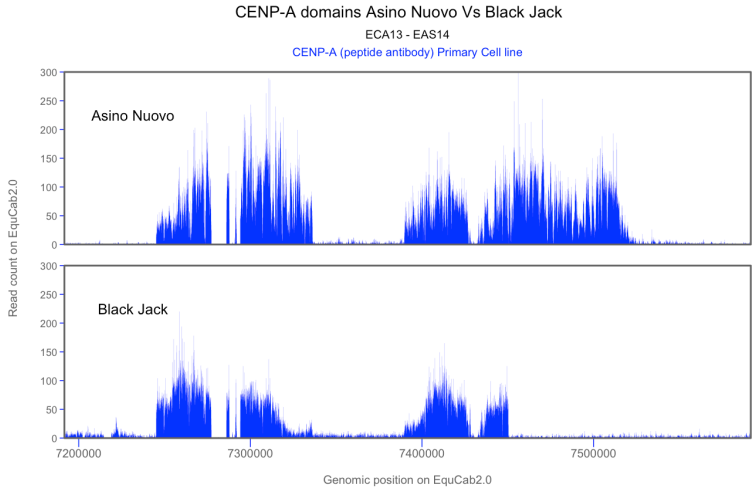


Figure 8.4 Positional variation in unrelated individuals – ECA9/EAS12, ECA11/EAS13
CENP-A domains are localised in different positions in two unrelated donkey individuals. ChIP-seq data from AN (top) and BJ (bottom) were aligned to *EquCab2.0*. ECA9/EAS12 (A) shows very little positional variation between two individuals. (B) ECA11/EAS13 positional variation between the two donkey individuals is observed.

A



B

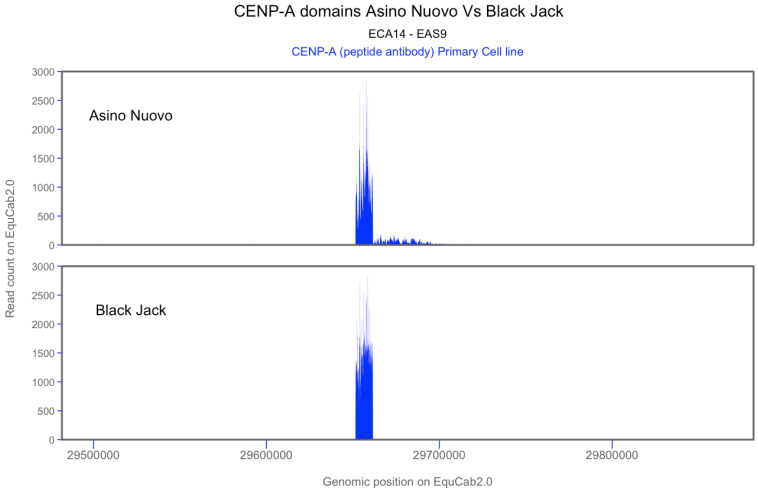
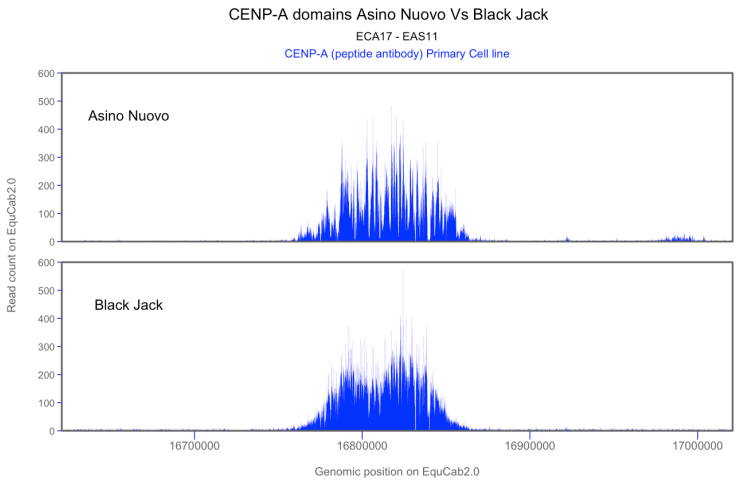


Figure 8.5 Positional variation in unrelated individuals – ECA13/EAS14, ECA14/EAS9
CENP-A domains are localised in different positions in two unrelated donkey individuals. ChIP-seq data from AN (top) and BJ (bottom) were aligned to *EquCab2.0*. ECA13/EAS14 (A) shows positional variation between two individuals. The right allele in BJ occupies a smaller span than in AN. (B) ECA14/EAS9 positional variation between the two donkey individuals is observed. The Gaussian-like subdomain is not present in BJ, indicating the spike peak contains all centromere function.

A



B

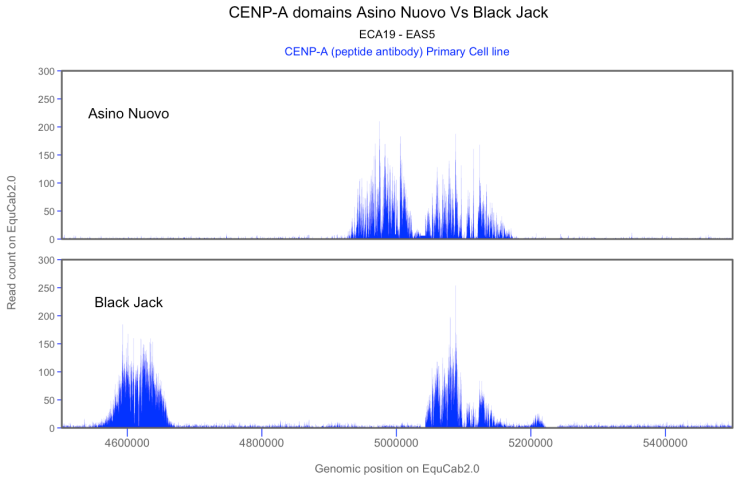
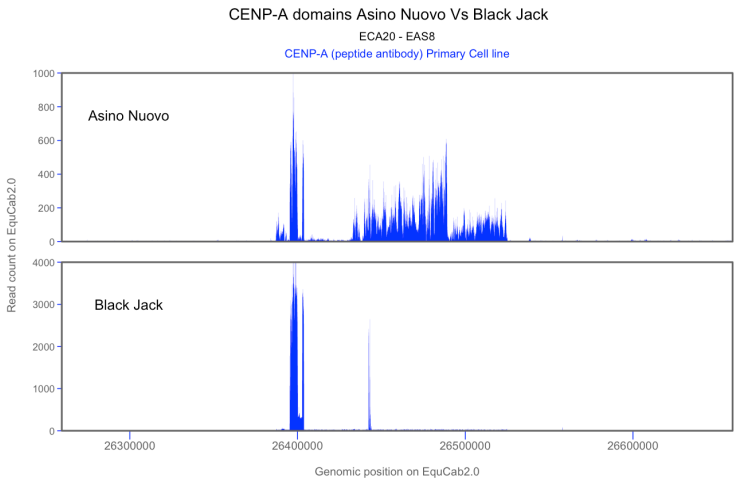


Figure 8.6 Positional variation in unrelated individuals – ECA17/EAS11, ECA19/EAS5
CENP-A domains are localised in different positions in two unrelated donkey individuals. ChIP-seq data from AN (top) and BJ (bottom) were aligned to *EquCab2.0*. ECA17/EAS11 (A) shows very little evidence positional variation between two individuals. (B) ECA19/EAS5 positional variation is observed. The left allele in BJ is ~45 kb apart from the right allele. In AN, both alleles are adjacent to one another.

A



B

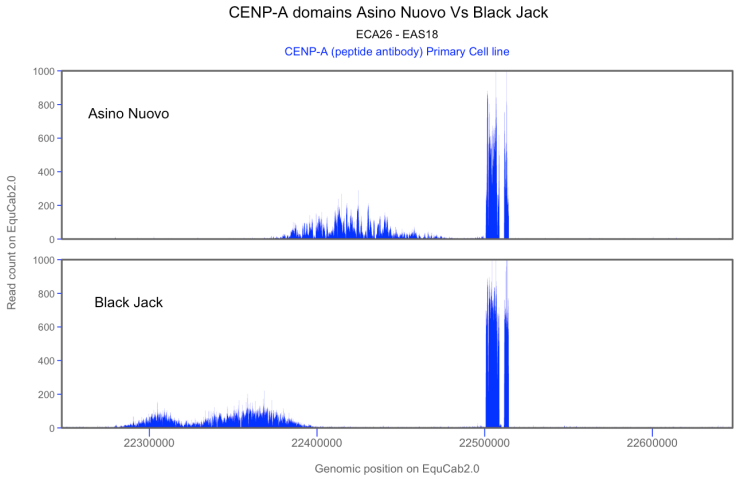
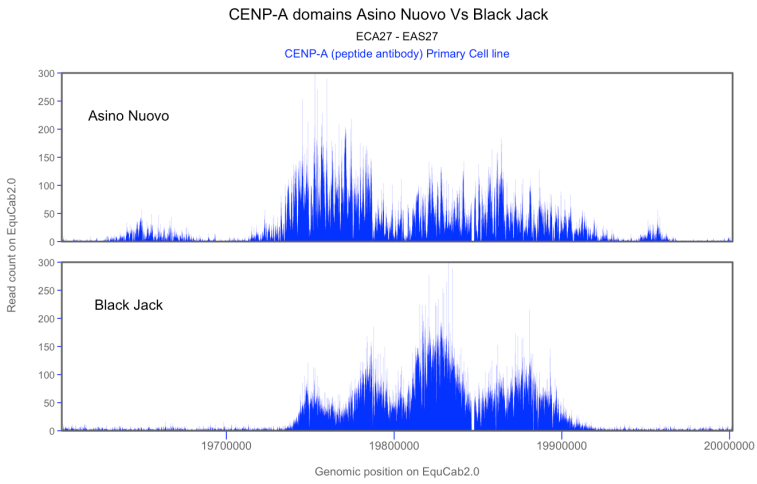


Figure 8.7 Positional variation in unrelated individuals – ECA20/EAS8, ECA26/EAS18
CENP-A domains are localised in different positions in two unrelated donkey individuals. ChIP-seq data from AN (top) and BJ (bottom) were aligned to *EquCab2.0*. ECA20/EAS8 (A) shows centromere function is contained within the spike in BJ and spread over two domains in AN. (B) ECA26/EAS18 positional variation is observed. The left allele in BJ is ~100 kb apart from the right allele. In AN, both alleles are adjacent to one another. The spikes in both AN and BJ co-localize to the same region.

A



B

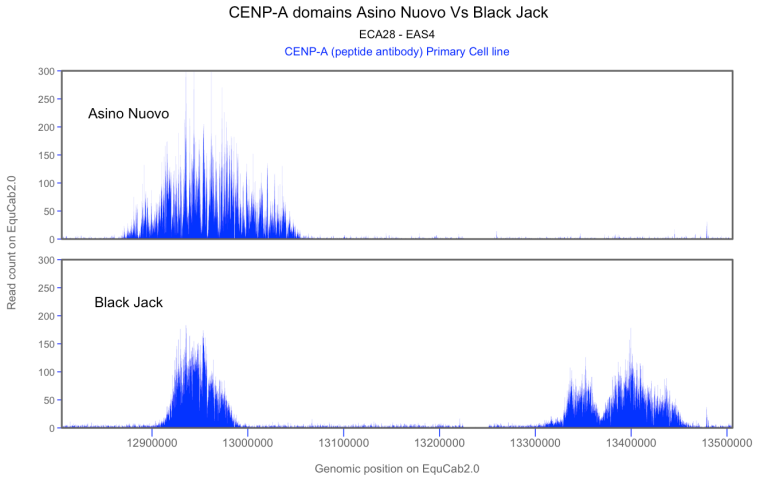
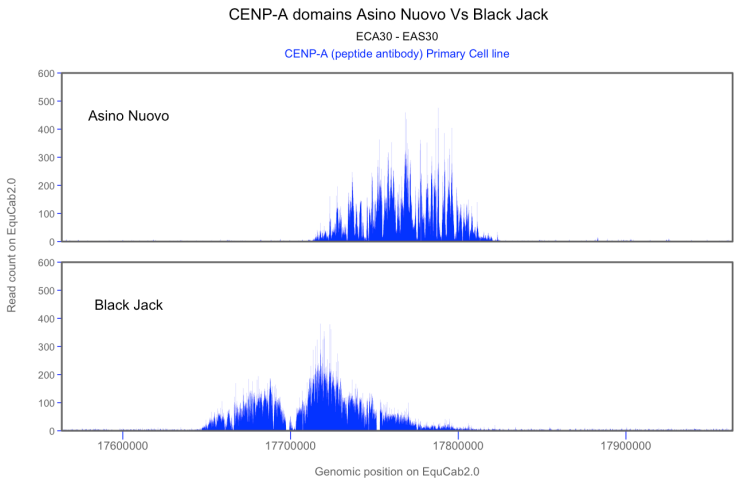


Figure 8.8 Positional variation in unrelated individuals – ECA27/EAS27, ECA28/EAS4
CENP-A domains are localised in different positions in two unrelated donkey individuals. ChIP-seq data from AN (top) and BJ (bottom) were aligned to *EquCab2.0*. ECA27/EAS27 (A) shows positional variation between both donkeys. Both domains profile a multi-domain distribution indicating overlapping centromere alleles. (B) ECA28/EAS4 positional variation is observed. The centromere in AN is contained in one region with overlapping centromere alleles, however, in BJ the centromere alleles are separated by ~300 kb.

A



B

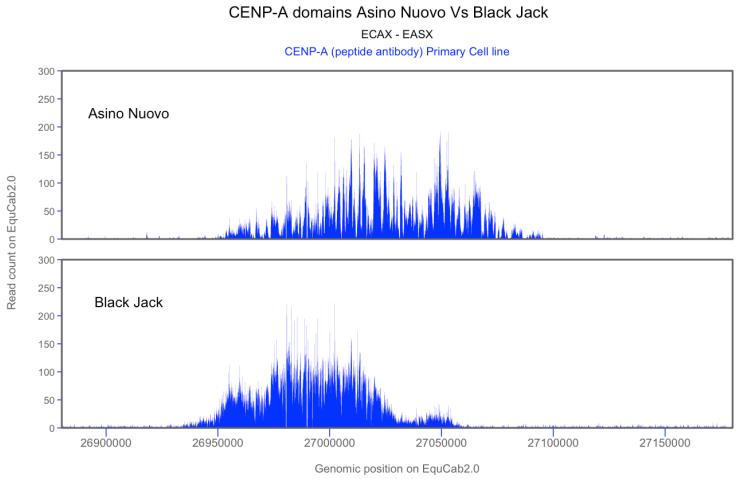


Figure 8.9 Positional variation in unrelated individuals – ECA30/EAS30, ECAX/EASX
CENP-A domains are localised in different positions in two unrelated donkey individuals. ChIP-seq data from AN (top) and BJ (bottom) were aligned to *EquCab2.0*. ECA30/EAS30 (A) shows a leftward shit in the BJ centromere compared to AN. (B) ECAX/EASX positional variation is observed. A leftward shift is observed in BJ compared to AN.

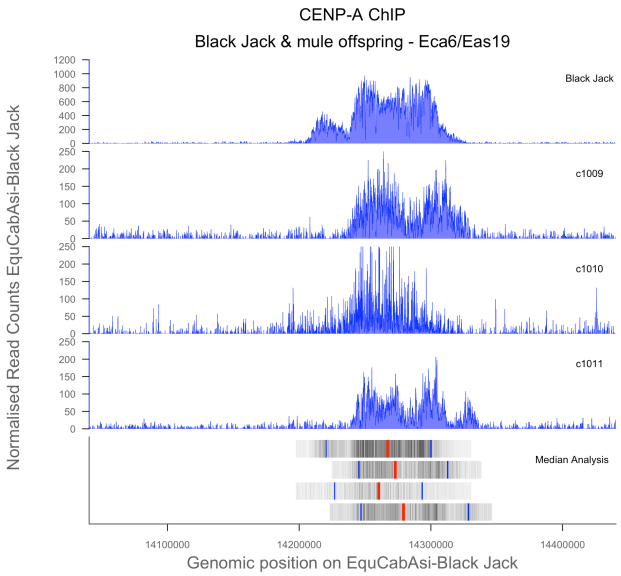


Figure 8.10 Centromere sliding analysis – ECA6/EAS19
Sliding analyses showing centromere movement across family datasets in ECA6/EAS19. A 7-12 kb displacement was observed in ECA6/EAS19. Displacement determined by median analysis in bottom panel. Red bars in bottom panel denote peak centre of gravity calculated by determining the positional median. Blue bars represent centromere boundaries determined by calculating the exact positions at which 5% and 95% of the integrated ChIP signal are located. ChIP signal across each domain displayed as heatmap of read counts (grey).

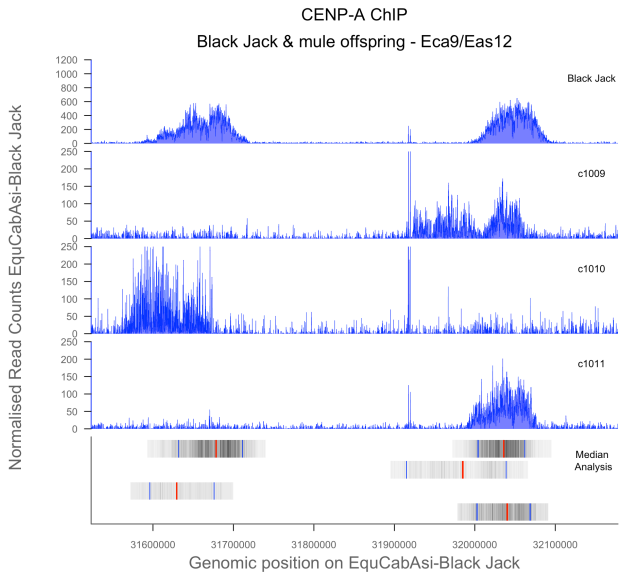


Figure 8.11 Centromere sliding analysis – ECA9/EAS12
Sliding analyses showing centromere movement across family datasets in ECA9/EAS12. An 11-52 kb displacement was observed across ECA9/EAS12 between paternal CENP-A and corresponding offspring alleles. C1009 and c1010 showed a 52 kb and 57 kb displacement respectively. Displacement determined by median analysis, represented in bottom panel.

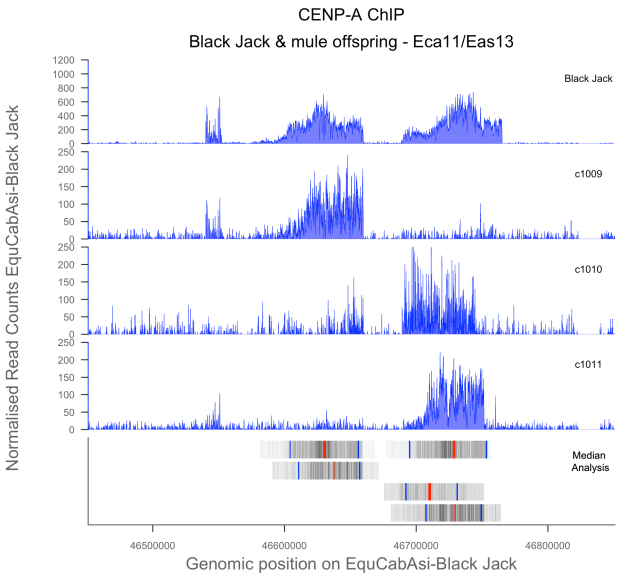


Figure 8.12 Centromere sliding analysis – ECA11/EAS13
Sliding analyses showing centromere movement across family datasets in ECA11/EAS13. A 5-20 kb displacement was observed across ECA11/EAS13 between paternal CENP-A and corresponding offspring alleles. Displacement determined by median analysis, represented in bottom panel.

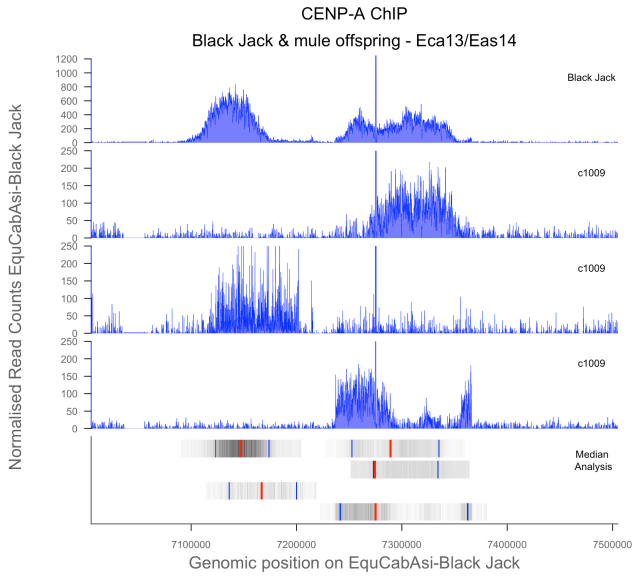


Figure 8.13 Centromere sliding analysis – ECA13/EAS14
Sliding analyses showing centromere movement across family datasets in ECA13/EAS14. A 16-21 kb displacement was observed across ECA13/EAS14 between paternal CENP-A and corresponding offspring alleles. Displacement determined by median analysis, represented in bottom panel.

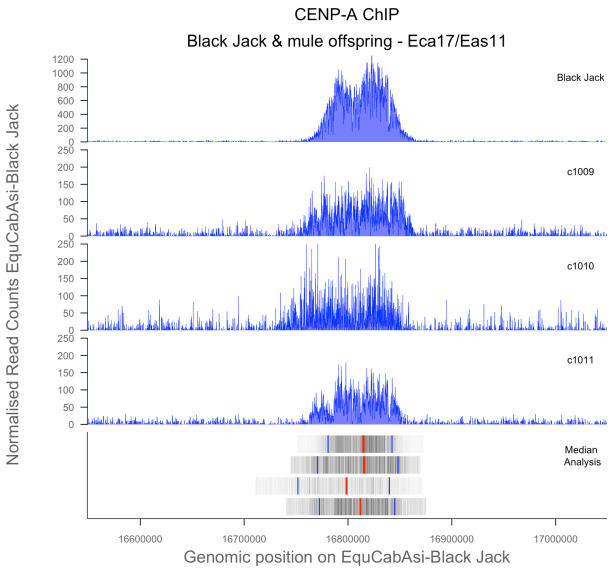


Figure 8.14 Centromere sliding analysis – ECA17/EAS11
Sliding analyses showing centromere movement across family datasets in ECA17/EAS11. A 1-19 kb displacement was observed across ECA17/EAS11 between paternal CENP-A and corresponding offspring alleles. C1009 and c1011 showed a small displacement of 1 kb and 5 kb compared to a larger displacement of 19 kb inc1010. Displacement determined by median analysis, represented in bottom panel.

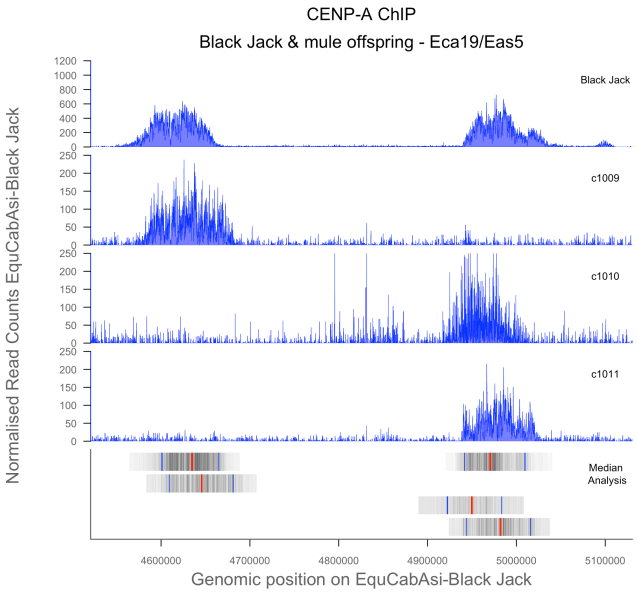


Figure 8.15 Centromere sliding analysis – ECA19/EAS5
Sliding analyses showing centromere movement across family datasets in ECA19/EAS5. A 1-18 kb displacement was observed across ECA19/EAS5 between paternal CENP-A and corresponding offspring alleles. C1011 showed a small displacement of 1 kb. C1009 showed a 13 kb displacement compared to a larger displacement of 18 kb in c1010. Displacement determined by median analysis, represented in bottom panel.

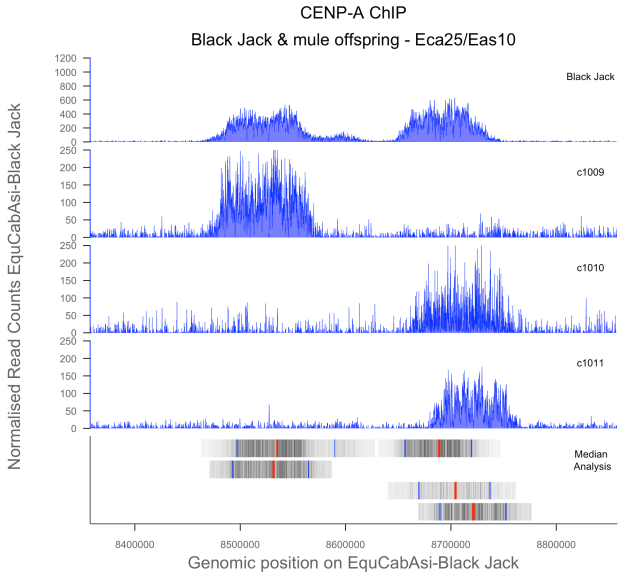


Figure 8.16 Centromere sliding analysis – ECA25/EAS10
Sliding analyses showing centromere movement across family datasets in ECA25/EAS10. A 4-26 kb displacement was observed across ECA25/EAS10 between paternal CENP-A and corresponding offspring alleles. C1009 showed a small displacement of 4 kb. C1010 showed a 17 kb displacement compared to a larger displacement of 26 kb in c1011. Displacement determined by median analysis, represented in bottom panel.

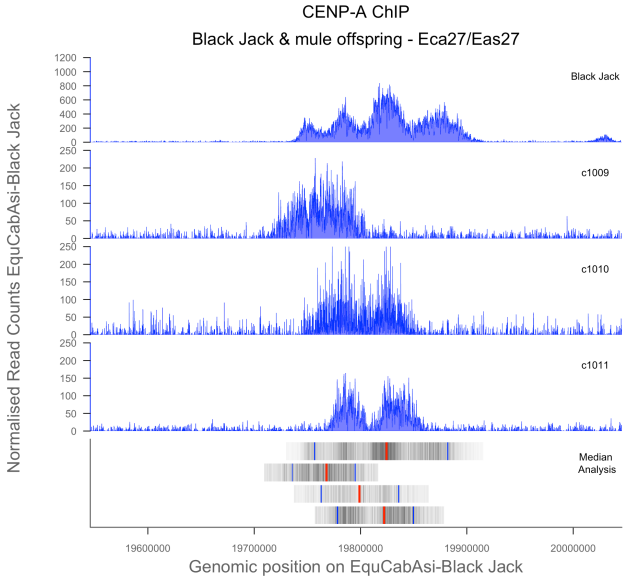


Figure 8.17 Centromere sliding analysis – ECA27/EAS27
Sliding analyses showing centromere movement across family datasets in ECA27/EAS27. A 5-62 kb displacement was observed across ECA27/EAS27 between paternal CENP-A and corresponding offspring alleles. C1011 showed a small displacement of 5 kb. C1010 showed a 26 kb displacement compared to a larger displacement of 62 kb in c1011. Displacement determined by median analysis, represented in bottom panel.

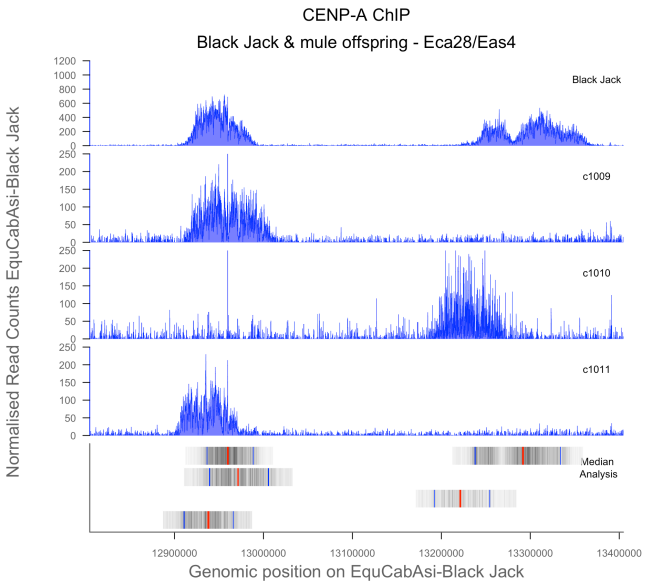


Figure 8.18 Centromere sliding analysis – ECA28/EAS4
Sliding analyses showing centromere movement across family datasets in ECA28/EAS4. A 11-78 kb displacement was observed across ECA28/EAS4 between paternal CENP-A and corresponding offspring alleles. C1011 showed the smallest displacement of 11 kb. C1011 showed a 26 kb displacement compared to a larger displacement of 78 kb in c1010. ECA28/EAS4 showed the largest displacement in the entire dataset. Displacement determined by median analysis, represented in bottom panel.

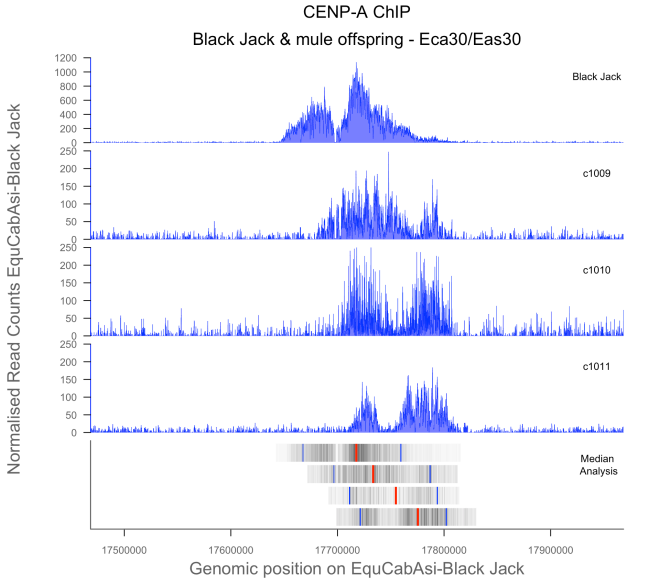


Figure 8.19 Centromere sliding analysis – ECA30/EAS30
Sliding analyses showing centromere movement across family datasets in ECA30/EAS30. A 19-57 kb displacement was observed across ECA30/EAS30 between paternal CENP-A and corresponding offspring alleles. C1009 showed the smallest displacement of 19 kb. C1010 showed a 46 kb displacement compared to a larger displacement of 57 kb in c1010. ECA30/EAS30 is one of the high displacement centromeres. Displacement determined by median analysis, represented in bottom panel.

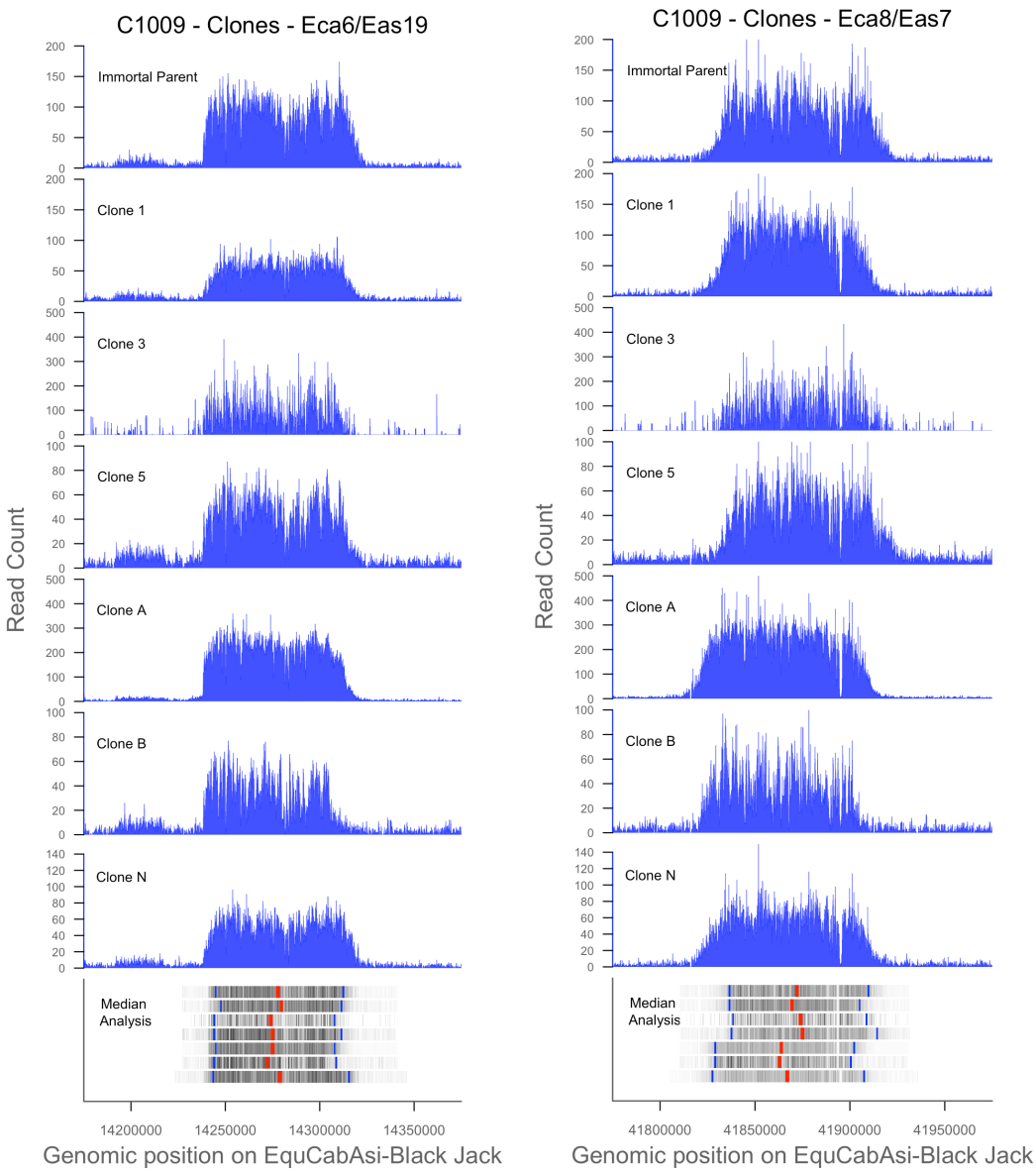


Figure 8.20 Centromere displacement in clonal cell lines – ECA6/EAS19 & ECA8/EAS7
Displacement analyses showing varying levels of centromere movement. Panels display seven independent CENP-A ChIP-seq experiments, performed on asynchronous populations, derived from single cell clones including the parental cell line. Displacement observed in ECA6/EAS19 (left) between 1-6 kb and 2-10 kb in ECA8/EAS7 (right). Red bars in bottom panel denote peak centre of gravity calculated by determining the positional median. Blue bars represent centromere boundaries determined by calculating the exact positions at which 5% and 95% of the ChIP signal are located. ChIP signal across each domain displayed as heatmap (grey).

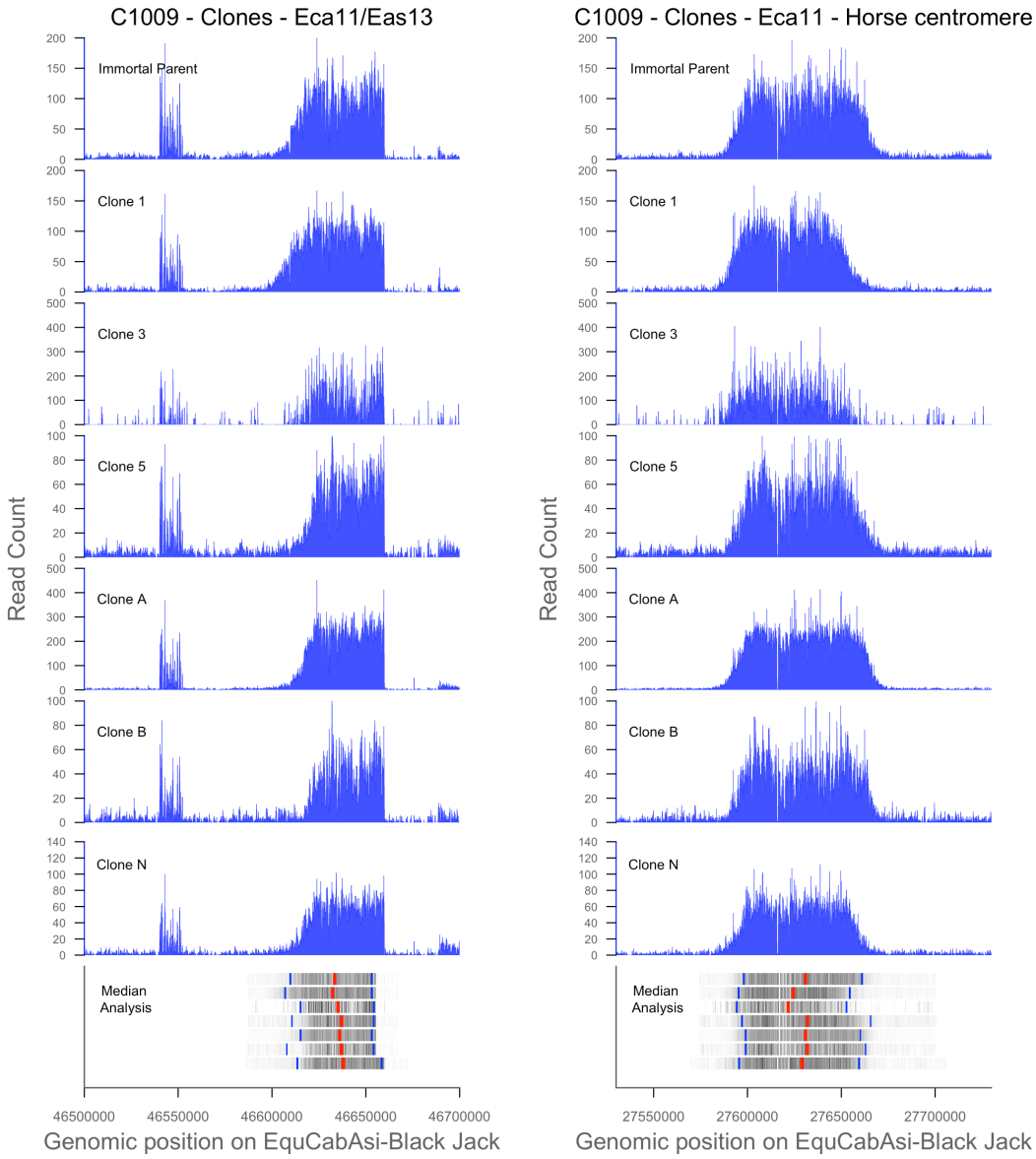


Figure 8.21 Centromere displacement in clonal cell lines – ECA11/EAS13 & ECA11- horse
Displacement analyses showing varying levels of centromere movement. Panels display seven independent CENP-A ChIP-seq experiments, performed on asynchronous populations, derived from single cell clones including the parental cell line. Displacement observed in ECA11/EAS13 (left) between 1-4 kb and 1-10 kb in ECA11-horse (right). Red bars in bottom panel denote peak centre of gravity calculated by determining the positional median. Blue bars represent centromere boundaries determined by calculating the exact positions at which 5% and 95% of the ChIP signal are located. ChIP signal across each domain displayed as heatmap (grey).

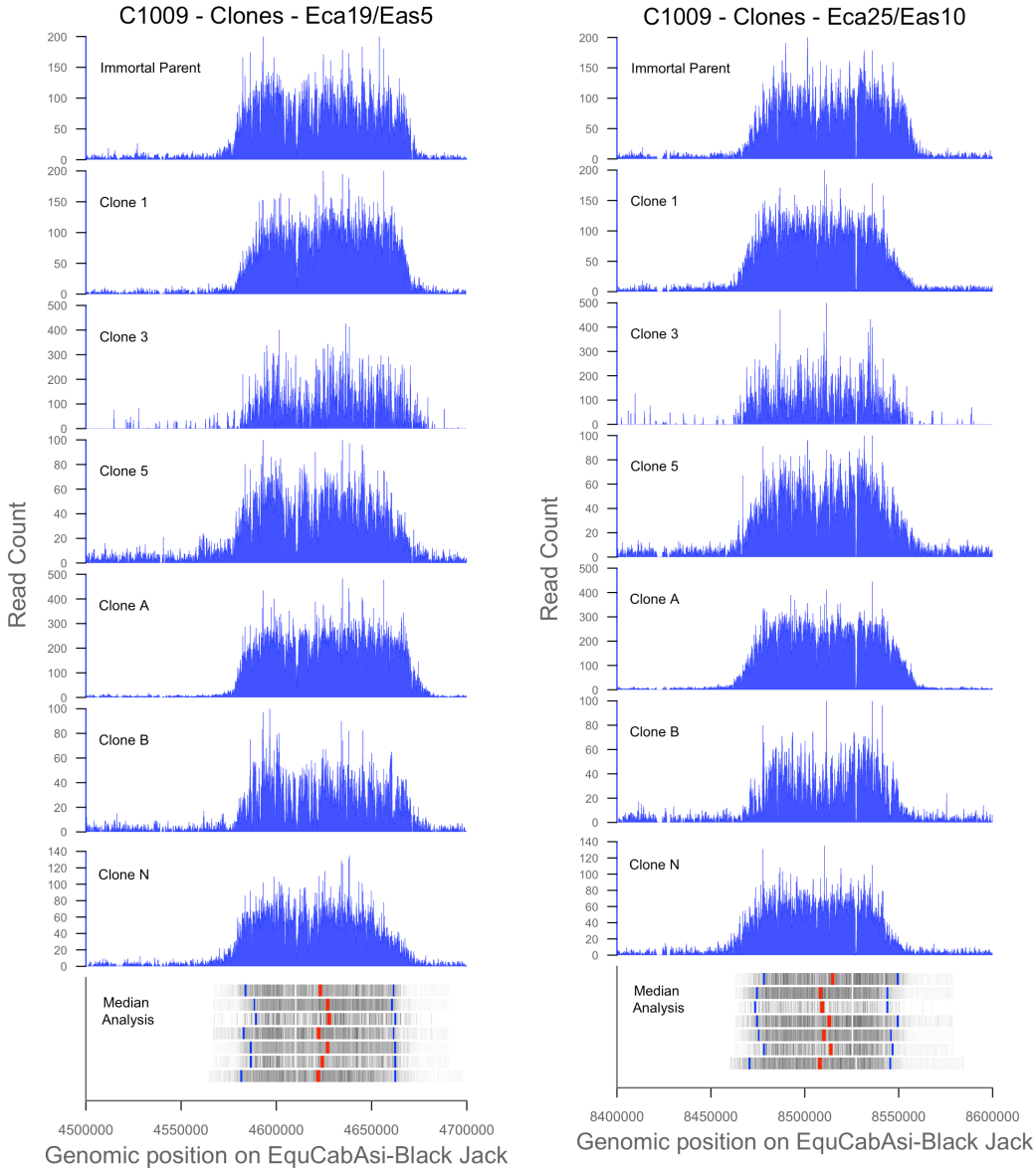


Figure 8.22 Centromere displacement in clonal cell lines – ECA19/EAS5 & ECA25/EAS10
Displacement analyses showing little variation in centromere movement. Panels display seven independent CENP-A ChIP-seq experiments, performed on asynchronous populations, derived from single cell clones including the parental cell line. Displacement observed in ECA19/EAS5 (left) between 1-5 kb and 1-8 kb in ECA25/EAS10 (right). Red bars in bottom panel denote peak centre of gravity calculated by determining the positional median. Blue bars represent centromere boundaries determined by calculating the exact positions at which 5% and 95% of the ChIP signal are located. ChIP signal across each domain displayed as heatmap (grey).

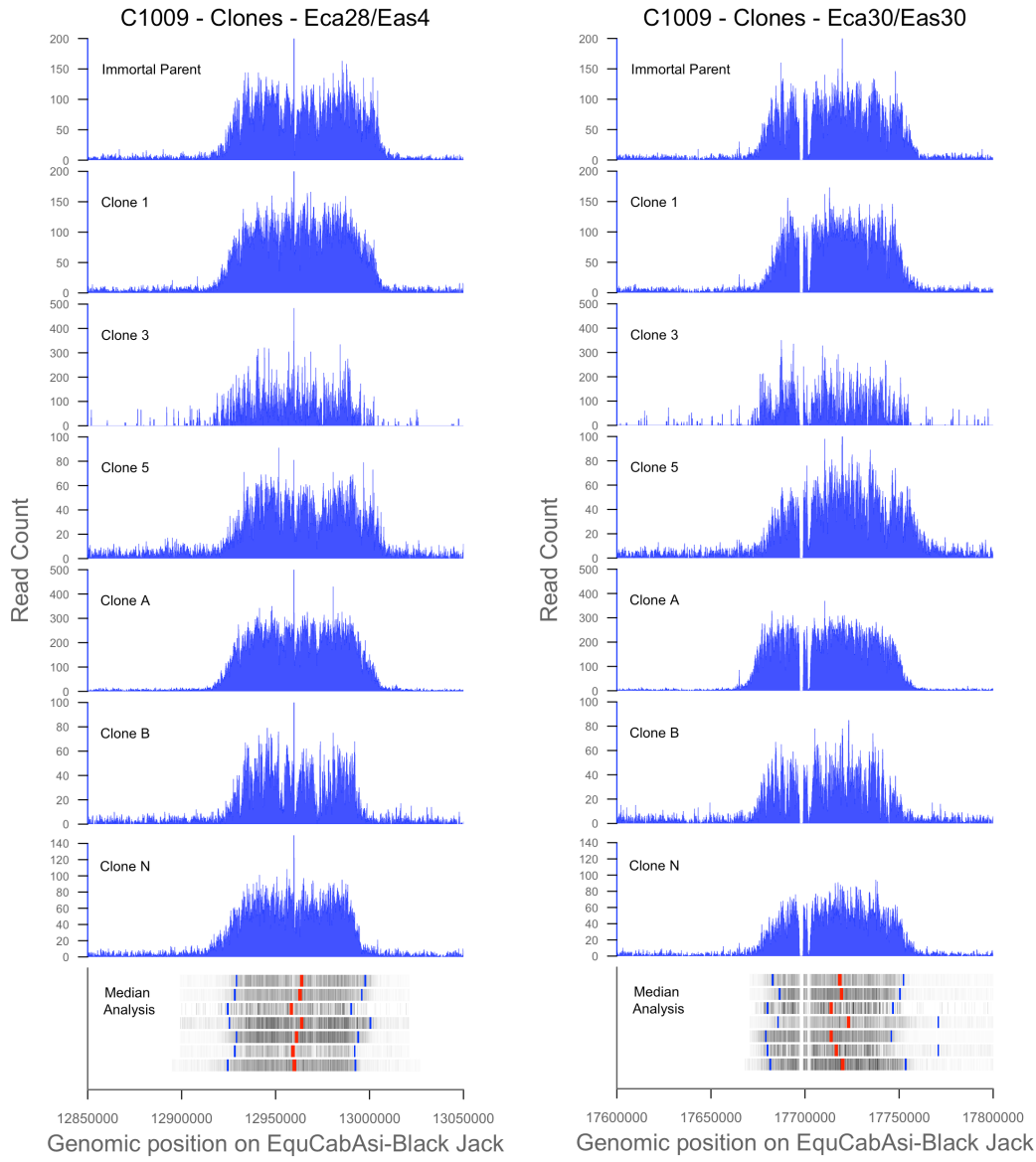


Figure 8.23 Centromere displacement in clonal cell lines – ECA28/EAS4 & ECA30/EAS30
Displacement analyses show little variation in centromere movement. Panels display seven independent CENP-A ChIP-seq experiments, performed on asynchronous populations, derived from single cell clones including the parental cell line. Displacement observed in ECA28/EAS4 (left) between 1-6 kb and 1-5 kb in ECA30/EAS30 (right). Red bars in bottom panel denote peak centre of gravity calculated by determining the positional median. Blue bars represent centromere boundaries determined by calculating the exact positions at which 5% and 95% of the ChIP signal are located. ChIP signal across each domain displayed as heatmap (grey).

8.2 Appendix II

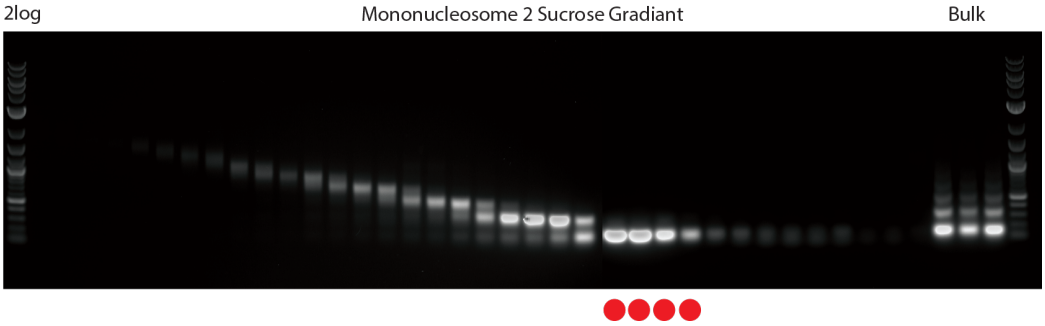


Figure 8.24 Mononucleosome 2 - Sucrose gradient
Samples processed for Mono2 ChIP experiment are denoted by red dots. Indicated on right is the input bulk chromatin used on sucrose gradient.

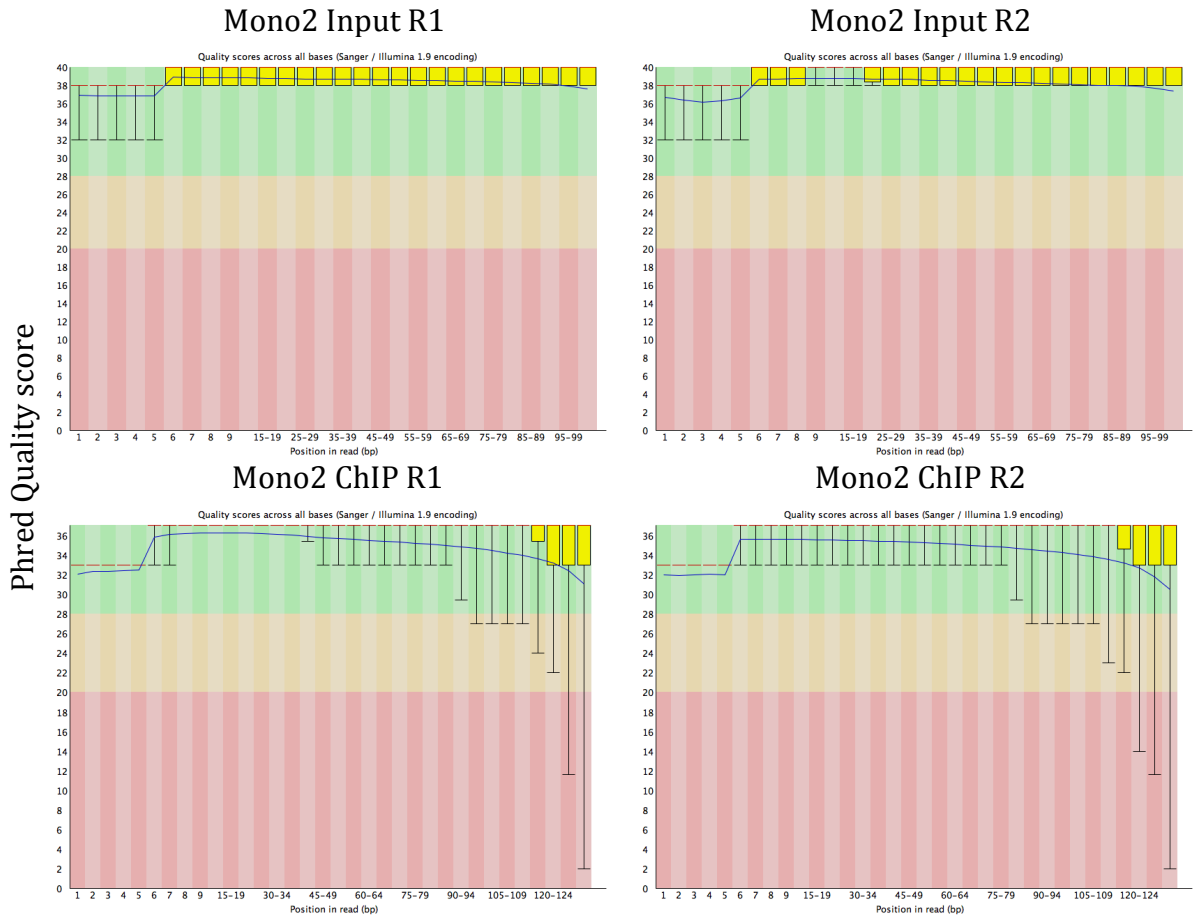


Figure 8.25 Per-base quality scores Mono2

Per base quality metrics for CENP-A ChIP and Input Mono2 libraries. Plots generated using *FastQC*. Phred score is defined in terms of the estimated probability of error. Phred score of 30 is the equivalent to a 1/1000 chance of a called base being incorrect, by the equation q = -10 x $\log_{10}(p)$. The lower quartile (bottom of yellow box) and median (red line) values are above the Phred values 10 and 25 respectively, indicating that the FASTQ reads for both Mono2 CENP-A datasets were of good quality for alignment (Andrews, 2010).

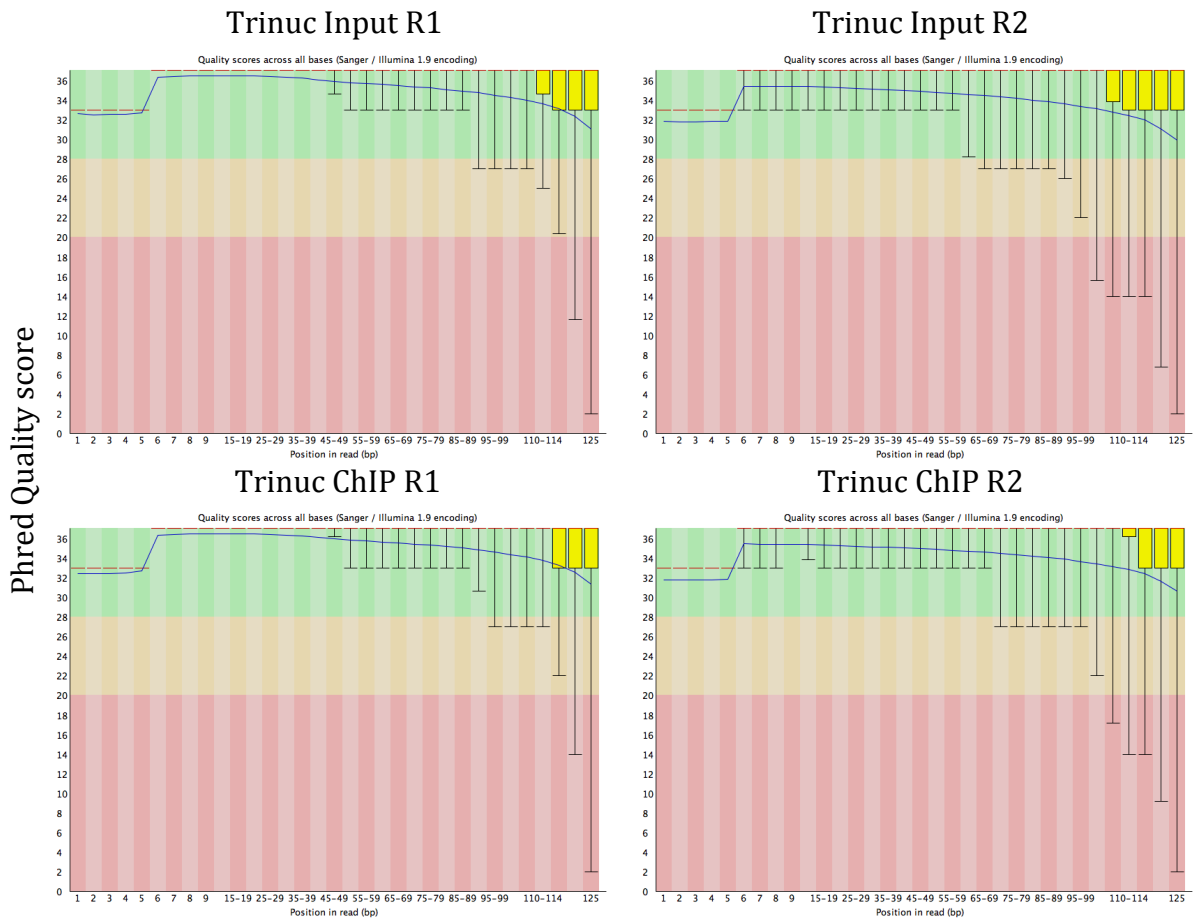


Figure 8.26 Per-base quality scores Trinuc
Per base quality metrics for CENP-A ChIP and Input Trinuc libraries. Plots generated using *FastQC*. Phred score is defined in terms of the estimated probability of error. Phred score of 30 is the equivalent to a 1/1000 chance of a called base being incorrect, by the equation q = -10 x $\log_{10}(p)$. The lower quartile (bottom of yellow box) and median (red line) values are above the Phred values 10 and 25 respectively, indicating that the FASTQ reads for both Trinuc CENP-A datasets were of good quality for alignment (Andrews, 2010).

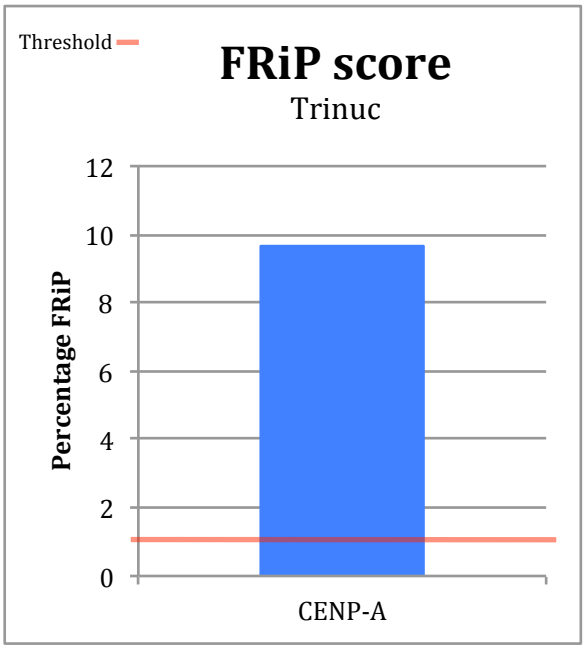


Figure 8.27 FRiP test – Trinuc
FRiP (Fraction of reads in peaks) analysis of the CENP-A Trinuc dataset (post alignment). Trinuc has a value of ~9.8% and therefore above the desired threshold of 1% indicating successful immunoprecipitation

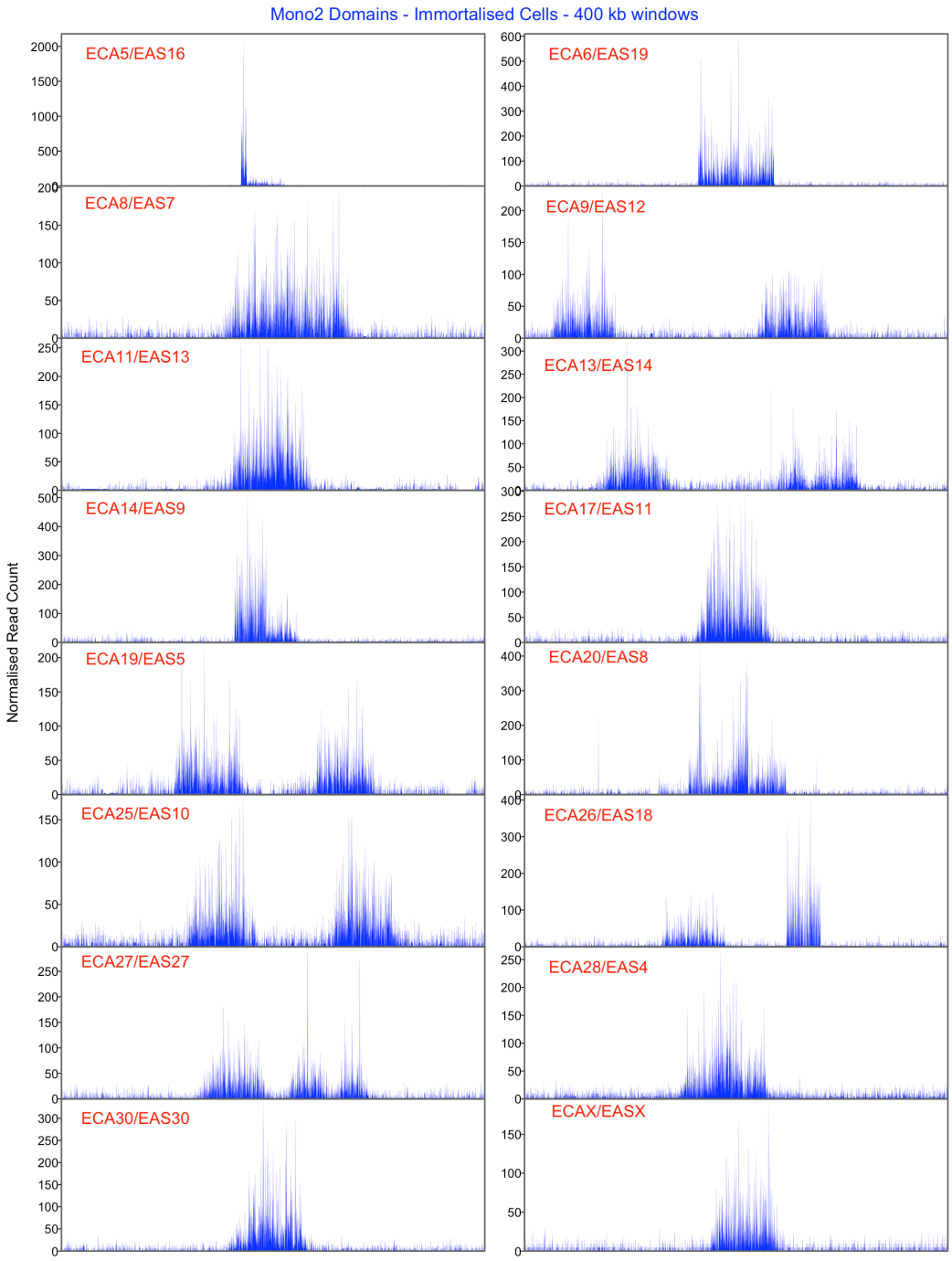


Figure 8.28 CENP-A profiles Mono2
CENP-A distributions from Mono2 dataset shown in Figure 8.28. CENP-A signal is reflective of the lower FRiP score and low recovery in centromere specific DNA. The signal is not as abundant compared to Mono1 or Trinuc.

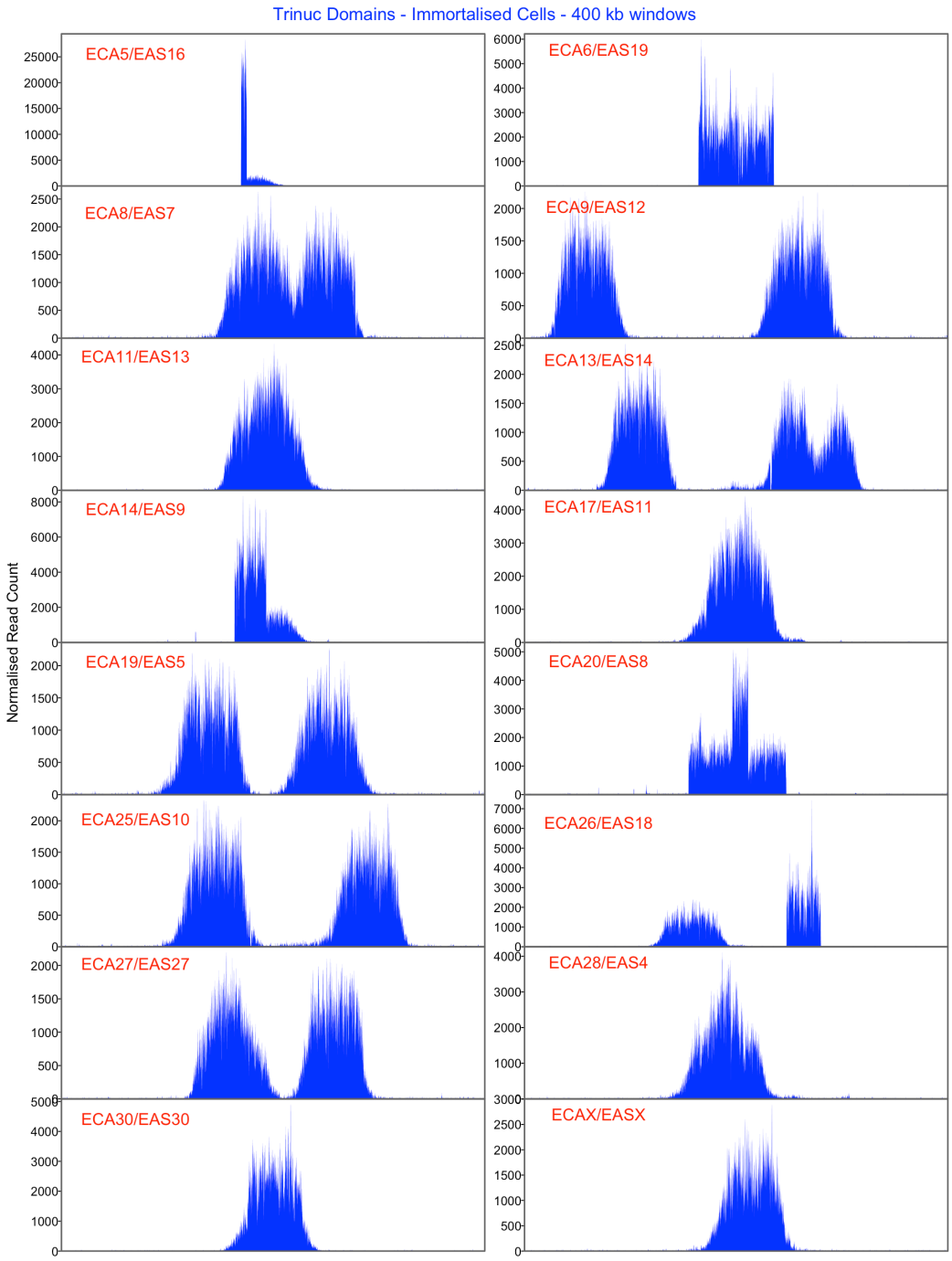


Figure 8.29 CENP-A profiles Trinuc
CENP-A distributions from Trinuc dataset shown in Figure 8.29. CENP-A signal is reflective of the high FRiP score and high recovery in centromere specific DNA.

8.3 Appendix III

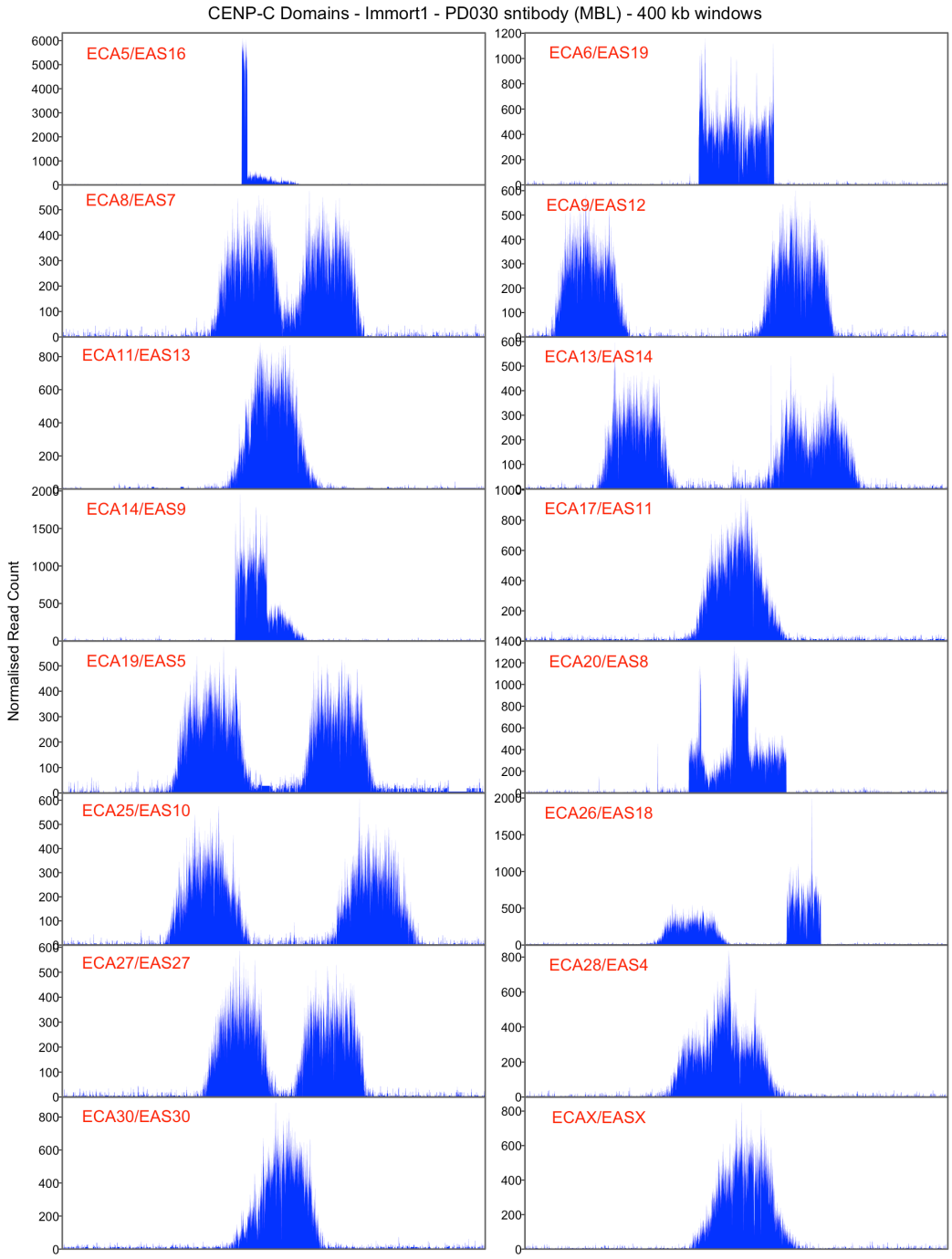


Figure 8.30 CENP-C profiles Immort1
CENP-C distributions from Immort1 dataset display CENP-C across 16 satellite-free centromere domains. Domains are displayed in 400 kb views and strong signal is reflective of high FRiP score for immort1.

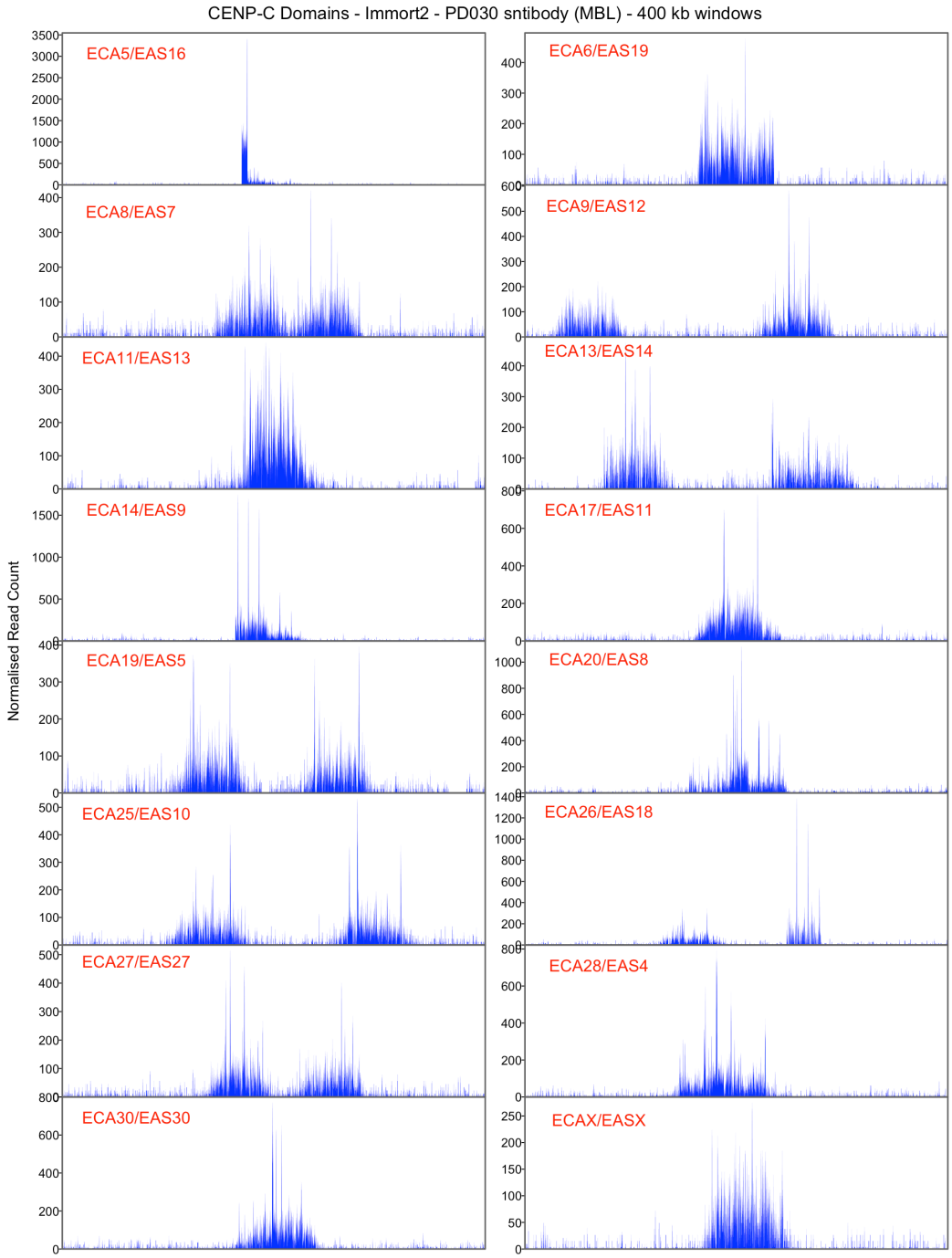


Figure 8.31 CENP-C profiles Immort2
CENP-C distributions from Immort2 dataset display CENP-C across 16 satellite-free centromere domains. Domains are displayed in 400 kb views and the less abundant signal is reflective of low FRiP score for CENP-C Immort2.