ALMA MATER STUDIORUM – UNIVERSITÀ DI BOLOGNA

PHD PROGRAM IN CELLULAR MOLECULAR AND INDUSTRIAL BIOLOGY

Program n. 1: Cell Biology and Physiology

XXXIII Cycle

Scientific Area Code BIO\18

STRUCTURAL AND FUNCTIONAL ANALYSIS OF CENTROMERIC CHROMATIN

PhD Candidate: Monica ZOLI

PhD Program Coordinator

Prof. **Michela RUGOLO**

Supervisor

Prof. **Giuliano DELLA VALLE**

Final Exam 2011

ABSTRACT

Animal neocentromeres are defined as ectopic centromeres that have formed in non-centromeric locations and avoid some of the features, like the DNA satellite sequence, that normally characterize canonical centromeres. Despite this, they are stable functional centromeres inherited through generations. The only existence of neocentromeres provide convincing evidence that centromere specification is determined by epigenetic rather than sequence-specific mechanisms. For all this reasons, we used them as simplified models to investigate the molecular mechanisms that underlay the formation and the maintenance of functional centromeres.

We collected human cell lines carrying neocentromeres in different positions. To investigate the region involved in the process at the DNA sequence level we applied a recent technology that integrates Chromatin Immuno-Precipitation and DNA microarrays (ChIP-on-chip) using rabbit polyclonal antibodies directed against CENP-A or CENP-C human centromeric proteins. These DNA binding-proteins are required for kinetochore function and are exclusively targeted to functional centromeres. Thus, the immunoprecipitation of DNA bound by these proteins allows the isolation of centromeric sequences, including those of the neocentromeres. Neocentromeres arise even in protein-coding genes region. We further analyzed if the increased scaffold attachment sites and the corresponding tighter chromatin of the region involved in the neocentromerization process still were permissive or not to transcription of within encoded genes.

Centromere repositioning is a phenomenon in which a neocentromere arisen without altering the gene order, followed by the inactivation of the canonical centromere, becomes fixed in population. It is a process of chromosome rearrangement fundamental in evolution, at the bases of speciation. The repeat-free region where the neocentromere initially forms, progressively acquires extended arrays of satellite tandem repeats that may contribute to its functional stability. In this view our attention focalized to the repositioned horse ECA11 centromere. ChIP-on-chip analysis was used to define the region involved and SNPs studies, mapping within the region involved into neocentromerization, were carried on. We have been able to describe the structural polymorphism of the chromosome 11 centromeric domain of *Caballus* population. That polymorphism was seen even between homologues chromosome of the same cells. That discovery was the first described ever.

Genomic plasticity had a fundamental role in evolution. Centromeres are not static packaged region of genomes. The key question that fascinates biologists is to understand how that centromere plasticity could be combined to the stability and maintenance of centromeric function. Starting from the epigenetic point of view that underlies centromere formation, we

decided to analyze the RNA content of centromeric chromatin. RNA, as well as secondary chemically modifications that involve both histones and DNA, represents a good candidate to guide somehow the centromere formation and maintenance. Many observations suggest that transcription of centromeric DNA or of other non-coding RNAs could affect centromere formation. To date has been no thorough investigation addressing the identity of the chromatin-associated RNAs (CARs) on a global scale. This prompted us to develop techniques to identify CARs in a genome-wide approach using high-throughput genomic platforms. The future goal of this study will be to focalize the attention on what strictly happens specifically inside centromere chromatin.

TABLE OF CONTENTS

ABSTRACT _____ *1*

TABLE OF CONTENTS _____ *4*

INTRODUCTION _____ *7*

1. THE CENTROMERE _____ **8**

1.1 CENTROMERE FUNCTIONS _____ 10

1.2 HUMAN CENTROMERIC DNA _____ 10

1.3 THE CENTROMERE-KINETOCHORE COMPLEX: THE METAPHASIC CHROMOSOME _____ 12

2. CENTROMERIC PROTEINS _____ **14**

2.1 IS CENP-A THE EPIGENETIC MARKER OF CENTROMERES? _____ 15

2.1.2 CENP-A CONTAINING NUCLEOSOMES _____ 18

2.1.3 THE EPIGENETIC MODIFICATIONS OF CENTROMERIC CHROMATIN _____ 19

2.1.4 CENP-A DEPOSITION: CELL CYCLE TIMING AND REGULATORS _____ 21

2.2 CENP-C _____ 22

3. Non-coding transcripts and Centromeres _____ **24**

3.1 CENTROMERIC TRANSCRIPTS AND EVOLUTION _____ 25

3.2 RETROELEMENTS _____ 27

4. NEOCENTROMERE _____ **28**

4.1 NEOCENTROMERES CLASSIFICATION _____ 30

4.2 NEOCETROMERE: A MOLECULAR VIEW _____ 30

4.3 NEOCENTROMERES AND TUMORS _____ 31

4.4 ECA11: THE EVOLUTIONARY NEOCENTROMERE OF DOMESTIC *HORSE* _____ 32

RESULTS _____ *33*

1. EVOLUTIONARY NEW CENTROMERE _____ **34**

1.1 CHARACTERIZATION OF ECA11 NEOCENTROMERE IN HSF _____ 34

1.2 THE PECULIARITY OF ECA11 CENTROMERE _____ 35

1.4 ANALYSYS OF CENP-A BINDING DOMAIN _____ 37

1.4 SINGLE NUCLEOTIDE POLYMORPHISM ANALYSIS ON HSF *HORSE* CELL LINE _____ 39

1.5 SINGLE NUCLEOTIDE POLYMORPHISM ANALYSIS ON HSF-G AND HSF-D HORSES CELL LINES _____ 41

2. HUMAN NEOCENTROMERES _____ **43**

2.1 HUMAN CHROMOSOME 6 NEOCENTROMERE _____ 43

2.1.1 SEQUENCE ANALYSIS OF NEO6 CENP-A/C DOMAIN: BTN3A2 GENE 45
2.2 HUMAN CHROMOSOME 9 NEOCENTROMERE 46
2.3 PORTNOI AND 2887 NEOCENTROMERES 48
3. CHROMATIN ASSOCIATED RNAs (CARs) 50
3.1 ISOLATION OF TIGHTLY CHROMATIN ASSOCIATED RNAs 51
3.1.1 CARs FROM INTERPHASIC HUMAN CELLS 51
3.1.2 THE MITOTIC CARs 54
3.1.3 CENTROMERIC SPECIFIC CARs 55
3.2 HIGH-THROUGHPUT SEQUENCING OF RNA 58
3.2.1 DATA ANALYSIS 59
DISCUSSION* *69
1. CHARACTERIZATION OF NEW HUMAN NEOCENTROMERES 70
1.2 NEOCENTROMERE DO NOT DEPEND BY THE PRIMARY DNA SEQUENCE 71
1.3 NEOCENTROMERE EVEN ARISE IN PROTEIN-ENCODING REGIONS. THEIR FORMATION DO NOT REPRESS THE GENES WITHIN 72
2. THE CURIOUSE CASE OF HORSE ECA11 CENTROMERE 73
2.1 THE ECA11 NEOCENTROMERE SHIFTING ALONG THE HORSE CHROMOSOME 11 73
2.2 THE FIRST CASE OF STRUCTURAL POLYMORFISM 74
2.3 CENTROMERE REPOSITIONING: A KEY EVENT IN EVOLUTION 76
3. CHROMATIN ASSOCIATED RNA (CARs) IN MITOTIC AND INTERPHASE CELL IDENTIFY MANY INTRONIC AND INTERGENIC TRANSCRIPTS 77
MATERIAL & METHODS* *81
1. CELL CULTURE 82
2. CHROMATIN IMMUNOPRECIPITATION (ChIP) 82
2.1 Native-Chromatin ImmunoPrecipitation (N-ChIP) 83
2.2 Cross-Linked Chromatin ImmunoPrecipitation (X-ChIP) 84
3. PRIMER DESIGN and PCR REACTION 86
3.1 PRIMERS FOR *HORSE* NEOCENTROMERES 87
3.2 PRIMER FOR THE SEQUENCING OF CENTROMERIC SNPs (*SINGLE NUCLEOTIDE POLYMORFISM*) IN *HORSE* 88
3.3 HUMAN NEOCENTROMERES 90
3.4 REAL-TIME PCR 91
4. MICROARRAY 92

4.1 STATISTICAL ANALYSIS OF BINDING PEAKS _____ 92

5. CARs ISOLATION _____ 93

5.1 CARs extraction from interphase cells _____ 93

5.2 CARs PURIFICATION _____ 96

5.3 CARs from Centromeric chromatin _____ 96

5.3.1 Protocol 1 _____ 96

5.3.2 Protocol 2 _____ 98

5.4 CARs FROM MITOTIC CHROMOSOME _____ 99

5.5 SOLEXA TECHNOLOGY _____ 100

5.6 TABLES _____ 101

5.6.1 INTHERPHASIC CARs BELONGING TO snoRNA _____ 101

5.6.2 MITOTIC CARs BELONGING TO snoRNA _____ 104

BIBLIOGRAPHY _____ 105

INTRODUCTION

1. THE CENTROMERE

"Each cell can only arise from the division of a cell before" the concept expressed by Wilson in 1925, reveals the crucial role of the two divisional processes: meiosis and mitosis. To exercise this extraordinary ability to divide the cells have numerous mechanisms to achieve the high degree of accuracy required. The way how the centromere is able to regulate cell division remains one of the greatest enigmas of the genome.

The Centromere, consisting of DNA and proteins, is a highly differentiated structure of the chromosome. In higher eukaryotes it can be divided into three domains: the pairing domain , the central domain and the kinetochore. The centromere-kinetochore complex plays essential functions in different aspects of mitosis and meiosis: it is responsible of the sister chromatids pairing, is the site of attachment to the microtubules of the mitotic spindle, it controls the cell cycle transition metaphase-anaphase and regulates the movement of chromosomes along the poles of the dividing cells. DNA sequences that control the functions of the centromere are very different even between closely related species. Unlike organisms such as *C. elegans* that contain olocentromeres , a particular type of centromere which assembles the kinetochore along the entire length of the chromosome, the centromeres of higher eukaryotes can be seen as a primary constriction on metaphase chromosomes. The DNA present in these regions typically consist of large blocks of repetitive sequences known as satellite sequences and has been very difficult to study the importance of specific classes of sequences in the determination of centromere function because of the size and complexity of these regions. In contrast, proteins that make up the centromere-kinetochore complex appear to be highly conserved even between phylogenetically distant organisms. This inconsistency between the conservation and variability of protein sequences that make up the centromere is at the base of the central paradox of the biology of the centromere, which seeks to clarify how DNA regions are subject to rapid sequence divergence can contact a group of kinetochore proteins conserved to ensure a critical and fundamental cellular function such as cell division and thus the inheritance of genetic information (Henikoff et al., 2001). Consequently, the analysis of protein and sequences that create an active centromere and the interactions that mediate and regulate the formation of a functional complex are still under investigation.

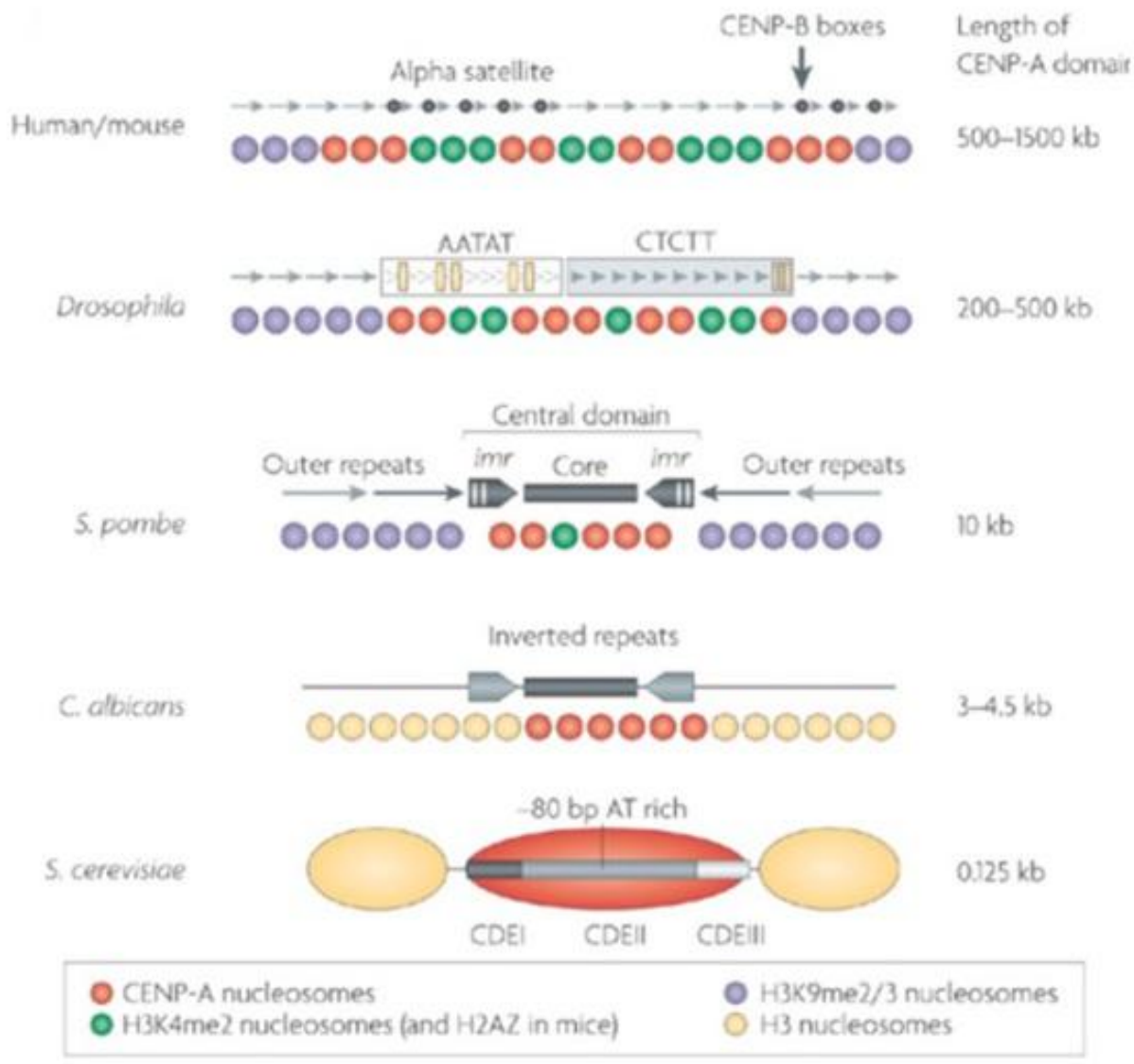


Figure 1.1 Example of different types of centromeres in different organisms (Allshire and karpen 2009)

There are three main types of centromeres structure that differ for complexity (fig 1.1):

1. POINT CENTROMERES: only present in S. cerevisiae, defined by a chromosomal region known as "CEN" of about 125bp organized in three conserved regions: CDEI, and CDEII CDEIII (Pluta et al., 1995).

2. REGIONAL CENTROMERES: they are found in higher eukaryotes, with various levels of complexity, and they are characterized by the presence of repeated sequences that occupy vast regions of chromosome: 40-120Kb in *S. pombe* with a central core surrounded by long tandem repeats; 250Kb in Drosophila with small repeats of only 5,7 or 10bp, interspersed

with transposable elements and up to several megabases in humans (Clarke and Carbon, 1985) (see §1.2).

3. OLOCENTROMERES: centromeres delocalized along the entire chromosome. They are typical of some invertebrates such as *C. elegans* (Felsenstein and Emmons, 1987) and some plants.

1.1 CENTROMERE FUNCTIONS

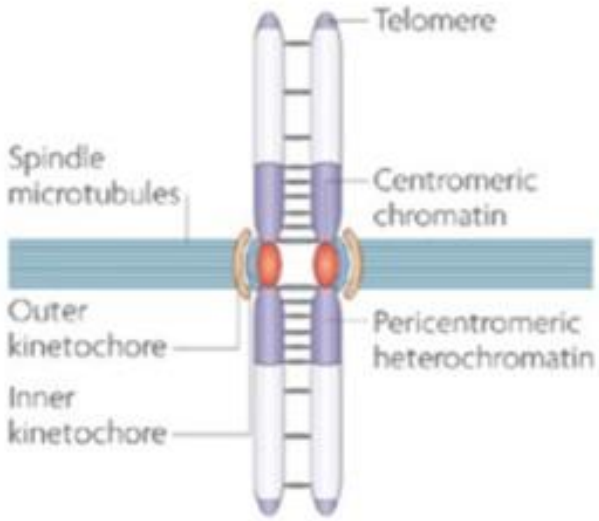


Fig. 1.2 Mitotic chromosome in cell division.

The centromere is the structure responsible for chromosome movement in eukaryotes. Its function is based in 3 process:

- Sister chromatid cohesion which establishes chromosome polarity at mitosis and meiosis (Dej and Orr-Weaver, 2000).
- Check-point of anaphase starts when all chromosomes are properly joined to microtubules (Rieder and Salmon, 1994).
- Kinetochore assembly and microtubule anchorage in centromeric heterochromatin (Rieder and Salmon, 1998).

1.2 HUMAN CENTROMERIC DNA

All primate chromosomes studied to date contain α-satellite DNA. α-satellite was originally characterized in the African Green Monkey genome as a satellite DNA family based on divergent 170-bp monomers arranged in a tandem, head-to-tail fashion resulting in an overall directionality. This type of α-satellite is termed monomeric and has been identified only within the pericentromeric regions of 21 of the 24 human chromosomes (Rudd and Willard, 2004). Each human chromosome is also characterized by a chromosome-specific higher-order array of α-satellite (Fig 1.3). Each higher-order array is composed of a tandemly arranged repeat unit consisting of an integral number of α-satellite monomers. Higher-order arrays can

span over 3–5 Mb and are highly homogenous, consisting of the same higher-order repeat unit occurring hundreds or thousands of times within a given centromere locus (Willard, 1989).
The stretch of monomeric α- satellite is frequently interrupted by interspersed elements [long interspersed element (LINE), short interspersed element (SINE), long terminal repeat (LTR) retrotransposons (Figure 1.3) and has disrupted directionality such that blocks of monomers of common directionality are defined by changes in orientation relative to adjacent blocks.

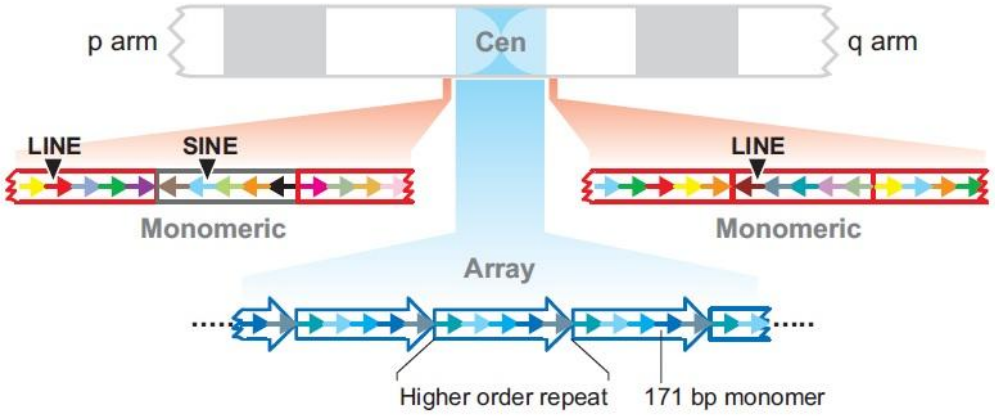


Figura 1.3 Centromeric (CEN) DNA organization. A typical human chromosome is schematically depicted, emphasizing the pericentromeric and CEN (*blue*) satellites in the ideogram. Each small arrow represents a single satellite monomer. In the pericentromeric regions, blocks of tandem satellite monomers from a single family (indicated by *red* versus *gray boxes* occasionally contain embedded interspersed repetitive elements [e.g., long interspersed elements (LINEs) and short interspersed elements (SINEs)].Adjacent satellite blocks can exist in the same or opposite orientations. In the CEN region, higher-order repeat units of α-satellite (this unit comprised of five monomers) are indicated with large blue arrows (Schueler and Sullivan, 2006)

The highly divergence in sequence make difficult his study. To date CENP-B is the only centromeric protein that bind in a sequence-specific manner a 17 bp long sequence within α-satellite known as "CENP-B Box". It is supposed to bind extensively along a-satellite (up to 4Mbs).
Centromere regions contain distinct epigenetic marks (see § 2.1.3) including dense DNA hypermethylation, normally associated to a transcriptional repression.
For years the alphoid DNA was considered of primary importance for the formation of a functional centromere. Human artificial chromosome stable in mitosis were create only if alpha satellite and telomeric sequence were present (Harrington et al., 1997; Saffery et al., 2001). Nowadays many are the evidences that sustain that alpha-satellite is neither necessary nor sufficient in order to have functional centromere: the existence of neocentromere and the functional silencing of canonical centromere in dicentric chromosome support the epigenetic nature of centromerization.

1.3 THE CENTROMERE-KINETOCHORE COMPLEX: THE METAPHASIC CHROMOSOME

Centromere is defined by a primary constriction visible in the metaphase chromosome and consist of more than 90 proteins that working together provide the site of attachment of spindle microtubules for the proper segregation of chromosomes.
The centromere-kinetochore complex consists of three main domains: pairing domain, central domain and the kinetochore (Fig 1.4).

- **PAIRING DOMAIN**

It consists of DNA and proteins, it promote the link between the two sister chromatids that occur before the bipolar attachment of chromosomes to the spindle. In this domain there are the INCENP protein (INner CENtromere Proteins) and the CLIP protein (Chromatid LInking Proteins) able to localize, during metaphase, the contact between the centromeres of sister chromatids.

- **CENTRAL DOMAIN**

It's mainly characterized by highly condensed constitutive heterochromatin which is anchored to the kinetochore. The constitutive heterochromatin is composed of highly repetitive satellite DNA and proteins associated with it. In primates, the most abundant family of repetitive DNA is DNA α-satellite (§ 1.2) which is distributed throughout the central domain (Pluta et al., 1990). The centromeric protein CENP-B is bound to a specific sequence of 17 bp (CENP-B box) inside the α-satellite and localizes in this domain. Two other proteins (HMG-I and PARP or pJα), which have been implicated in binding to 'α-satellite, are within the central domain.

- **KINETOCHORE**

Electron-microscopy studies defined the kinetochore as a electrondense trilaminar disk composed of two layers, the inner plate and outer plate, separated by a layer transparent to electrons (intermediate zone). This macromolecular structure described is only visible from late prophase until the end of mitosis, suggesting that the complex is assembled and disassembled periodically during each mitotic process.

The outer plate is associated with a fibrous structure. It consist mainly of protein, CENP-E and CENP-F are examples of protein that localize in the outer plate and both are directly involved in chromosome movement because of their interaction with the microtubules of the mitotic spindle. The inner plate is composed of DNA and proteins, among these CENP-A, CENP-C, CENP-H, CENP-I, CENP-G, Mis12.

To date more than 16 non histonic protein are known to localize in the centromere-kinetochore complex, but their function is still object of studies.

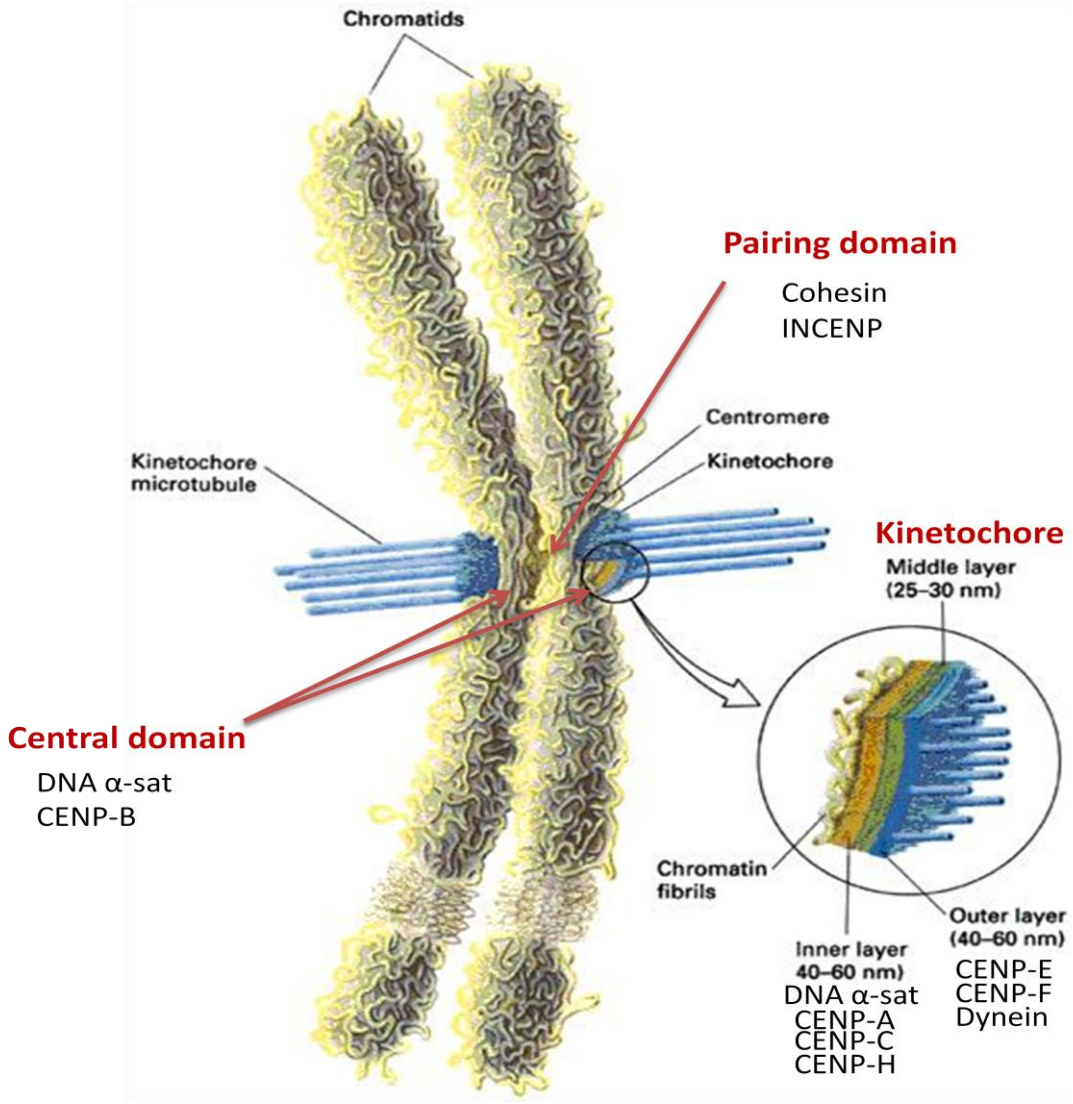


Figure 1.4 Centromere-kinetochore complex of the eucariotic chromosome in mitoses

2. CENTROMERIC PROTEINS

The proteins that constituent the kinetochore are of particular interest because they define its functions and are involved in the determination and propagation of epigenetic centromere (Amor et al., 2004). They are localized to the centromere throughout the cell cycle and are associated, for the most part, in the inner kinetochore plate. They are assembled in hierarchical but interdependent succession that still under definition. Homologues of these proteins have been identified in all higher eukaryotes: their evolutionary conservation, even at the level of primary sequence, clarify their importance.

Recent biochemical purification studies in human and chicken cells have identified two major centromere-associated complexes (Foltz et al., 2006): CENP-B, -C, -H, -M, -N, -T and -U were found to be associated with one or a few CENP-A nucleosomes, forming the nucleosome-associated complex (NAC). All these proteins except for CENP-B, are essential and exclusive to all active centromeres including neocentromeries. Other proteins isolated in these studies did not associate directly with CENP-A, and instead formed a more distal complex called CAD (CENP-A distal). NAC proteins are needed to recruit CAD components, which in turn are required for assembling subsets of outer kinetochore proteins during mitosis (Fig 2.1).

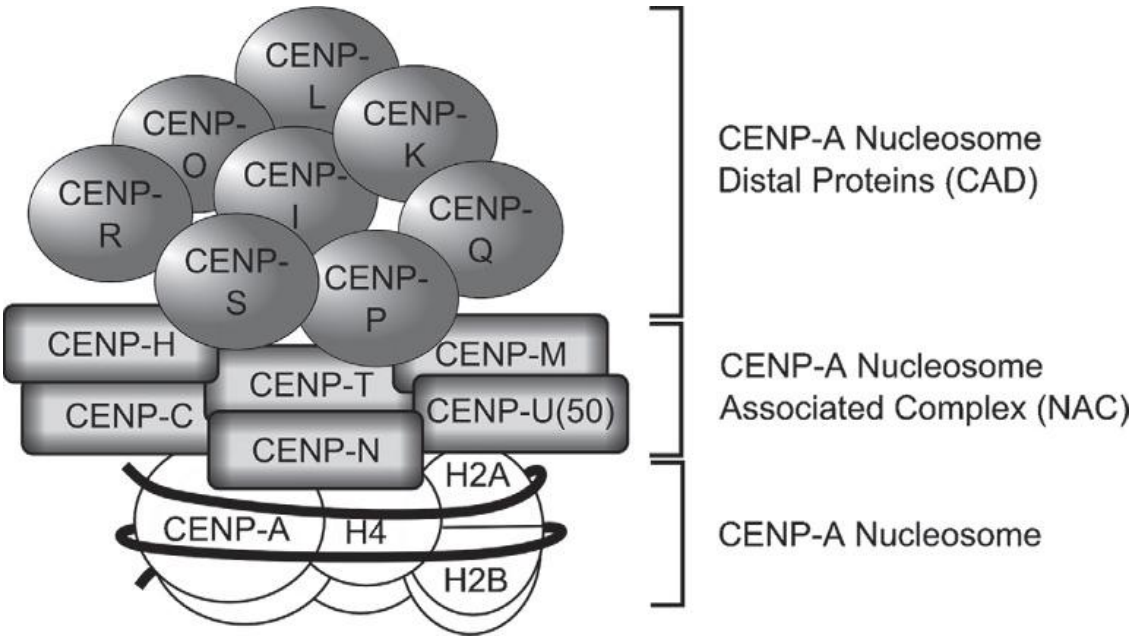


Fig. 2.1 The CENP-A nucleosome associated complexes. CENP-A nucleosomes co-purify with members of the CENP-A NAC that are constitutively found at centromeres (Foltz et al., 2006; Okada et al., 2006) . A more distal complex, CENP-A CAD , contains several additional constitutive centromere components (Panchenko and Black, 2009)

2.1 IS CENP-A THE EPIGENETIC MARKER OF CENTROMERES?

CENP-A is the centromere-specific histone H3 variant and is the main component of the assembly of a functional centromere, it is a constitutive and essential protein of 140amino acids. It's molecular weight is 17 kDa. Identified by the use of serum from patients with autoimmune CREST, it is found only in functional centromeres. The study of the localization of CENP-A kinetochore was made possible by the use of specific antibodies directed against the CENP-A N-terminal region of the human protein (amino acids 17-24) which shows no homology with histone H3 (Trazzi et al., 2009; Warburton et al., 1997). Studies with knockout mice have shown that CENP-A protein is essential for life (Howman et al., 2000). While mice heterozygous for CENP-A + / - are healthy and fertile, the homozygous CENP-A -/ - do not survive beyond 6.5 days after conception (dpc) and show severe mitotic problems including formation of micronucleus and macronucleus, connections between the nuclei and nuclear swelling, ipercondensation and fragmentation of chromatin. Interphase cells of 5.5 days embryos shows a total lack of CENP-A, while CENP-B and CENP-C are missing in the nucleus. Because of its structural role in the nucleosome is thought that the association of CENP-A to the kinetochore is one of the first events that characterize the formation of the kinetochore during interphase (Howman et al., 2000).

CENP-A localizes to the kinetochore inner plate and overlap with of CENP-C. Both CENP-C (Sugimoto et al., 1999) than CENP-A (Sugimoto et al., 2000) still remain associated with the centromere throughout the cell cycle. Immunoprecipitation experiments have shown that CENP-A is associated with the α-satellite DNA (Vafa and Sullivan, 1997) but they could not identify a consensus sequence binding. The fact that CENP-A is present in all analyzed neocentromeres but not in the inactive centromeres of dicentric chromosomes, was the first suggestion for the epigenetic nature of its association with the centromere (Vafa and Sullivan, 1997; Warburton et al., 1997).

Histones are highly conserved and contains sites for a plethora of post-translational modifications, particularly in their divergent N-terminal tails. These modifications are correlated with different functional states, such as transcriptional activity or silencing, and are involved in chromatin assembly and disassembly. Surprisingly little is known about this post-traslational modifications of CENP-A. The only modification described for CENP-A is the

phosphorylation of serine 7 in human cells by Aurora B that has been implicated in promoting the proper localization of Aurora B (Zeitlin et al., 2001). The fact that the N-terminal domain seems to be dispensable for CENP-A targeting argues against the possibility of the involvement of modification of the tail for proper localization and function. CENP-A is present in all eukaryotes examined to date and its depletion, as previously introduced, involves the mis localization of most kinetochore proteins including CENP-C (Blower and Karpen, 2001; Howman et al., 2000; Meluh et al., 1998) whereas depletion of most kinetochore proteins has no effect on CENP-A localization (Blower and Karpen, 2001). The overexpression of the protein leads to its incorporation in ectopic sites on chromosome arms and the relocation of CENP-C, but not the formation of a functional centromere (Van Hooser et al., 2001) so the presence of CENP-A alone is not sufficient to promote the assembly of the kinetochore. These observation suggest that CENP-A is both a structural and functional foundation for the kinetochore, and that it lies at or close to the apex of the pathway that is responsible for kinetocore formation. However recent studies have identified proteins that are co-dependent with CENP-A for centromere localization, suggestiong that CENP-A is not alone at the top of the hierarchy.

Gene silencing studies have supported this ipotesis showing that CENP-A have a central role in directing the hierarchy association of kinetochore through three distinct, although interrelated, pathways of assembly primarily dependent on CENP-C, CENP-I and Aurora B (Liu et al., 2006) . A novel genomic-wide Drosophila screening for genes that regulate the centromeric localization of CENP-A/CID protein reveal an interdependency between CID, CENP-C and CAL1 for centromere propagation (Erhardt et al., 2008). Centromere localizations of CENP-A, CENP-C, and Cal1 in drosophila were mutually dependent, because RNAi depletion of any single protein disrupted or diminished the localization of the other two. Inner kinetochore formation may involve a co assembly process of CENP-A, -C, and the protein Cal1 discovered in this screening (Goshima et al., 2007). Levels of CAL1 are reduced on metaphase centromeres and increase with CENP-A loading in late anaphase to telophase (§2.1.4). No homologs for the CAL1 have been described in other organisms, but in vertebrate cells is shown a similar interdependence between CENP-H- CENP-I complex and CENP-A (Okada et al., 2006).

Sequencing and structural studies of CENP-A showed a C-terminal region with the 57% identity with the H3 Histone FoldDomain (HFD) and a N-terminal region, not conserved, varying in length from a minimum of 27 to a maximum of 196 amino acids (Fig. 2.2; (Earnshaw et al., 1986; Palmer et al., 1991).

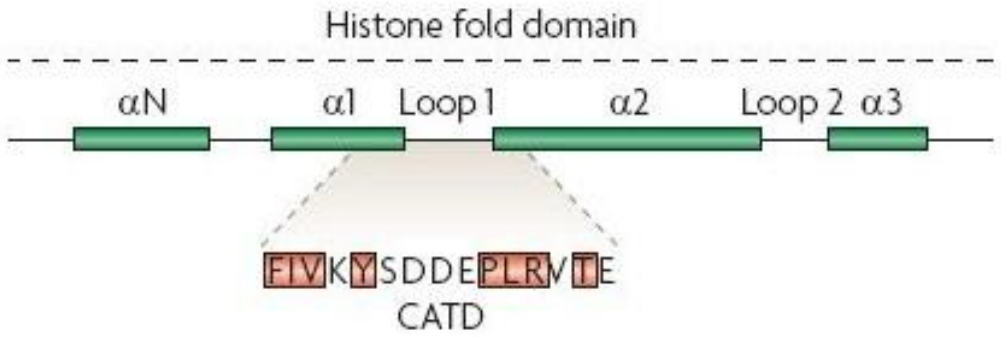


Fig. 2.2 Generic scheme of the CENP-A protein

In humans, the protein consists of a N-terminal domain of 40amino acids and a C-terminal domain of 100 amino acids with a 59% identity with histone H3.The latter domain is necessary for protein localization to the centromere (Sullivan et al., 1994). The secondary structure of the C-terminal region display common features with the same histone H3 protein (Shelby et al., 1997) both contain three α-helices (I, II and III) and two regions β-sheet (Loop1 and Loop2). Interchanging Loop1 or helix II (regions that show greater divergence within the C-terminal domain) between CENP-A and histone H3, result in a loss of the ability of CENP-A to specifically localize to centromere. Moreover crystallography studies show that histone H3 contacts the DNA by this two specific features (Richmond et al., 1993; Shelby et al., 1997). The HelixII mediates histone H3 interaction with H4 to form the H3/H4 tetramer in the nucleosome core, while in CENP-A is required for the specific localization to the centromere in immunofluorescence assay and is also responsible for the homodimers formation of (Shelby et al., 1997). This histone-histone interaction is essential for the function of CENP-A. The demonstration that CENP-A can replace histone H3 in a nucleosome reconstructed in vitro was of crucial importance to address how CENP-A contributes to the formation of a nucleosomal structure containing 120-150bp, α-satellite and equimolar amounts of the H2A, H2B and H4 histones (Yoda et al., 2000).

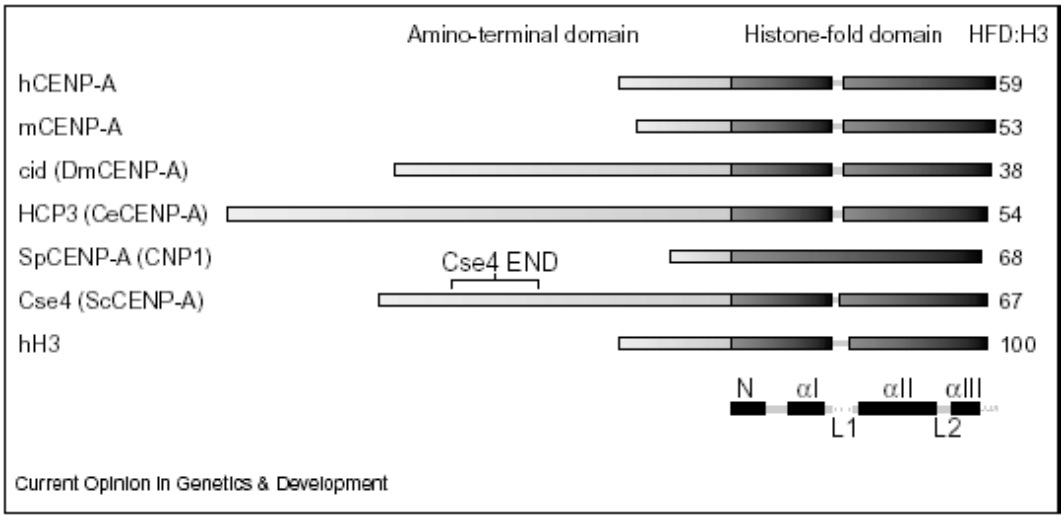


Fig 2.3 Diagrams of the CENP-A structures in different eucariotic organisms. The sequences are: hCENP-A, human; mCENP-A, mice; cid, *Drosophila*; HCP3, *C. elegans*; SpCENP-A, *S. pombe*; Cse4, *S. cerevisiae*; hH3, human histone H3 (Sullivan et al., 2001)

2.1.2 CENP-A CONTAINING NUCLEOSOMES

CENP-A nucleosome are structurally distinct from H3 nucleosomes: the conformation of the interface with H4 is more compact. The first loop (L1) and second alpha helix(2) within the HFD of the CENP-A protein are the responsible of that tightly structure. This region is known as the CENP-A targeting domain (CATD) and if introduced in the histone H3 is sufficient to direct it to the centromeres and this chimeric histone can rescue the viability of CENP-A depleted cells (Black et al., 2004; Black et al., 2007). Very recently, CENP-A nucleosomes assembled in vitro have been shown to wrap DNA in a right-handed manner, opposite to the left-handed wrapping of H3 nucleosomes (Furuyama and Henikoff, 2009). Thus, the physical properties of [CENP-A-H4]2 and [H3-H4]2 tetramers are distinct and might contribute to functional differences and thereby propagate the epigenetic mark from one generation to the next.

The composition of the CENP-A nucleosomes is still unclear. Despite some reports of unusual CENP-A nucleosomes in fly, the affinity purification of CENP-A nucleosomes from human and fly cells suggest that they mainly exist as homotypic octamers that contain [CENP-A-H4-H2A-H2B]2; anyway more detailed information about the components and properties of CENP-A nucleosomes, and whether they change during cell cycle are needed.

2.1.3 THE EPIGENETIC MODIFICATIONS OF CENTROMERIC CHROMATIN

Immunocytochemistry analysis on extended chromatin fibers interestingly showed the simultaneous presence of CENP-A and histone H3 (Fig. 2.4a). This founding stimulating the hypothesis of an alternating arrangement of classic nucleosomes (with histone H3, H2B, H2A, H4) with nucleosomes containing CENP-A. This linearly interspersed pattern was found in humans, in *Drosophila* (Blower et al., 2002) in rice (Yan and Jiang, 2007) and in one human neocentromere (Alonso et al., 2010). The interspersed H3 domains within centromeres are modified in an epigenetic pattern that is distinct from both euchromatin and eterochromatin. These domains contain the H3K4me2 modification which is normally associated with open (active) chromatin, but they lack the acetylated residues of the tail that also usually mark open chromatin. Even in *S.pombe*, centromeres seem to adhere to this arrangement. The central kinetochore domain, composed of inner repeats and a central core, is packaged in approximately 10kb of chromatin that is composed mainly of the CENP-A homologue (Cnp1) and that seems to contain some histone H3K4me2 (Castillo et al., 2007).
A very recent analysis of the centromere chromatin of the alphoid HAC in human cells reveal the presence of the hypermethylated H3K36 (Bergmann et al., 2010) that is consistent with the observed transcription of centromere type I α-satellite repeats. Recent studies in yeast highlight an important role for H3K36 methylation in the maintenance of chromatin architecture: co-transcriptional methylation of H3K36 is linked to the recruitment of a HDAC containing complex that maintains a hypoacetylated state, antagonizes H3K4 trimethylation and suppresses spurious intragenic transcription (Lee and Shilatifard, 2007).
The function of this particular post-traslational modifications is not known but this could be a particular features of centromere and could serve like a marker for the centromeric recruitment of CENP-A and/or participate in assembling the cylindrical three-dimensional structure of centromeric chromatin in mitotic chromosome. Nucleosome containing CENP-A appears to primarily associate with α-satellite type I found in the outer centromere complex (Ando et al., 2002) while the α-satellite type II, which is located inside make contact with purely structural proteins such as INCENPs, the Aurora B kinase and cohesion. The entire region seems to be folded in mitotic chromosomes so that all CENP-A is on the surface of the chromosome, with H3 residing underneath (Fig. 2.4). This arrangement could ensure that CENP-A is exposed on opposite sides of cohesed sister centromeres, which would promote

the formation of sister kinetochores that interact with microtubules from opposite poles promoting the bi-orientation (Fig 2.4 b and c).

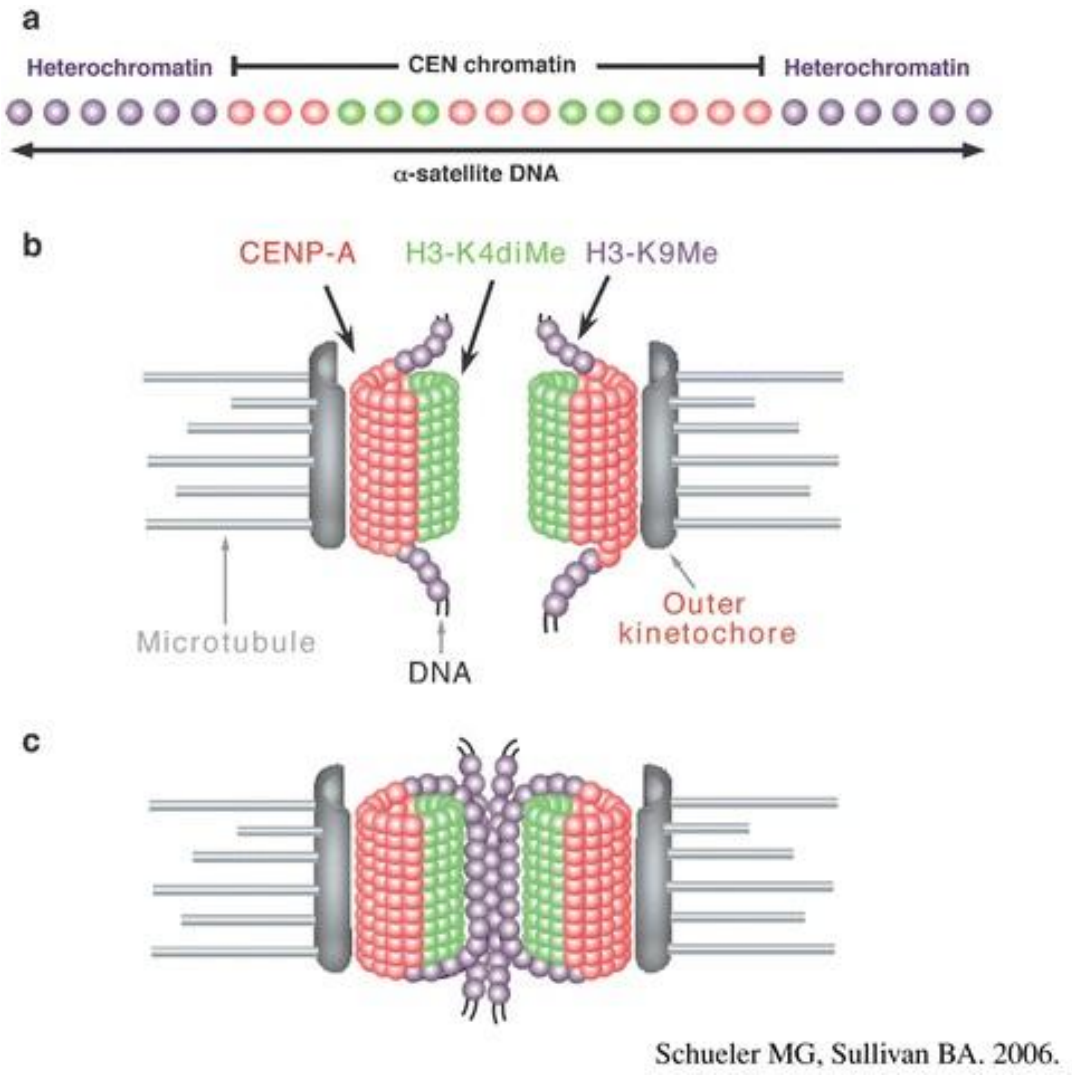


Fig. 2.4 Unique organization of centromere regions in humans. (a) On linear, two-dimensional chromatin fibers, with histone CENP-A (red) interspersed with H3K4me2 (green) to form a domain of CEN chromatin on human α-satellite DNA. The heterochromatin (purple) fanks one or both sides of CEN chromatin domain. (b) At metaphase, when mitotic chromosomes condense, the interspersed domains promote coiling of the DNA so that stacks of CENP-A nucleosomes are presented to the poleward face of the chromosome where they can interact with other kinetochore proteins. H3-containing nucleosomes are oriented between sister kinetochores. (c) Heterochromatin defined by nucleosomes containing H3-K9 methylation (purple) is assembled into a domain that is distinct from CEN chromatin.

.

2.1.4 CENP-A DEPOSITION: CELL CYCLE TIMING AND REGULATORS

Maintenance of centromere identity requires incorporation of new CENP-A during or after replication of centromeric DNA. In *S. cerevisiae* all pre-existing CENP-A is replaced by newly synthesized CENP-A during S phase (Pearson et al., 2004). Early studies in human cells showed that deposition of CENP-A is uncoupled from DNA replication: while centromere duplication takes place in mid-to late S phase, CENP-A levels are low at this time (Shelby et al., 2000). The incorporation of newly synthesized CENP-A occurs in fact in telophase/early G1 (Jansen et al., 2007; Schuh et al., 2007). Nowadays what specifically bring the CENP-A to the centromere is still unknown. The incorporation of histones into chromatin involves numerous events. Each of these steps (histone genes transcription and translation, proteins modifications, import into the nucleus and nucleosome assembly) provides an opportunity for regulation of CENP-A localization to centromeres by contributing factors. Factors required for centromere localization of CENP-A have been identified by genetic screens and subsequent analyses. Progress has been made in identifying *trans*-acting proteins that are required for CENPA localization in eukaryotes. Despite their sequence and structural similarities, histone H3 variants are deposited by distinct chaperones. In humans, several purification approaches led to the identification of constitutive centromeric components (CENPs) and associated factors whose presence at centromeres depends on CENP-A. Very recently, throught affinity purification and mass-spectrometry analysis, two independent groups identified HJURP as a partner of pre-deposited CENP-A.

HJURP is required to promote deposition of human CENP-A in human cells, and recognizes the CATD of CENPA. Furthermore, it localizes to centromeres at the time of CENP-A loading (Dunleavy et al., 2009; Foltz et al., 2009). HJURP cannot be detected on mitotic chromosome but bind centromere in G1. All these evidences make HJURP the likely CENP-A specific chaperone although its molecular mechanism of action remains to be elucidated. Interestingly, HJURP had been previously implicated in double strand break repair and shown to bind Holliday junction-like DNA in vitro (Kato et al., 2007) and base excision repair proteins have been proposed to contribute to CENP-A loading in *Xenopus* extracts (Zeitlin et al., 2005). A speculation could be that mitotic recombination of centromeric sequences, a process known to occur both in yeast and mammalian cells, could be a source of unresolved junctions recognized by HJURP upon exit from mitosis (Jaco et al., 2008). Alternatively, DNA breaks could arise from the late resolution of catenanes in anaphase, as centromeric

DNA is severely stretched, and be repaired in a process coupled to CENP-A loading with the participation of HJURP (Wang et al., 2008).

Moreover a determinant for the specific loading of CENP-A at centromere loci could be the pattern of specific histone modification that promotes a chromatin status permissive for CENPA deposition and maintenance or, alternatively, by prevent CENP-A eviction. A good candidate could be the *S. pombe* Mis18. It has been shown to be necessary for CENP-A deposition and it helps to maintain the hypoacetylated state of histones in the centromere central domain (Hayashi et al., 2004). In human cells, Mis18 is targeted to centromeres upon exit from mitosis and precedes HJURP, but a dependency of HJURP binding on Mis18 function has not been demonstrated (Foltz et al., 2009; Fujita et al., 2007). Components of the human Mis18 complex, consisting of Mis18A, Mis18B and Mis18-binding protein, are particularly interesting factors required for centromere formation.

2.2 CENP-C

CENP-C is an essential kinetochore protein identified for the first time by using the autoimmune serum (CREST) of patients with scleroderma (Saitoh et al., 1992). It's a marker of functional centromeres because of its association with active centromere on dicentric chromosomes (Page et al., 1995) and neocentromeres (Choo, 1997). It localize in the inner kinetochore and is fundamental for both its structure and functionality. CENP-C act roles in the segregation of chromosome, in the correct assembly of functional kinetochore and in the transition metaphase/anaphase.

From siRNA experiment CENP-C is necessary for the structural organization of the kinetochore and should direct the binding of proteins that make direct contact with the mitotic spindle (Liu et al., 2006). Gene silencing of CENP-C result in a significant destruction of the structure of the typical trilaminar kinetochore. The pathway governed by CENP-C, and CENP-A , CENP-H and CENP-I dependent, seems to drive the compaction and the size of the kinetochore plate and its loss is associated with the formation of kinetochore dimensionally smaller, probably because it made by a small number of modular units which a consequent outer plate reduce in size (Liu et al., 2006). Completely kinetochore destruction has been detected after CENP-C si-RNA knock down in Drosophyla (Orr and Sunkel, 2010) where, differently from all other organisms, CENP-A (Cid in Drosophyla) deposition to centromere is CENP-C depend.

On the other hand, CENP-C seems to be necessary but not sufficient to induce the formation of a functional centromere, because, while the lack of CENP-C causes a temporary block of the transition metaphase/anaphase, its ovrespression is associated with errors in chromosome segregation and consequently blocking of cells in mitosis (Fukagawa et al., 1999).

Furthermore siRNA knockdown of CENP-C in mammalian cells led to a significant loss of DNA methylation, marked changes in the histone code (higher H3K9 dimetilation at the alpha satellite was detected) and reduced DNMT3B (DNA methyltransferase) binding at centromeric and pericentromeric regions, elevated mitotic chromosome instability and enhanced centromeric transcription (Gopalakrishnan et al., 2009). The suggestion is that this interaction could have a role in the epigenetic marks of centromeres.

Despite its importance it is not yet clear how CENP-C localize to centromere and perform its function. Comparative analysis of the CENP-C homologues isolated from other species shows that the central and C-terminal region of the human CENP-C display different degrees of conservation Particularly, while the central domain is poorly conserved, the C-terminal domain contains two regions that are highly conserved from yeast to mammals, suggesting that such regions might be preserved during evolution to exert critical centromere functions (Talbert et al., 2004). Based on their degree of conservation with the yeast Mif2 protein, the two regions have been named Mif2p homology domain II and III (Brown, 1995). The Mif2p homology domain II (aa 737/759 of human CENP-C), also called CENP-C motif (Talbert et al., 2004), is present in all CENP-C homologues, though its specific function has not been defined.

The creation of mutant of the protein has allowed the analysis of individual portions of CENP-C. ChIP assay to functional mutant of CENP-C have shown that CENP-C is able to bind the alpha satellite DNA of human centromere in two different domain: the central domain (aa 410-537) (Politi et al., 2002) and in the C-terminal domain (aa 638-943) (Trazzi et al., 2002). The same domain in immunochitochemistry assay have been shown to specifically localize to centromere (Trazzi et al., 2002) and in co-immunoprecipitation assay to direct bind the DNMT3B (Gopalakrishnan et al., 2009). Crosslinking analysis have shown that the C-terminal domain is required for CENP-C dimerization and/or oligomerization (Trazzi et al., 2009). Moreover using both coimmunoprecipitation and bimolecular fluorescence complementation assays, Trazzi et al showed in 2009 that the C-terminal region of CENP-C, containing the evolutionarily conserved Mif2p homology domains III, is required for the interaction with CENP-A at the centromere position and the same domain of CENP-C directly interact with the histon H3.

SiRNA knockdown of CENP-C led to a significant loss of DNA methylation, marked changes in the histone code and reduced DNMT3B binding at centromeric and pericentromeric regions.

3. NON-CODING TRANSCRIPTS AND CENTROMERES

Centromeres have long been thought to comprise noncoding and transcriptionally inactive DNA. However, recent evidence suggests that eukaryotic centromeres produce a variety of transcripts and genes within centromeres can be efficiently transcribed. The transcription of satellites has been observed in numerous eukaryotic species across a broad range of phyla, from yeast to human (Bonaccorsi et al., 1990; Bouzinba-Segard et al., 2006; Epstein et al., 1986; Fukagawa et al., 2004; Lachner and Jenuwein, 2002; Lee et al., 2006; Lehnertz et al., 2003; Li and Kirby, 2003; Neumann et al., 2007; Rudert et al., 1995; Topp et al., 2004; Volpe et al., 2003; Volpe et al., 2002). The wide-spread conservation of satellite transcription is consistent with a conserved regulatory role for these transcripts in gene regulation or chromatin modification (Ugarkovic, 2005).

RNA derived from centromere could bound to chromatin and may have a targeting and/or stabilizing role providing, for example, low-affinity contacts that facilitate higher-order interactions among chromatin proteins because the intrinsic property of RNA that is more flexible than protein and can tolerate rapid sequence divergence while still maintaining function. These centromeric transcripts may function in one of the three ways (O'Neill and Carone, 2009) :

1. They may facilitate post-transcriptional gene regulation (Li and Kirby, 2003) potentially through the RNA-induced silencing complex (RISC).
2. They may participate in the RNA induced transcriptional silencing complex (RITS), a pathway in which siRNAs are involved in heterochromatin recruitment (Volpe et al., 2003).
3. Alternatively, in a manner analogous to the *Xist* transcript in mammalian X-inactivation, they may recruit heterochromatin assembly factors such as histone deacetylases and Polycomb group proteins (Heard, 2005).

Although the mechanisms are unknown, evidence that satellite transcripts participate in heterochromatin assembly and/or nucleosome recruitment at centromeres is accumulating.

3.1 CENTROMERIC TRANSCRIPTS AND EVOLUTION

In *Shizosaccharomyces pombe* centromeres, dsRNAs transcribed from the *dh* and *dg* repeats in the pericentric *otr* region produce siRNAs that are bound to the RITS complex and bring about H3 lysine-9 methylation through the RNA interference pathway (RNAi) (Volpe et al., 2003; Volpe et al., 2002) . The deletion of the RNAi processing factors Argonaute, Dicer, and RNA dependent RNA polymerase in *S.Pombe* results in the loss of centromere function including lagging chromosomes, loss of sister chromatid cohesion, and the loss of histone 3-lysine 9(H3K9) methylation.

In maize, transcripts have been identified from both strands of the 156 bp CentC centromere-specific repeat as well as the centromere-specific CRM retroelement, each of which coimmunoprecipitates with the CENP-A antibody. Although no siRNAs were found in this study (Topp et al., 2004). In A.*thaliana* the loss of function mutants of RNAi components led to the disappearance of these centromeric small RNAs, but dimethyl-H3K9 levels were unaffected and centromere function was apparently normal (May et al., 2005). Otherwise siRNAs have been identified for CentO repeats, the analogous centromere-specific repeat in rice (Lee et al., 2006), indicating that the RNAi pathway may be involved in centromere transcript processing in plants. Thus, a complex interaction of RNAs, modified histones, and DNA define the genomic locations that act as centromeres. Recent work in mouse, human and in tammar (Carone et al., 2009) suggests that this may also be true of mammalian centromeres. Obliteration of dsRNA in mouse results in the loss of centromere foci in interphase nuclei (Maison et al., 2002). Mouse cells null for *dicer*, the gene encoding the enzyme responsible for cleaving dsRNA into siRNAs, show a similar centromere defect (Kanellopoulou et al., 2005; Peters et al., 2001), implicating an RNA silencing pathway in centromere function in mammalian cells through dsRNA processing. Fukagawa et al. (2004) used human–chicken somatic cell hybrids to demonstrate that *dicer* conditional loss of function mutant cells lack centromeric heterochromatin and exhibit an accumulation of centromere satellite transcripts, implicating the need for *dicer* to cleave them into smaller RNAs. From these studies, it has been proposed that centromere satellite transcripts have a role in kinetochore assembly in mammals through kinetochore demarcation and heterochromatin establishment (Fukagawa et al., 2004; White and Allshire, 2004).

The transcription of centromere sequences appears to be under strict regulation in human and mouse cells. Stresses, such as heat shock, nutrient deficiency, apoptosis, and chemical shock

result in genetic instability that ultimately leads to aneuploidy, loss of sister chromatid cohesion, and abnormal chromosome segregation. These defects are directly correlated with aberrant transcription of centromere satellites. In mouse, 120 nt transcripts for the minor satellite accumulate under stress conditions that ultimately lead to abnormal centromere function (Bouzinba-Segard et al., 2006).

Similar aberrant transcript accumulation has been found for satellite III (satIII) satellites in human cells under stress conditions (Valgardsdottir et al., 2005). Loss of function of Dicer in chicken-human somatic cell hybrids (Fukagawa et al., 2004) and in mouse embryonic stem cells (Kanellopoulou et al., 2005; Murchison et al., 2005) also results in the accumulation of long centromere satellite transcripts. The mechanism through which transcription of centromeric satellite sequences is promoted is still unknown. It has been proposed that transcriptional control through retroelements may facilitate the satellite sequence transcription observed in a broad range of vertebrate species (Diaz et al., 1981; Ugarkovic, 2005).

An interesting involvement of centromeric specific sequence in centromere functionality is give by the marsupials. A specific endogenous retroelements, known as KERV, is specific of centromeres of all Macropodines. KERV is characterized by open reading frames for *gag, pro* and *pol* bounded by two identical long- terminal repeats (LTRs) (Ferreri et al., 2004). To investigate the role of transcription and small noncoding RNA, RNA depletion experiments followed by immunocytochemistry localization of centromere and heterochromatin proteins were done. This experiments indicated that RNA is necessary for the recruitment of centromere (CENP-A and CENP-B) and heterochromatin (tri-methyl H3K9) proteins (Carone et al., 2009). Further investigation into the RNA species involved in this association and the transcripts produced from known centromeric sequences and, in particular small noncoding RNA, indicated that small RNA transcripts produced from *M. eugenii* centromeres are not in the size range of siRNA (21–23 nt) as seen for plant and yeast satellite sequences. In contrast, the small RNA produced from the wallaby centromeres are 34–42 nt, a previously unknown size class termed crasiRNAs (centromere repeat associated small interacting RNAs) (Carone et al., 2009).

Centromeric transcripts clearly play a critical role in centromere identity and function. The association of the accumulation of full leght centromere satellite transcripts with centromere failure in stressed cells (Bouzinba-Segard et al., 2006; Valgardsdottir et al., 2005) argues that these transcripts must be processed in some way to function properly but the process by this full length transcripts are processed remains unknown.

3.2 RETROELEMENTS

Dawe (2003) and Wong and Choo (2004) have hypothesized that retroelements and their associated machinery may be integral to centromere functioning based upon three different lines of evidence. First, in plants some transposable elements have a genomic distribution restricted to the centromere. In rice there are centromere-specific retroelements (CRR) (Cheng et al., 2002). Centromere retroelements (CRs) in both maize and rice associate preferentially with CENP-A (Nagaki et al., 2005; Nagaki et al., 2003; Zhong et al., 2002). Similar retrotransposon specificity for centromeres has been identified in many other plant species (reviewed in (Jiang et al., 2003). LTRs act as strong promoters and can retain their transcriptional potential once the sequence becomes integrated into the genome. As they age, these LTR sequences lose their ability to promote transcription through genetic drift and mutation caused by host defense mechanisms. The retention of transcriptional machinery within the CR retroelement LTRs has led (Jiang et al., 2003) to hypothesize that production of RNA transcripts by these LTRs facilitates the establishment of CENP-A domains in the demarcation of the active centromere.

Second, in several cases divergent repeat arrays within centromeres retain features of the retroelements from which they were derived (Wong and Choo, 2004). Thus, satellites found in centromere domains may be derived from retroelements.

Third, at least one centromere protein may have been derived from transposable element machinery. The amino acid sequence of CENP-B, a DNA-binding protein involved in the establishment of centric heterochromatin, could be a candidate. The homologs of CENP-B in *S. pombe* , Cbh1 and Cbh2, both bind otr repeats (Nakagawa et al., 2002). This interaction, likely mediated through siRNAs produced from *dg* and *dh* repeats (Fig 1.1) , is crucial for the establishment of H3K9 methylation at the centromere (Volpe et al., 2003).

While the Dawe/Wong and Choo hypothesis has garnered robust support in plants (Neumann et al., 2007; Topp et al., 2004; Zhong et al., 2002), very little work has been done to test this theory directly in mammals. However, a recent studies by Chueh et al. (2005 and 2009) describes a positive correlation between neocentromere formation and transposable elements in humans, implicating LINE-1 in centromere initiation.

4. NEOCENTROMERE

The term “neocentromere” were first use by Rhoades and Vilkomerson in the 1942 when they found the first one in maize. Then neocentromeres have been reported in other plants and animals.

Neocentromeres in animals are not confined to meiosis as that of maize and are defined like ectopic centromeres that have formed in non-centromeric locations and avoid some of the DNA features that normally characterize canonical centromeres. In contrast to plant neocentromeres that lack fundamental centromere proteins and interact with microtubules in a very different manner, animals neocentromere have been shown to bind all known essential centromere proteins. Human neocentromeres complitly lack of α-satellite. An absolute requirement for the integrity of the cells is to have only one functional centromere. The creation of a new centromere is, therefore, an extraordinary event (Amor and Choo, 2002).

The first human neocentromere that lacked any α-satellute sequence (Voullaire et al., 1993) was detected in 1993 on a mitotic marker chromosome during the routine karyotyping of a boy with learning difficulties. This was a new type of neocentromere, quite different from the first neocentromeres described in maize. This marker, designated “mardel(10)” (Fig. 4.1) was derived from a de novo complex rearrangement of chromosome 10 that had resulted in loss of the original centromere. Despite the complete absence of α-satellite DNA, the neocentromere was able to form a primary constriction and assemble a functional kinetochore that was stable in mitosis. It was one of the first evidence to support the epigenetic nature of the centromerization. Approximately 93 human neocentromeres have been reported to 2010, typically they are located on rearranged marker chromosomes that have similarly lost their centromeres. In 21 of the human chromosomes has been found neocentromere but certain regions appear to have high propensity to form neocentromeres such as chr 13q, 15q, and 16q. Neocentromeres have also been detected in human cancers and have been produced experimentally in *Drosophila*, *S. pombe* and *C. albicans.*

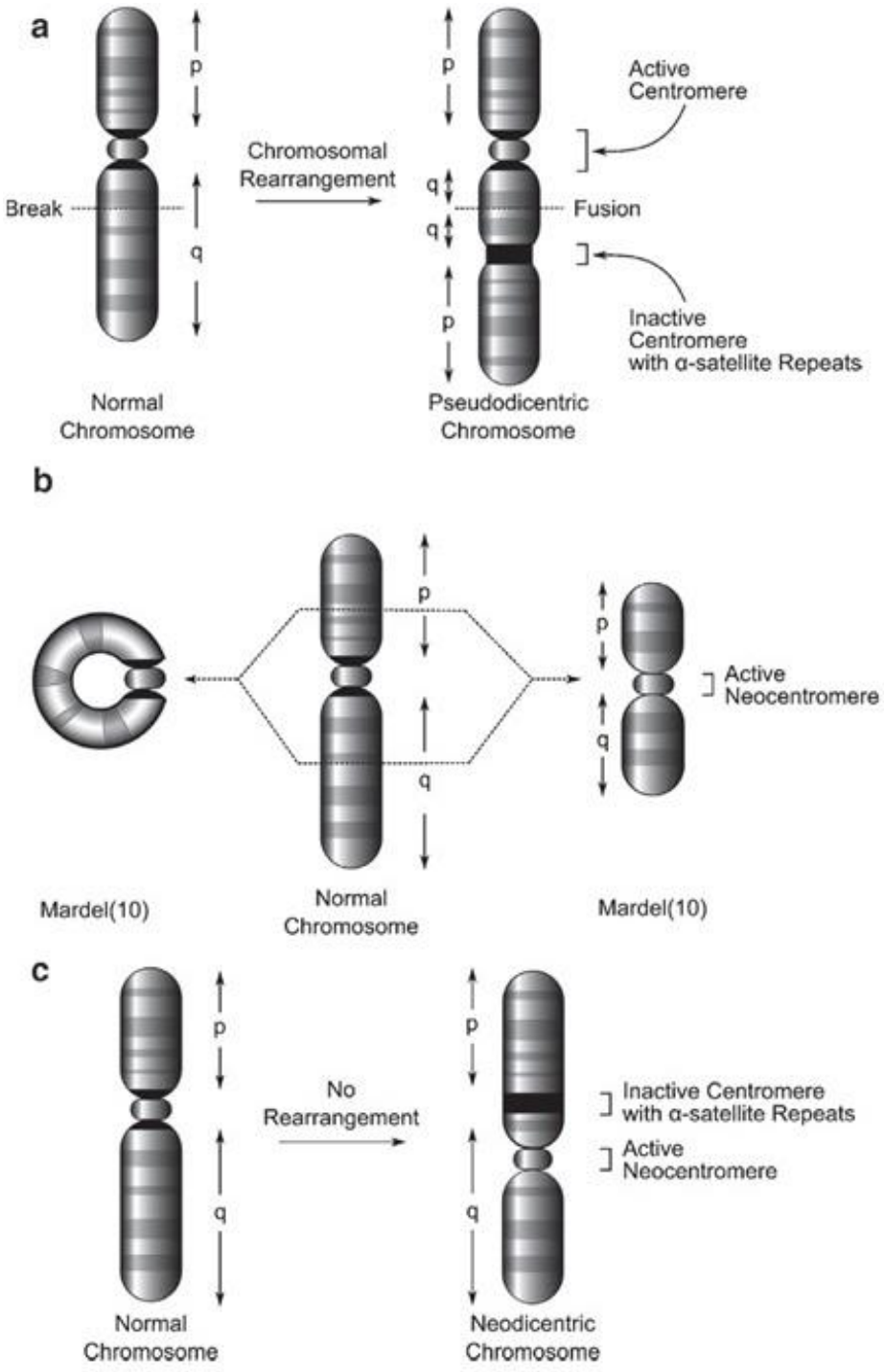


Fig. 4.1 Neodicentric chromosome, clinical and evolutionary neocentromere. (**a**) Dicentric chromosomes typically arise through chromosome fusion. When this happens, the dicentric chromosome may achieve mitotic stability and avoid breakage on the spindle by inactivating one of its centromeres. This forms a pseudodicentric chromosome that contains two distinct a -satellite loci (*shaded in black*), but only one of which acts as a functional centromere. (**b**) The Mardel(10) clinical neocentromeric chromosome (*righ*) was the acentric product of an internal recombination event that looped out the endogenous centromere (circular mini-chromosome, rdel(10), *left*). (**c**) Evolutionary centromere repositioning on neodicentric chromosomes occurs when the functional centromere relocates to a non-alphoid locus in the absence of any DNA rearrangements (Panchenko and Black, 2009).

4.1 NEOCENTROMERES CLASSIFICATION

Nowadays we could classify neocentromeres in two mainly categories:

- ENC (Evolutionary New Centromere) are also termed repositioned centromeres and were discovered in evolutionary studies that clearly showed that the centromeres can reposition along the chromosome during evolution with absent phenotypic evidences. They were documented in a variety of eukaryotes (Cardone et al., 2006; Ventura et al., 2004).
- HCN (Human Clinical Neocentromeres): perfectly functioning analphoid centromeres which emerge in ectopic chromosomal regions (Amor and Choo, 2002; Warburton, 2004). Most HCN arise in acentric, supernumerary chromosomal fragments whose mitotic survival is rescued by the neocentromere. These extrachromosomes results in phenotypic abnormalities that bring these patients into the clinical setting. Neocentromere of this class are found only in human because of the very efficient clinical filter.

It has been suggested that ENC and HCN, could be regarded as two faces of the same phenomenon because of the region where they emerged apparently harbor inherent potentiality to form novel centromeres (Capozzi et al., 2008; Cardone et al., 2006; du Sart et al., 1997; Ventura et al., 2004).

4.2 NEOCETROMERE: A MOLECULAR VIEW

Neocentromeres are of interest for structural, functional and evolutionary information because of their lack of the repetitive elements they are important tool for identifying the main elements of centromere determination.

Protein that discretly and constitutively localize to functional centromeres, such as CENP-A, CENP-C and CENP-H, are present on the neocentromeres and are absent in the inactivate α-satellite DNA centromere. Otherwise CENP-B that specific bind a sequence of the α-satellite DNA do not relocate in the neocentromere and remains at the silenced canonical centromere (Warburton et al., 1997)

CENP-A chromatin immunoprecipitation (ChIP) on ChIP (microarray) analysis of different neocentromeres was used to closely map the relocate site in the genome. CENP-A–binding domain where determined for three different neocentromeres, at 10q25(Lo et al., 2001a);

20p12 (Lo et al., 2001b), and (Satinover et al., 2001) and the range in size was from 330 kb to 500 kb, and at the level of the DNA sequence the only noticeable similarity between the three domains is an increase in AT content. The sequence analysis failed to show similarities or tandemly repeated DNA in common between neocentromeric region (Alonso et al., 2003); (Cardone et al., 2006). Altogether this support the epigenetically determination of neocentromeres with little involvement of the primary DNA sequence. ChIP on ChIP analysis of the neocentromere 13q32 showed precise colocalization of CENP-C and CENP-H with CENP-A organized into distinct major and minor domains (Alonso et al., 2007). Similar analysis failed to find H3K4me2 domain associated with the CENP-A domain in neocentromere (Alonso et al., 2010). Large heterochromatin domain appears to be a ubiquitous feature of metazoan centromeres involved in retention of centromeric sister chromatid cohesion. ChIP on ChIP analysis using antibodies to HP1α and H3K9me3 (specific for pericentromeric chromatin) followed by FISH, where done to further investigate the 13q32 neocentromere. Any significant signal for heterochromatin where detected around the neocentromere region (Alonso et al., 2010).

This discovery could mean that neocentromeres could indeed start off minimal to no heterochromatin structure and still be functional, the fixation of the neocentromere in a species could be accompanied by an expansion of centromeric sequences and heterochromatin at the new centromere, which may be required for increased mitotic stability (Ventura et al., 2007). A documented example of evolutionary centromere fixed in population that is adding the repetitive centromeric element is the cen8 of rice (Yan et al., 2006).

4.3 NEOCENTROMERES AND TUMORS

Although neocentromere formation is in general a rare occurrence, certain cancers are associated with the formation of a complex, rearranged chromosome containing neocentromere. The best characterized link between neocentromeres and a specific form of cancer is found in the “atypical lipomas and well-differentiated liposarcomas” (ALP-WDLPS); adipocyte tumors of borderline malignancy belonging to the heterogeneous group arising from adipose tissue (lipomatous tumors). An apparently primary cytogenetic aberration in these tumors is the presence of supernumerary marker chromosomes, either in the form of rings or remarkable “giant rod”–shaped chromosomes. The rings and giant rods have functional centromeres, as demonstrated by the binding of centromere proteins (such as

CENP-C) and by mitotic stability (Sirvent et al., 2000). ALP-WDLPS, therefore, represents the first example of a tumor class for which the formation of analphoid neocentromeres is a predictable outcome. Neocentromere formation presumably provides a mechanism to impart mitotic stability and, thus, a selective advantage to the neoplastic cells, on what might otherwise be highly unstable acentric supernumerary marker chromosomes (Amor and Choo, 2002).

4.4 ECA11: THE EVOLUTIONARY NEOCENTROMERE OF DOMESTIC *HORSE*

The genus *Equus* radiated into 8 or 9 species around three million years ago (Carbone et al., 2006). Members of the family equidae exhibit diverged karyotypes and variable centromeric positioning.

One unexpected feature of the horse genome landscape was the identification of an evolutionary new centromere (ENC) on chromosome 11 (ECA11), captured in an immature state. Several ENCs have been generated in the genus *Equus* by centromere repositioning. ENCs are believed to form initially by unknown mechanisms in repeat-free regions and then progressively acquire extended arrays of satellite tandem repeats that may contribute to functional stability (Ventura et al., 2007). The centromere of ECA11 resides in a large region of conserved synteny in many mammals, where the horse is the only species with a centromere present, strongly suggesting that this centromere is evolutionarily new. The ECA11 centromere is the only horse centromere lacking any hybridization signal in fluorescence in situ hybridization experiments probing with the two major horse satellite sequences, as if it had not had enough time to acquire satellite DNA. The absence of satellite signals in the ECA11 centromere suggests that this ENC may not have yet “matured” to the point of being endowed with satellite DNA (Wade et al., 2009).

RESULTS

1. EVOLUTIONARY NEW CENTROMERE

Centromere of the chromosome 11 of the *Equus caballus* has been classified as an example of evolutionary neocentromere (Wade et al., 2009). It is in an immature state because of its lacking of highly repeated horse sequences that normally constitute centromeres. This part of the work were done in collaboration with Elena Giulotto, University of Pavia, that had provided us all the horses cell lines used for that work and that firstly characterized by immunofluorescence in situ hybridization ECA11.

1.1 CHARACTERIZATION OF ECA11 NEOCENTROMERE IN HSF

In order to map that centromeric region, ChIP-on-chip experiment (see §2.2 protocol of Mat. & Met.) on the horse fibroblasts cell lines HSF were performed using two rabbit polyclonal antibodies directed against human CENP-A or CENP-C centromeric proteins. These well phylogenetically conserved DNA-binding proteins are required for kinetochore function and are exclusively targeted to functional centromeres. Horse centromeric protein are recognized by the anti-human CENP-A and CENP-C (Fig 1.1).

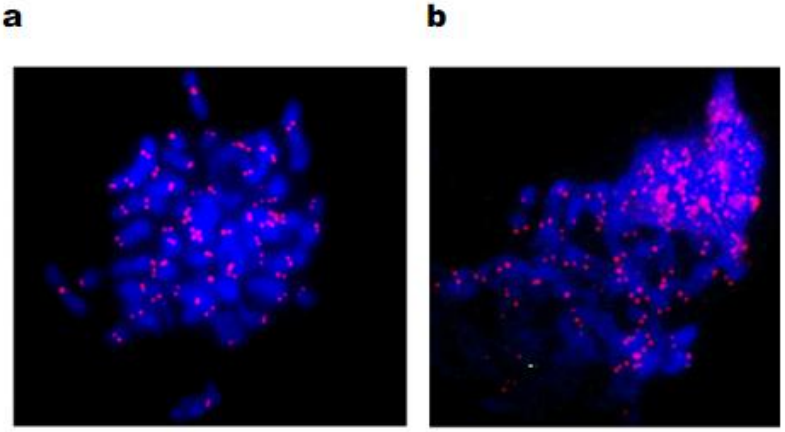


Fig. 1.1
Immuno-localization of CENP-A and CENP-C on equine chromosomes in HSF. Antibodies against human centromeric proteins CENP-A (left) and CENP-C (right) bind horse centromeres (red fluorescence signals). (Wade et al., 2009)

The formaldehyde crosslinked chromatin of fibroblasts was sonicated and then immunoprecipitated using antibodies. The isolated DNA was then amplified using the Whole genome amplification kit (Sigma) and 4ug were hybridized on Nimblegen tiling array (mat e met). The resolution of that array is in average of 100bp. The binding domain of CENP-A and CENP-C correspond to the functional units of the ECA11 centromere.

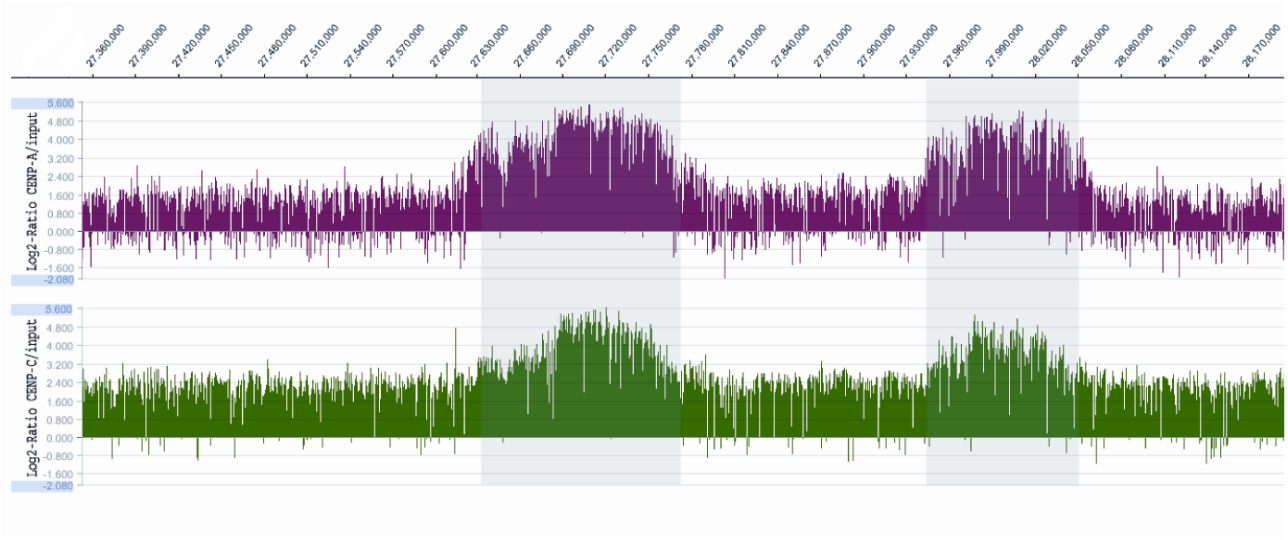


Fig 1.2. Partial view of the ChIP-chip analysis data on chromosome 11, using anti-CENP-A and anti-CENP-C antibodies. Results are presented as the log2 ratio between the hybridization signal obtained with immunoprecipitated DNA using anti-CENP-A (purple) or CENP-C (green) antibodies and that from the input DNA sample. The x-axis shows the genomic position of each oligo on chromosome 11. The data are visualized by the SignalMap software (NimbleGen Systems, Inc.). The CENP-A and CENP-C domains (the shaded area) clearly map at 27.643.412-27.779.345 kb and 27.950.821-28.049.577 kb (Wade et al., 2009)

In the HSF cell line the ECA11 centromere localizes between 27.643.400 and 28.049.600 bp of the chromosome 11 of the horse, a big region of about ~400 kb. Both centromeric proteins bind two distinct but perfectly overlapping region of the genome (Fig. 1.2). The first domain is between 27.643.412-27.779.345 bp (~136 kb) and the second domain is between the position 27.950.821-28.049.577 bp (~99 kb). This results were part of the work "Genome sequence, comparative analysis, and population genetics of the domestic horse" published in Science in November 2009 (Wade et al., 2009).

1.2 THE PECULIARITY OF ECA11 CENTROMERE

The peculiarity of the double binding domain of centromeric proteins (never seen before in literature) led us to investigate the ECA11 centromere in other fibroblast cell lines derived from different horses. The initial purpose was to validate this results.

Native chromatin-immunoprecipitation experiments (§2.1 of Mat & Met) were performed on the following horse fibroblasts cell lines: HSF-B, HSF-C, HSF-D, HSF-E, HSF-G. Nuclei from cells were digested with micrococcal nuclease (to have a range of ~180bp length fragments) and then immunoprecipitated with rabbit CENP-A polyclonal antibody. Whole genomic amplification (WGA) of the chipped anti-DNA was performed and that was hybridized onto a Nimblegen Custom Genomic array (§4 of Mat & Met).

Primers spanning over the CENP-A\C domain previously found in HSF were disegned (§ 3.1 of Mat. & Met.) and used to validate the enrichment of this N-ChIP experiments by Real Time PCR. Region inside and outside the CENP-A binding domain were chosen to normalize datas. For each of that chosen region the Log_2 Ratio CENP-A/INPUT was evaluated: positive values are associated to an enrichment that corresponds to a region of binding of the protein onto the genome. The Real Time PCR data are summarized in the Table 1.1.

	27,569,560-27,569,649	27,687,704 27,687,797	27,739,831-27,739,923	27,770,997-27,771,100	27,934,666-27,934,768	27,966,050-27,966,138	27,985,954-27,986,054	27,990,583-27,990,679	28,227,839-28,227,938
HSF-B	+	-	+	+	-	-	-	-	-
HSF-C	+	+	-	-	-	-	-	-	-
HSF-D	-	+	+	+	-	-	-	-	-
HSF-E	+	+	-	-	-	-	-	-	-
HSF-G	-	+	+	+	+	+	-	-	-

Table 1.1.
Real Time data summary. The + symbol indicates an enrichment of the Log2 Ratio CENP-A/INPUT for the region of the chr11 tested. Red color corresponds to the CENP-A binding domain in the ECA11 of HSF cell line.

Analyzing the real time data, the extension and the position of the binding region seem to change among cell lines. To better understand the meaning of that data, we deepen the investigation higher the resolution of the mapping.

Horse fibroblasts of each cell lines were processed as explained before: chromatin was digested with micrococcal nuclease and immunoprecipitated with rabbit α-CENP-A antibody. Whole Genome Amplification (Sigma) was perform on each sample in order to amplify isolated DNA to enrich the quantity necessary for the array hybridization. In this way 4 ug of immunoprecipitated DNA along with the specific INPUT sample where differently labeled and co-hybridizated onto NimbleGen Custom Tiling Array 385K.

The figure 1.3 shows the results of the array hybridization derived from each cell lines. The graph display the value of the log2 ratio of the fluorescent signal emitted by CENP-A immunoprecipitated DNA over its INPUT signal organized as a function of the genomic position. Different displayed colors correspond to different cell lines.

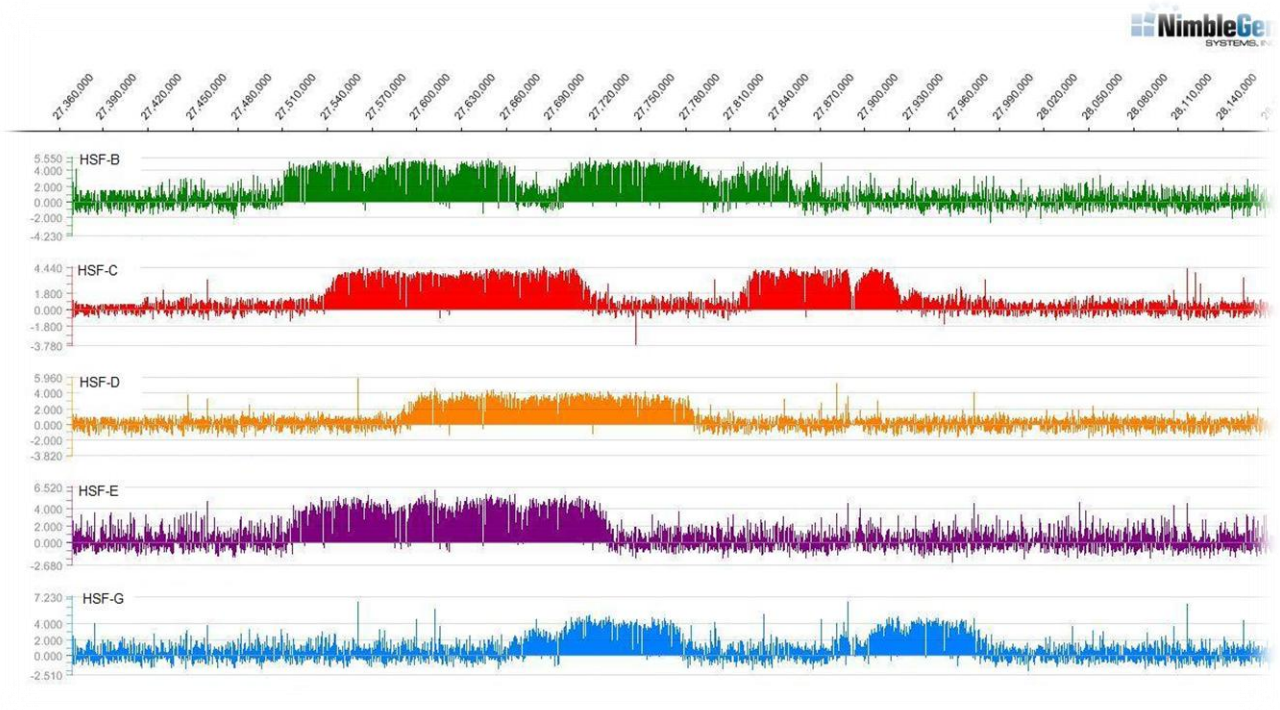


Fig. 1.3 Partial view of the ChIP-ChIP analysis data on ECA chromosome 11. Results are presented as the log2 ratio between the hybridization signal obtained with immunoprecipitated DNA using anti-CENP-A antibodies and that from the input DNA sample. The **x**-axis shows the genomic position of each oligo. Each lane and color represents a different horse cell line:in green the HSF-B, in red the HSF-C; in orange the HSF-D; in purple the HSF-E and in blue the HSF-G . Each cell line differ for the CENP-A binding domain.

1.4 ANALYSYS OF CENP-A BINDING DOMAIN

The raw Nimblegen data were analyzed by a statistical on-line server in order to detect peaks of signals that correspond to the binding sites of the protein onto the genome to finely locate the boundary. The name of the server we used was TAMALPAIS (Bieda et al., 2006) : http://chipanalysis.genomecenter.ucdavis.edu/cgi-bin/tamalpais.cgi

The way of working of the server is explain in detail in the section 4.1 of Materials & Methods. In this way we were able to define the boundary of the CENP-A peaks of binding for each horse cell lines tested.

As shown in Fig 1.4 in the HSF-B cell line (green) two CENP-A binding domains are distinguishable in the centromeric region. The first peak is between 27.514.628-27.666.137 bp (total length of 150 kb) and the second one is between 27.703.686-27.847.139bp (143 kb). As well, HSF-C cell line (red) shows two distinct CENP-A binding domains but they are shifted compared to those of the HSF-B: 27.541.127-27.719.151 bp (178 kb) and 27.819.572-

27.933.035 bp (113, 5 kb). On the contrary, both the HSF-D (orange) and the HSF-E (purple) cell lines show only one CENP-A peak but again they differ for the genomic position: in HSF-D the centromeric domain is between 27.594.542-27.780.073 bp (185,5 Kb) while in HSF-E the peak is shifted between 27.528.964-27.717.783 bp (188,8 Kb). In the HSF-G cell line (blue) two peaks are distinguishable: the first in the region between 27.679.870-27.770.190 bp (91 Kb of length) and the second between 27.904.271-27.967.926 bp (64 Kb).

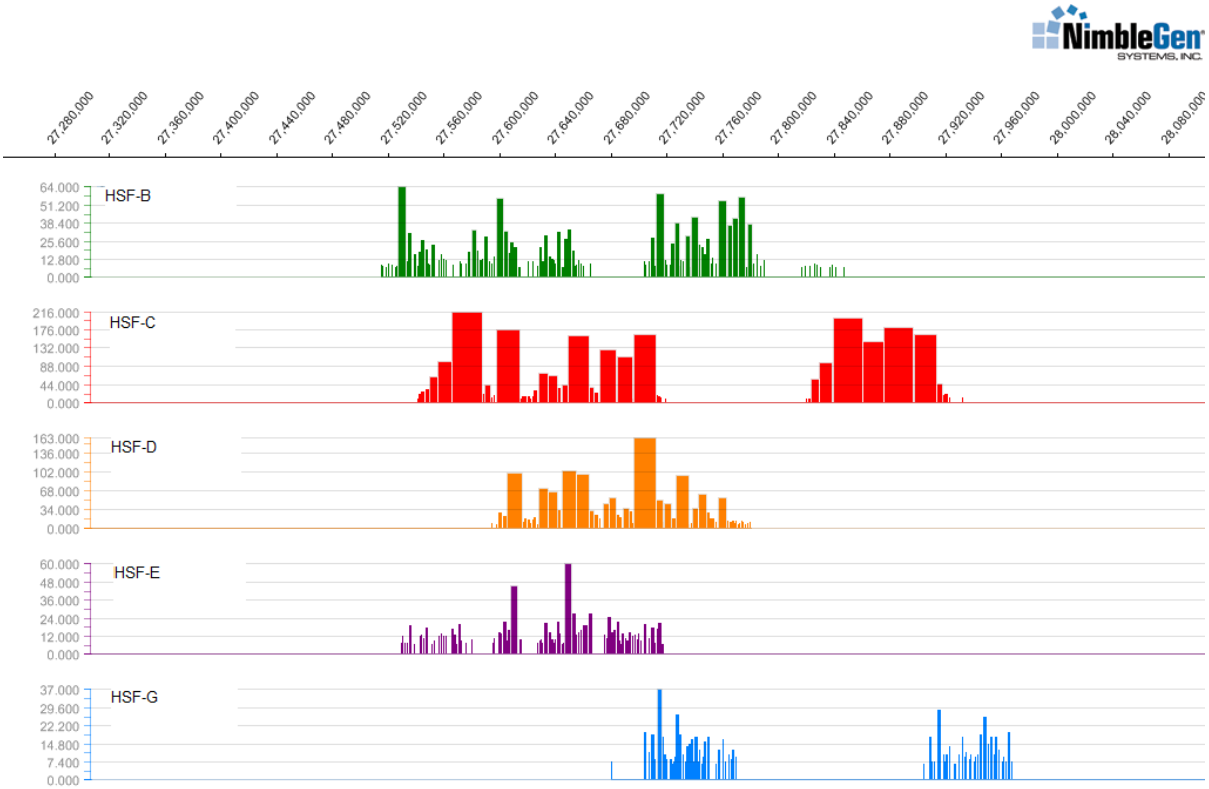


Fig. 1.4 TAMALPAIS statistical analysis of ChIP-on-chip data (T=0.2 P=0.0001). The **x**-axis shows the genomic position of each oligo. Each lane and colour represents a different horse cell line: in green the HSF-B, in red the HSF-C; in orange the HSF-D; in purple the HSF-E and in blue the HSF-G . Each cell line differs for the CENP-A binding domain

The data arrays analysis confirmed the huge variability of the ECA 11 centromeric position and dimension between cell lines derivated from different individual of the same species. The sequence analysis of this region revealed no protein coding sequences, normal levels of noncoding conserved elements, and typical levels of interspersed repetitive sequences, but no satellite tandem repeated sequences. We also found no evidence of accumulation of L1 transposons or KERV-1 elements, which were previously hypothesized to influence ENC formation.

1.4 SINGLE NUCLEOTIDE POLYMORPHISM ANALYSIS ON HSF *HORSE* CELL LINE

In order to better investigate the double CENP-A binding domain that differently characterized the centromeric ECA11 region of different horse cell lines, we decided to investigate SNPs *(single nucleotide polymorphisms)* identified in the sequence inside each peaks. We used the on-line information published on the http://www.broadinsitute.org/mammals/horse/snp web site. We first identified SNPs in the CENP-A binding domain of HSF cell line: we found 263 SNPs inside the first peak (27.643.412-27.779.345 bp) and 37 SNPs in the second peak (27.950.821-28.049.577 bp). Specific primers for regions of interest (to amplify 1kb of genomic DNA) were designed. The amplified fragment were then sequenced by Macrogen in order to determine the heterozygosity or the homozygosity of each SNPs in our specific cell line (Fig 1.5).

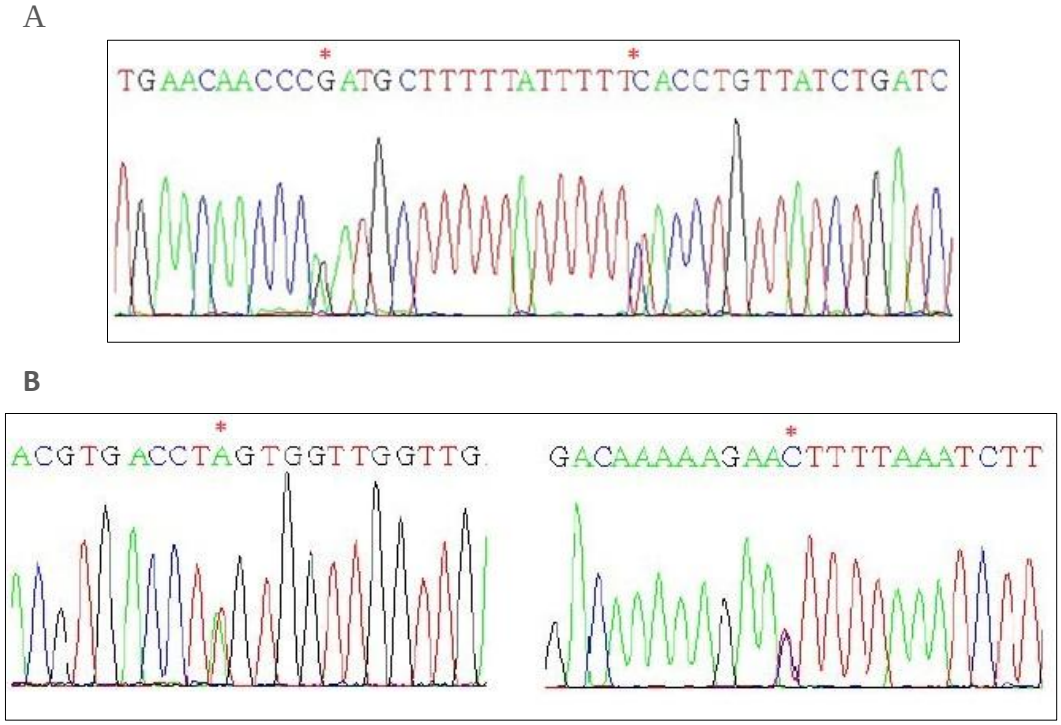


Fig. 1.5 Electropherograms of HSF genomic sequencing. The region tested was the centromeric region identified by our ChIP-on-chip data: A) genomic coordinate chr11: 27648523-27649126, primers used are the 523; B) genomic coordinate: chr11:27966426-27967111. Primer used are 426. The red stars identified the SNPs in heterozygosity.

A selection of SNPs shown to be in heterozygosity in HSF CENP-A domain were chosen and specific primers were designed. That primers should be able to amplify DNA fragments

obtained from N-ChIP experiments (that are fragmented to an average of 200bp). The purpose is to see if the heterozygosity of the SNP tested is maintained in the immunoprecipitated sample or not. The INPUT of ChIP experiment is representative of the total genome and acts as internal control.

Amplified PCR products were purified on Mini Elute PCR purification (QIAGEN) columns and 1ug of that with the primers of interest were send to Macrogen to be sequenced.

The figure 1.6 shows some electropherograms resulting from the sequencing of PCRs on chipped DNA. Notable all the heterozygosis tested in the genome (INPUT) became homozygosis in the centromeric immunoprecipitated sample (CENP-A domain). This mean that the two CENP-A domains (27.643.412-27.779.345 kb and 27.950.821-28.049.577 kb) are differently organized on homolog chromosomes. CENP-A is not present on both that region contemporary on the same chromosome. This reveals differences even between the two homologous chromosomes for the centromere position.

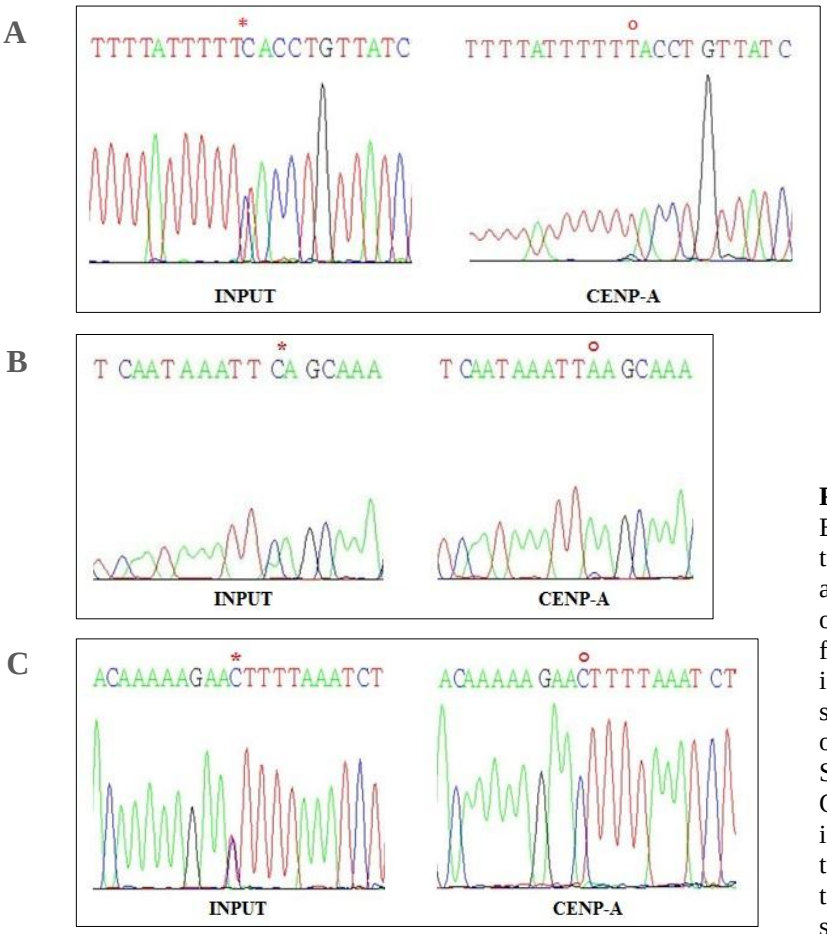


Fig. 1.6 Electropherograms. Each panel shows on the left the electropherograms associated to the INPUT and on the right the one of the fragment CENP-A immunoprecipitated. For both samples the genomic position of the SNPs showed are: A) SNP1picco1, B) SNP1picco3 C) SNP2picco1. The red star identified a heterozygosity that became homozygosity in the immunoprecipitated sample (red circles).

1.5 SINGLE NUCLEOTIDE POLYMORPHISM ANALYSIS ON HSF-G AND HSF-D HORSES CELL LINES

We decided to proceed with the SNPs status determination of the ECA 11 region of some other horse cell lines previously ChIP-on-chip mapped (fig 1.3).

We decided to use the HSF-G cell line that, as well as HSF already tested, showed a double CENP-A binding domain and the HSF-D that interestingly showed just one huge CENP-A domain. As already explained (§1.4) we chose the region inside the CENP-A binding domain enriched for SNPs in heterozygosis in horse population (published on the http://www.broadinsitute.org/mammals/horse/snp web site) to define the subset that are heterozygotic in the genome of the cell lines of interest. Then we proceeded with the SNPs status determination of that subset on the N-ChIP samples both the CENP-A immunoprecipitated and the Input. We analyzed a total of 8 SNPs in the HSF-G cell line: 4 SNPs for each CENP-A binding domain. The fig. 1.7 shows two different SNPs comparing the input and the immunoprecipitated samples. As for the HSF cell line, all the heterozygosis tested in the HSF-G genome (INPUT) became homozygosis in the centromeric immunoprecipitated sample (CENP-A).

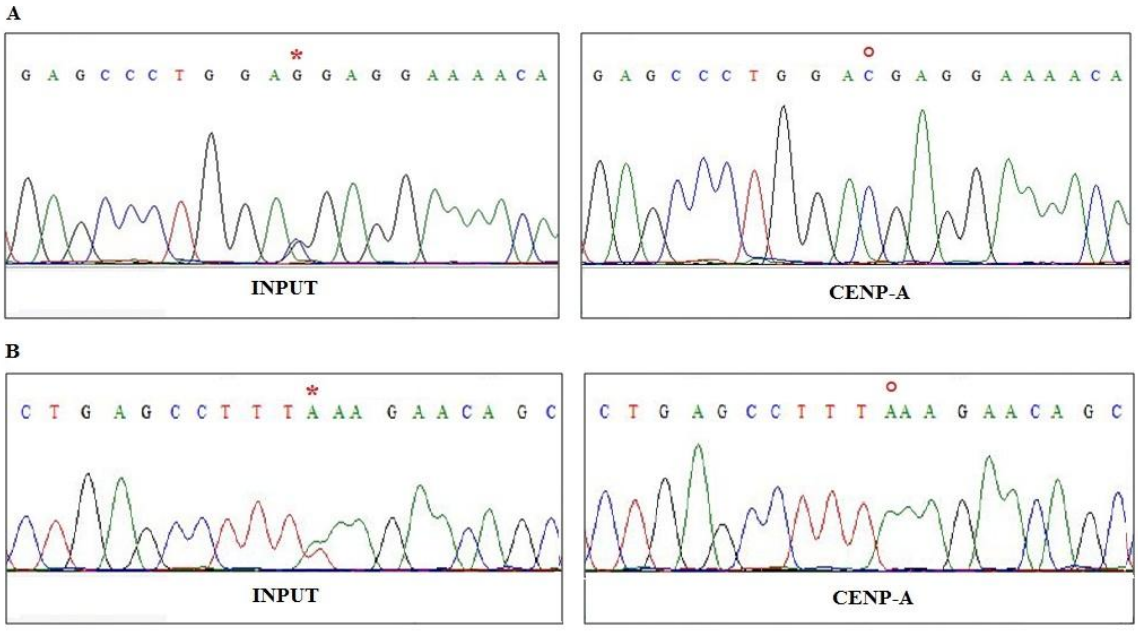


Fig. 1.7 Electropherograms on HSF-G horse cell line. Each panel: on the left is the electropherograms of the INPUT and on the right the CENP-A immunoprecipitated. The genomic position of the SNPs are: A) chr11:22.728.254 within the first CENP-A binding domain; B) chr11:27.966.662 within the second CENP-A binding domain. Red stars identified heterozygosities that became homozygosities in the CENP-A sample (red circles).

We then decided to use HSF-D cell line because its unique CENP-A binding region. We tested a total of 10 SNPs along that region (chr11: 27.594.542-27.780.073 bp) heterozygotic in that cell line, some map on the boundary other in the core of that region. What we found was interesting: the SNPs at each of the border became omozygotic in the CENP-A immunoprecipitated (fig 1.8) instead the SNPs mapped in the core still be eterozygotic in the immunoprecipitated (fig 1.9).

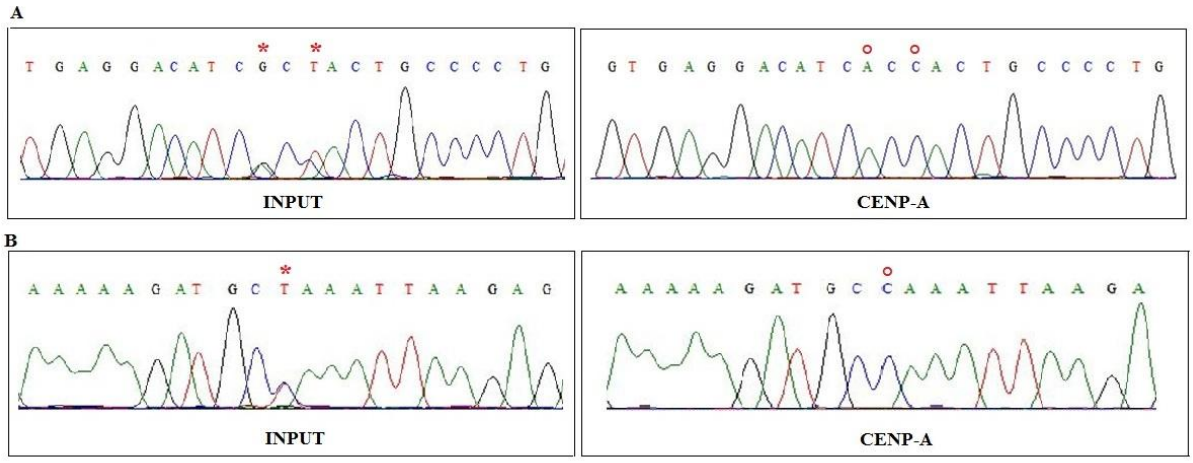


Fig 1.8 Electropherograms on HSF-D horse cell line. In each panel, on the left is the electropherograms of the INPUT and on the right the one of the CENP-A immunoprecipitated. For both samples the SNPs tested are at the genomic position: A) chr11:27.728.443 and 27.728.445; B) chr11:27.744615. Red stars identify heterozygosities that became homozygosities in the CENP-A (red circles).

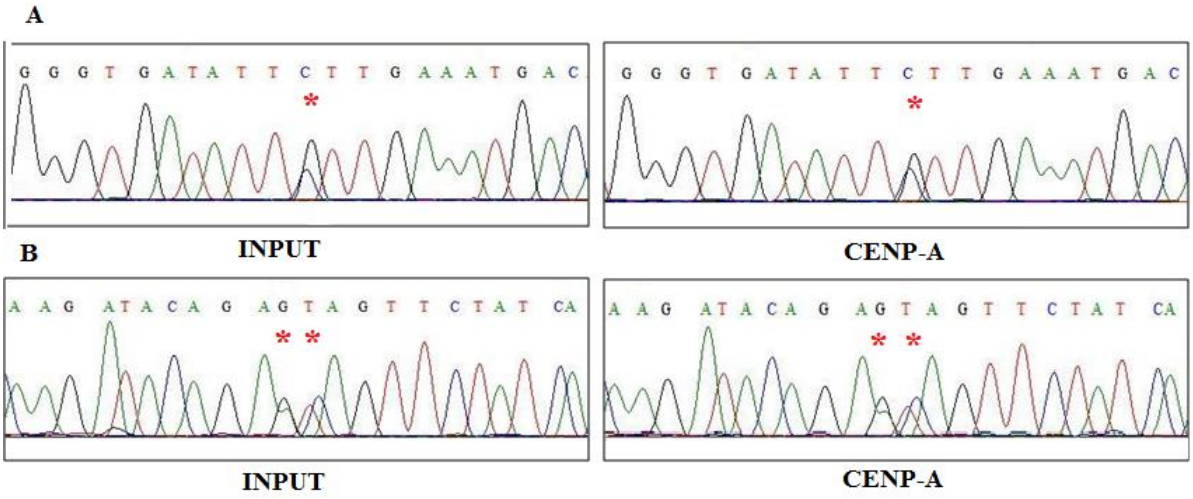


Fig. 1.9 Electropherogrms on HSF-D horse cell line. For each panel: on the left is shown the electropherograms associated to the INPUT sample and on the right the one of the CENP-A immunoprecipitated sample. The position of the SNPs shown are: A) chr11:27,691,524 ; B) chr11:27,691,648 and chr11:27,691,649. Each red star identified heterozygosities.

What we first consider a unique CENP-A binding domain was instead a partially overlapping CENP-A domain differing between the two homologous chromosomes. As well as in HSF and in HSF-G the CENP-A binding domain in ECA11 of HSF-D slightly differs from the two chromosomes.

2. HUMAN NEOCENTROMERES

Neocentromeres are of interest for structural, functional and evolutionary information. Because of their lack of the repetitive elements, they are an important tool for identifying the main elements of centromere determination.

Through the collaboration with Mariano Rocchi, University of Bari, we were able to collect four neocentromeric cell lines (HL-neo6,HL- neo9, HL-2887, HL-portnoi) that differ for the nature and the position of the marker centromere. The clinical filter (amniocentesis) is the best way to find patients that harbor neocentromere. Notice that neocentromere formation do not always associate to detectable clinical problems. The Mariano Rocchi lab. generally immortalizes a lymphoblastoid cell line from periferical blood of the patient and then locates the band where neocentromere arises with FISH (*Fluorent InSitu Hybridization*) experiments on methaphase chromosomes. This first step gives us a preliminary idea, because of the low resolution, of the interested region. On that lymphoblastoid cell line we proceed to an higher resolution mapping of neocentromere position at the sequence level throught ChIP-on-chip experiments.

2.1 HUMAN CHROMOSOME 6 NEOCENTROMERE

Neo6 cells (§1 Mat & Met) derive from patient who don't show any metabolic and behavioural deficits. Cells were crosslinked by adding formaldehyde to a 1% final concentration directly to the culture medium. Chromatin was immunoprecipitated with anti-CENP-A and anti-CENP-C polyclonal antibodies. These DNA-binding proteins are required for kinetochore function and are exclusively targeted to functional centromeres. Thus, the immunoprecipitation of the DNA bound to these proteins allows the isolation of centromeric sequences, including those of the neocentromere. The immunoprecipitated and purified DNA was amplified using the Whole Genome Amplification kit (Sigma-Aldrich) and hybridized to a NimbleGen custom tiling array (§ 4 Mat & Met.), which has an average resolution of about

100 bp. The enrichment of ChIP DNA, before and after amplification, was validated by real-time PCR with specific primers.

DNA-binding peaks were identified by using the statistical model TAMALPAIS (§ 4.1 Mat.& Met.). The analysis showed a clear-cut and unique peak at 6p22.1 (chr6:26,407–26,491 kb for CENP-A, and at chr6:26,415–26,491 kb for CENP-C) using very stringent conditions (98th percentile threshold and $P < 0.0001$).

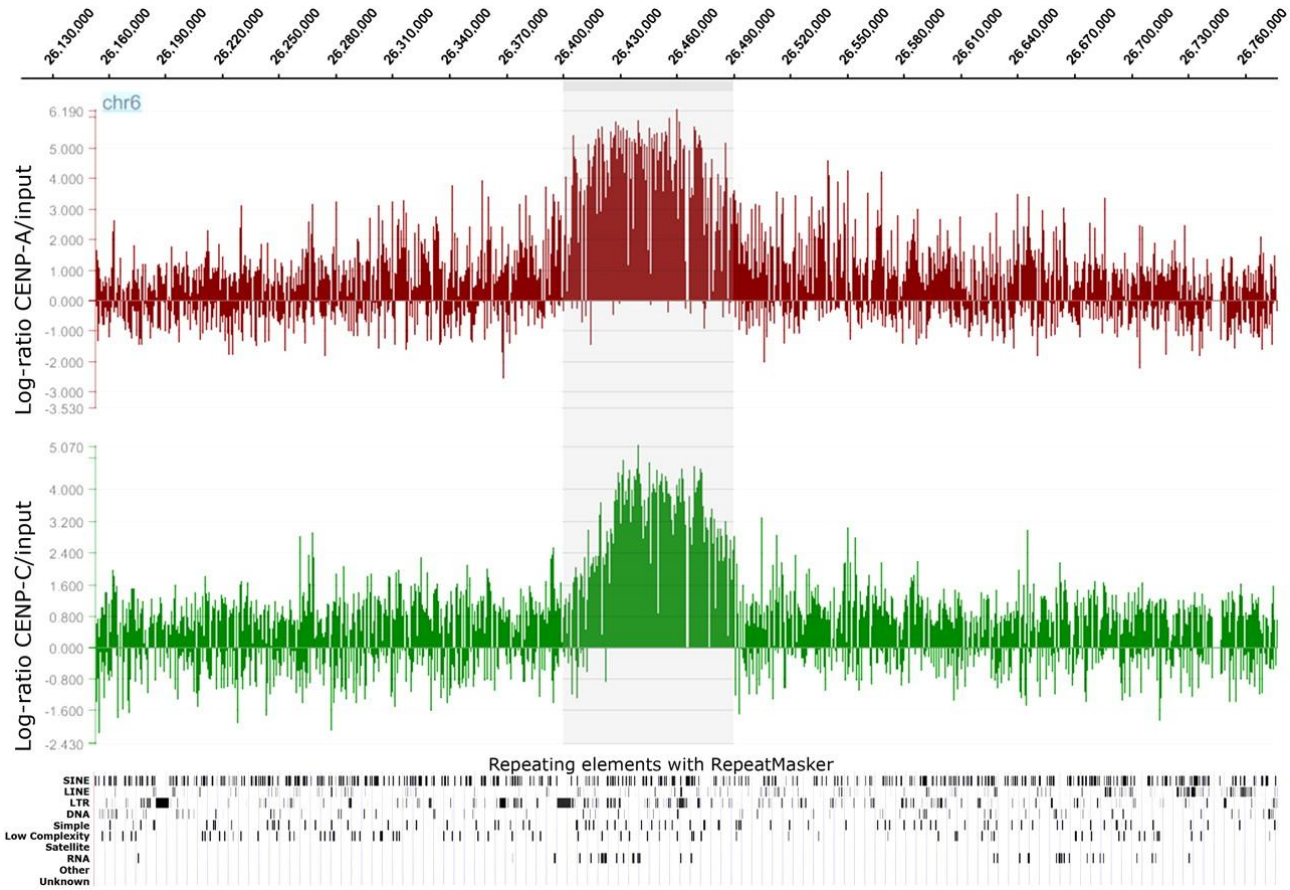


Fig 2.1 Partial view of the ChIP-chip analysis data on chromosome 6, using anti-CENP-A and anti-CENP-C antibodies. Results are presented as the log2 ratio between the hybridization signal obtained with immunoprecipitated DNA using anti-CENP-A or CENP-C antibodies and that from the input DNA sample. The **x**-axis shows the genomic position of each oligo on chromosome 6. The data are visualized by the SignalMap software (NimbleGen Systems, Inc.). Details of the microarray structure are reported at the NimbleGen site (http://www.nimblegen.com) . The CENP-A and CENP-C domains (the shaded area) clearly map at chr6:26,407–26,491 kb and chr6:26,415–26,491 kb, respectively. Below is shown the RepeatMasker analysis of the interspersed repetitive DNA elements as deducted by the UCSC Genome Browser. The "RNA" lane includes the tRNA elements.

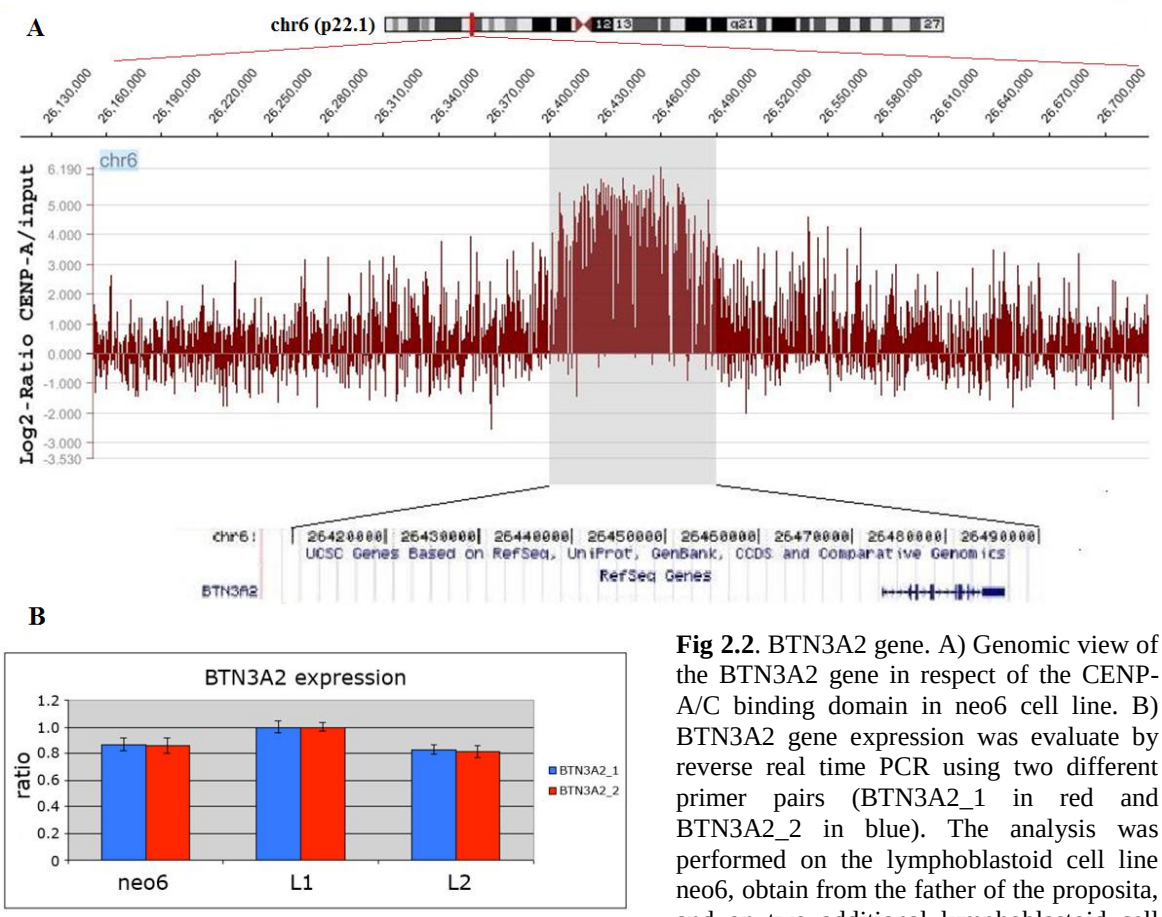
2.1.1 SEQUENCE ANALYSIS OF NEO6 CENP-A/C DOMAIN: BTN3A2 GENE

A very peculiar feature of the region chr6:26,394–29,064 kb, which includes the CENP-A/C domain (chr6:26,407–26,491 kb), is a massive clustering of tRNA (included in the "RNA" lane of the "Repeating Elements by RepeatMasker" track in UCSC browser, see the bottom of Fig. 2.1) The CENP-A/C domain, in addition, showed an AT content of 57.24% (average genome: 57.2%). The spanning of the different types of repeat elements in the CENP-A/C domain, in the flanking regions, on the entire chromosome 6, and in the human genome, is reported in Table 2.1.

	kb Interval					
Repeats	Percent in region 26,107–26,407	Percent in CENP-A/C domain 26,407–26,491	Percent in region 26,491–26,791	Average percent at 6p22.1 26,100–29,900	Average percent on human chr6	Average percent on human genome
SINE (total)	19.26%	19.76%	16.92%	14.79%	11.61%	13.64%
Alu						
Total	17.80%	17.62%	15.10%	13.31%	9.19%	10.77%
*Alu*J	2.40%	4.35%	2.99%	2.09%	2.03%	2.45%
*Alu*S						
Total	12.10%	12.12%	10.57%	8.93%	5.52%	6.44%
*Alu*Sx	6.21%	6.27%	5.89%	4.44%	2.85%	3.42%
*Alu*Y	2.47%	0.37%	0.70%	1.82%	1.31%	1.48%
MIRs	1.45%	2.14%	1.81%	1.48%	2.42%	2.87%
LINE (total)	6.18%	7.48%	14.40%	19.25%	22.00%	21.38%
LINE1	3.19%	4.50%	10.62%	16.74%	17.81%	17.65%
LINE2	2.97%	2.98%	3.78%	2.30%	3.23%	3.26%
LINE3	0.00%	0.00%	0.00%	0.16%	0.38%	0.35%
LTR (total)	16.27%	17.67%	10.22%	12.65%	8.76%	8.72%
MaLRs	3.36%	4.40%	4.10%	3.46%	3.72%	3.79%
ERVL	1.15%	4.41%	2.26%	2.74%	1.72%	1.60%
ERV-class I	7.73%	8.86%	3.18%	5.36%	3.01%	3.01%
ERV-class II	4.03%	0.00%	0.68%	1.10%	0.32%	0.31%
Total repeats	45.41%	50.19%	46.10%	51.66%	47.33%	48.80%

Table 2.1 Repeat element distribution in the CENP-A/C domain and in the two flanking region. Alu and LINE spanning in the CENP-A/C domain, in the 300 kb flanking on both sides, this domain, with reference to the entire chromosome 6 and to the entire genome (Capozzi et al., 2008)

Within the CENP-A/C domains there is the BTN3A2 gene (chr6:26,473,377–26,486,527). This gene encodes a member of the immunoglobulin superfamily, containing two Ig domains with similarity to Ig variable and Ig constant domains. The BTN3A2 expression, evaluated by reverse real-time PCR in the lymphoblastoid cell line derived from the father of the proposita, was found to be very similar to two other lymphoblastoid (L1 and L2) cell lines taken as a reference (see Fig 2.2). This result agrees with the previous studies on two neocentromere cases, which have shown that neocentromere formation does not affect the expression of genes that are located inside or near the CENP-A/CENP-C domain (Saffery et al., 2001) (Lam et al., 2006). This data have been published (Capozzi et al., 2008).



(Capozzi et al., 2009)

Fig 2.2. BTN3A2 gene. A) Genomic view of the BTN3A2 gene in respect of the CENP-A/C binding domain in neo6 cell line. B) BTN3A2 gene expression was evaluate by reverse real time PCR using two different primer pairs (BTN3A2_1 in red and BTN3A2_2 in blue). The analysis was performed on the lymphoblastoid cell line neo6, obtain from the father of the proposita, and on two additional lymphoblastoid cell lines (L1 and L2) as control. The value are the average of three distinct measurements.

2.2 HUMAN CHROMOSOME 9 NEOCENTROMERE

Unlike neo6 showed before, neo9 is classified like a human clinical neocentromere (HCN) because it arises in an acentric, supernumerary chromosomal fragments (organized in a ring of 12 Mb) whose mitotic survival is rescued by the neocentromere formation in position 9q33.1. Lymphoblastoid cells, derived from the mother and showing the ring in about 70% of the cells, were processed for chromatin immunoprecipitation experiments. Cells were crosslinked in situ by adding formaldehyde to a 1% final concentration directly to the culture medium. Chromatin was immunoprecipitated with anti-CENP-A and anti-CENP-C polyclonal antibodies (Trazzi et al., 2009), then ChIP DNA was amplified using the Whole Genome Amplification kit (Sigma) and hybridized to a NimbleGene Whole-Genome Tiling array (HG18Tiling Set 22), which has an average resolution of 100 bp. DNA binding peaks were identified by the statistical model TAMALPAIS (Bieda et al., 2006). The analysis showed a clear-cut and unique peak at chr9, 121.261–121.315 Mb (9q33.1) for both CENP-A

(Fig. 2.2.1a) and CENP-C (Fig. 2.3b), using very stringent conditions (98th percentile threshold and $P < 0.0001$). The region lies internally to the ring chromosome, as predicted by the cytogenetic analysis. The sequence of this region was analyzed for repeat content (RepeatMasker) (Fig. 2.3c). Density of the long interspersed nuclear elements (LINE1, LINE2 and LINE3) and of the mammalian-wide interspersed repeats (MIRs) within the CENP-A and CENP-C domains was about twofold higher (35.03%, LINE1; 5.17%, LINE2; 1.38%, LINE3; and 4.96%, MIRs) as compared to the human genome average (16.89%, LINE1; 3.22%, LINE2; 0.31%, LINE3; and 2.54%, MIRs).

This data have been published (Capozzi et al., 2008).

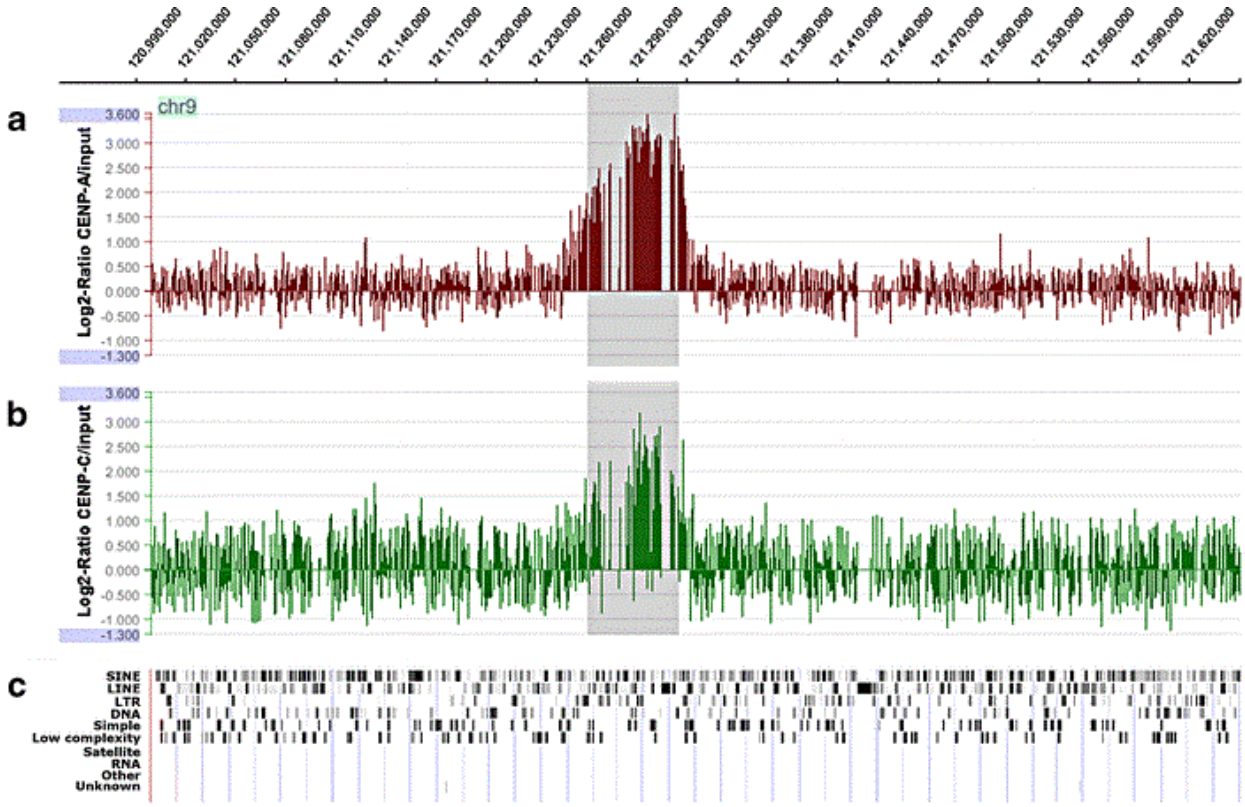


Fig 2.3 Partial view of the ChIP-on-ChIP analysis data on chromosome 9. Results are presented as the log2 ratio between the hybridization signal obtained with immunoprecipitated DNA using anti-CENP-A (a) and anti-CENP-C (b) antibodies and that given by the input DNA sample. The *X* axis shows the genomic position of each oligo on chromosome 9. The data are visualized by the SignalMap software (NimbleGene Systems, Inc.). The shaded region indicates the location of the CENP-A and CENP-C domains. c) RepeatMasker analysis of the interspersed repetitive DNA elements as shown by the UCSC Genome Browser.(Capozzi et al., 2008)

2.3 PORTNOI AND 2887 NEOCENTROMERES

We decided to analyzed two other cell lines harboring clinical neocentromeres that were already known in literature: HL-Portnoi and HL-2887.

HL-2887 is a lymphoblastoid cell line derived from peripheral blood of a moderate mentally retarded man that was referred in 1997 because of facial dysmorphism (Petit and Fryns, 1997). The marker chromosome originate from a deleted chr2 fragment whose mitotic rescue was preserved by neocentromere formation. The marker chromosome was present in 100% of the lymphocyte metaphases.

HL-Portnoi is a lymphoblastoid cell line derived from peripheral blood of a normal intelligence 22 years old man that was referred in 1999 because of pigmentary cutaneous anomalies . The patient was healthy and not dysmorphic. The marker was present in 30% of the lymphocyte metaphases. The marker of the G group size was acrocentric without satellites. By FISH experiments was determined that it was originated from chromosome 3.

Native chromatin, and not crosslinked chromatin, was immunoprecipitated (see §2.1 of Mat. & Met. To see the protocol) with anti-CENP-A polyclonal antibodies (Trazzi et al., 2009), then ChIP DNA was amplified using the Whole Genome Amplification kit (Sigma-Aldrich, St. Louis, USA) and hybridized to a NimbleGene Custom Genome Tiling array (HG18 releasing Mar.2006), which has an average resolution of 100 bp. DNA binding peaks were identified by the statistical model TAMALPAIS (§ 4.2 Mat & Met.*)*.

The analysis of the Portnoi chr3 showed a clear-cut and unique peak at the position chr3:191.564.181- 191.642.140 (3q28) (Fig. 2.4 A,B), using very stringent conditions (98th percentile threshold and $P < 0.0001$). Density of the long interspersed nuclear elements (LINE1) within the CENP-A domain was about twofold fewer (7,20%) as compared to the human genome average (16.89%). Mammalian-wide interspersed repeats (MIRs, 3,49%,) and the long interspersed nuclear elements (LINE2, 3,75%, and LINE3, 0,89%,) inside the binding region did not significantly differs from the average genome average (2.54%, MIRs and 3.22%, LINE2; 0.31%, LINE3).

Fig 2.4. Partial view of the ChIP-on-ChIP Portnoi analysis data on human chromosome 3. A) Results are presented as the log2 ratio between the hybridization signal obtained with immunoprecipitated DNA using anti-CENP-A antibodies and that given by the input DNA sample. The *X* axis shows the genomic position of each oligo on chromosome 3. The data are visualized by the SignalMap software (NimbleGene Systems, Inc.). **B)** DNA binding peak (chr3:191.564.181-191.642.140) identified by the statistical model TAMALPAIS (98th percentile threshold and $P < 0.0001$). **C**) Genomic localization of genes and RepeatMasker analysis of the interspersed repetitive DNA elements as shown by the UCSC Genome Browser

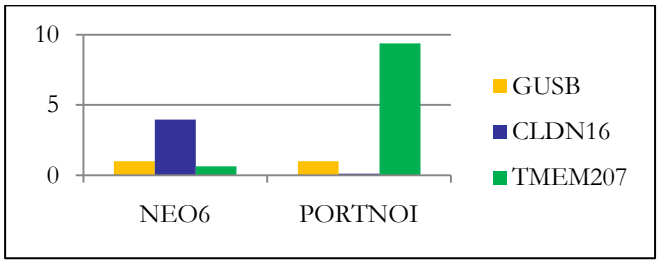


Fig. 2.5 CLDN16 (in blue) and TMEM207 (in green) genes expression. The housekeeping gene taken as reference was GUSB (yellow). Reverse real time PCR on total RNA from HL-portnoi and HL-neo6 cell lines were done. The expression of the two genes differently changed after neocentromerization. The value are the average of three distinct measurements.

The analysis of the 2887 chr2 showed a clear-cut and unique peak at chr2:41.603.764-41.680.907 (p22.1) (Fig. 2.6 b) using very stringent conditions (98th percentile threshold and $P < 0.0001$). The sequence analysis revealed that density of the long interspersed nuclear elements within the CENP-A domain was higher than that of the human genome average (21,60% compared to 16.89% for LINE1 and 6.12% against 3,22% for the LINE2).

Mammalian-wide interspersed repeats (MIRs) did not significantly differs from the genome average.

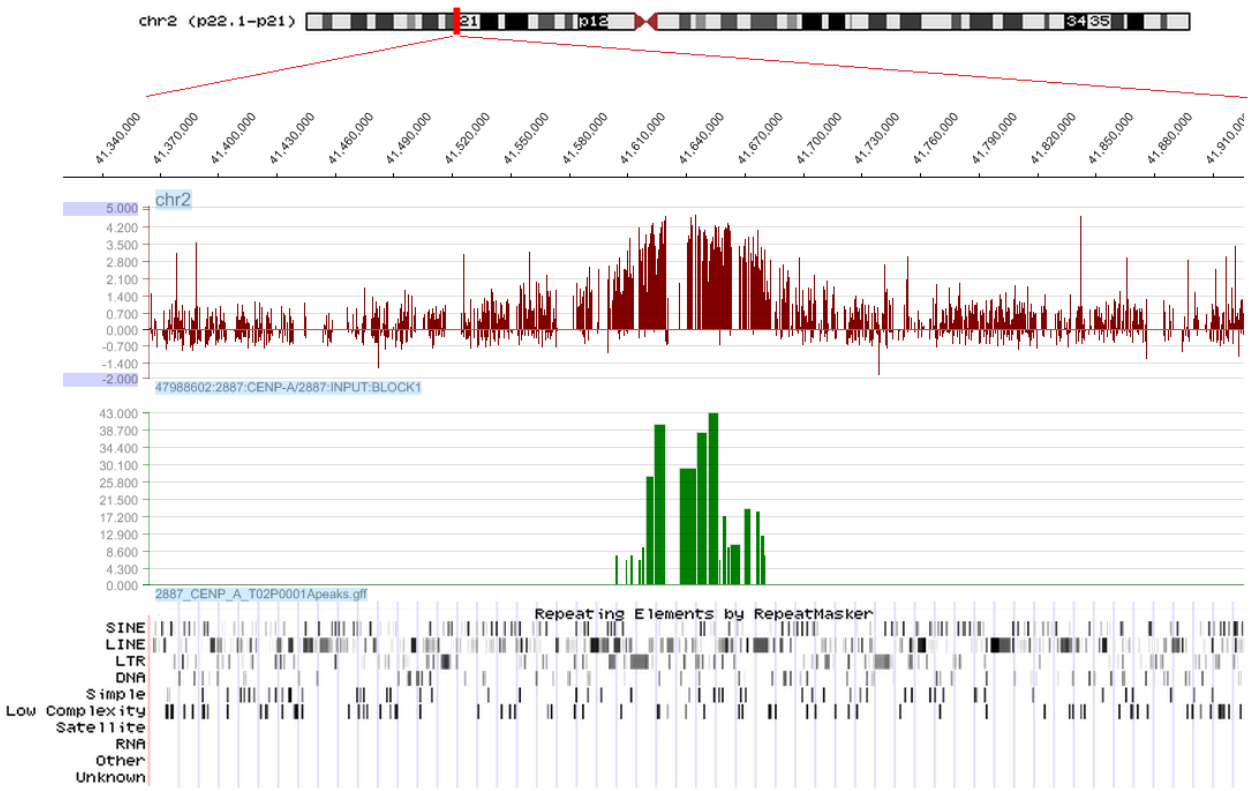


Fig. 2.6 Partial view of the ChIP-on-chip 2887 analysis data on human chromosome 2. Results in lane 1 are presented as the log2 ratio between the hybridization signal obtained with immunoprecipitated DNA using anti-CENP-A antibodies and that given by the input DNA sample. The *X* axis shows the genomic position of each oligo.In lane 2 the DNA binding peak (chr241.603.764-41.680.907) identified by the statistical model TAMALPAIS (98th percentile threshold and $P < 0.0001$). Below the RepeatMasker analysis of the interspersed repetitive DNA elements as shown by the UCSC Genome Browser.

3. CHROMATIN ASSOCIATED RNAs (CARs)

To date, there has been no thorough investigation addressing the identity of the chromatin-associated RNAs (CARs) on a global scale. This prompted us to develop a technique with the aim to identify CARs in a genome-wide approach using high-throughput genomic platforms. This part of the project was made by me in collaboration with the Dr Nick Gilbert, Cancer Research Institute, University of Edinburgh.

In this study, CARs were purified from interphase or mitotic blocked human fibrosarcoma cell lines (HT1080) by isolating soluble chromatin by digestion or sonication followed by

separation of different length chromatin fragments and RNA isolation and its high-throughput sequencing on the Illumina platform.

The sequencing of CARs revealed an association of many intronic and intergenic transcripts with chromatin, indicating that they may have structural and functional roles in chromatin organization through direct or indirect interactions with chromatin.

3.1 ISOLATION OF TIGHTLY CHROMATIN ASSOCIATED RNAs

3.1.1 CARs FROM INTERPHASIC HUMAN CELLS

Chromatin fiber structures can be separated by sucrose gradient sedimentation (Gilbert and Allan, 2001; Gilbert et al., 2004; Kimura et al., 1983). Sedimentation rate is determined by the mass (DNA length and protein composition) and hydrodynamic shape (conformation) of the fiber. A given length of DNA will sediment faster than bulk chromatin if it is packaged into a more compact regular chromatin structure (Gilbert and Allan, 2001) and slower if it is packaged in fibers whose structure is interrupted. So open chromatin and compact chromatin of the same length could be separated between the bulk size of the genome after sucrose gradient sedimentation of chromatin fibers.

We chose this technique to fractionate chromatin fibers in order to isolate the RNA fraction of interphase cells for the further deep sequencing.

A human fibrosarcoma cell line (HT1080) was first treat with alpha-amanitin. The block of the polymerase II prevent the isolation of RNA that are bound to chromatin because of their highly transcription rate and not because of a real structural and/or functional role in chromatin compaction. Nuclei were than digested with an engineered form of the murine DNA fragmentation factor (mDFF-T) (Gilbert and Allan, 2001; Xiao et al., 2007). We decided to use this apoptotic nuclease instead of micrococcal nuclease because of its intrinsic properties that enables its usage into chromatin structure studies. 1) DFF-T shows exquisite selectivity for linker region cleavage, 2) it exclusive generates double-stranded breaks, 3) it lacks exonuclease activity. The well documented RNase activity of micrococcal nuclease avoids its usage for the aim of this study. The recombinant form of the protein is expressed in S. cerevisiae in an engineered inactive form. The cleavage with TEV-protease highly activates the DFF endonuclease activity.

Because of the solenoid model of the 30nm fiber that compose chromatin, to analyze the structure that is propagated over extensive regions (50-100 nucleosomes), we optimized digestion condition to have average fragments length between 10-20 kb (Fig. 3.1 A)

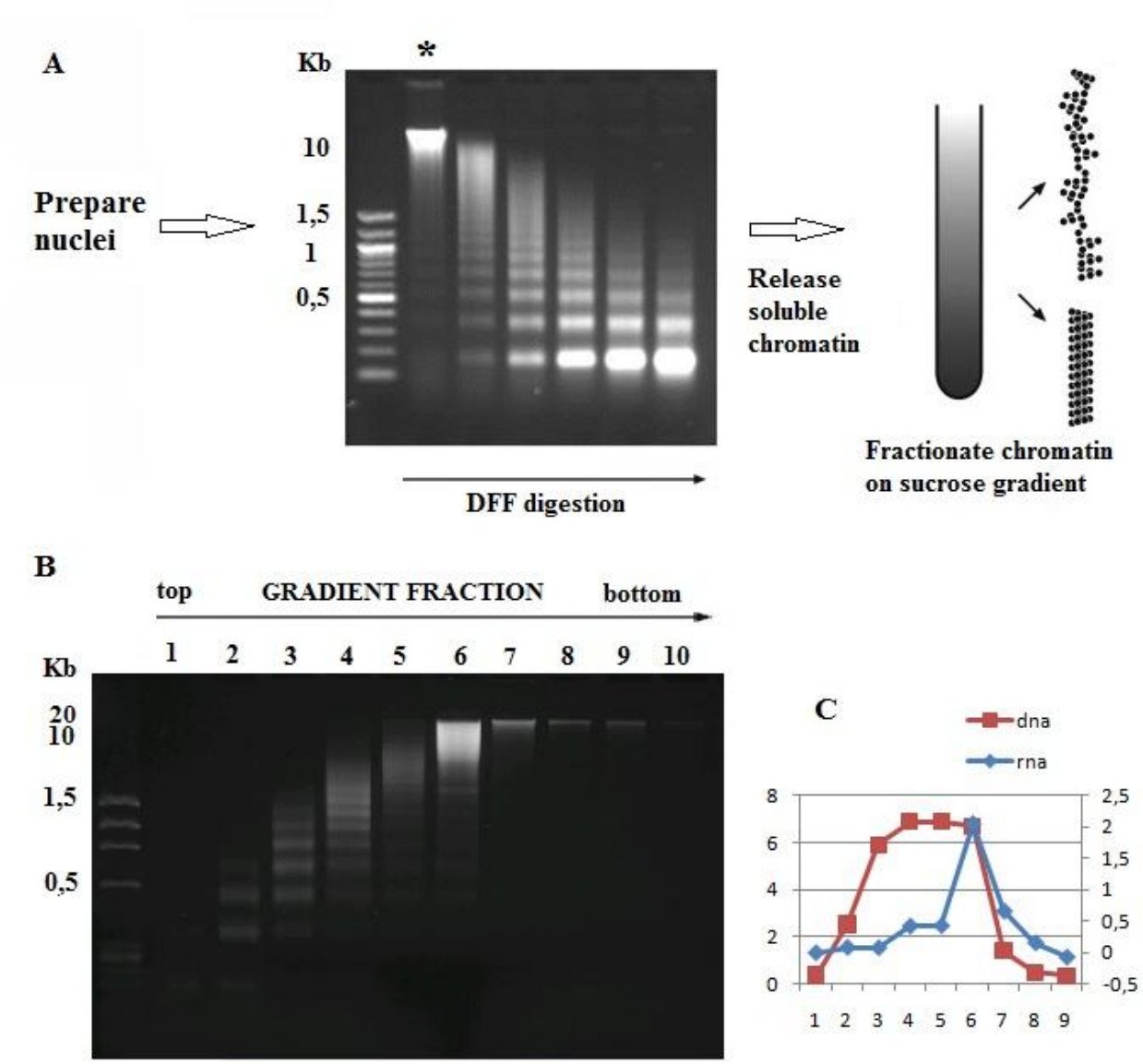


Fig 3.1 Sucrose Gradient Fractionation of Human Chromatin(A) DFF digestion of nuclei was used to produce chromatin fragments with a size range of ~10 kb. The soluble chromatin from the digestion marked by an asterisk, was run on a 50-20-10% sucrose step gradient. For two chromatin fragments of equal length (kb) the more open/disordered fragment (top) will sediment slower than the more compact/rigid one (bottom). (B) The gradient was fractionated from top to bottom and the DNA purified from each fraction was examined by agarose gel electrophoresis. Fraction 6 and 7 were collected for the CARs isolation.(C) Graphic of the DNA-RNA content for each fraction obtained from scintillation quantification. On y-axis is the quantity in µg (the scale on the left is for DNA and the right for the RNA), red line for DNA and blue line for RNA. Fractions 6 and 7 were collected and pulled together for CARs isolation.

In previous studies performed by Dr. Gilbert to ensure that we were not preferentially releasing particular parts of the genome before loading onto the gradient, total genomic DNA and DNA from the digested soluble chromatin were hybridized by FISH to metaphase

chromosomes. Without CotI suppression both hybridize strongly to centromeric and juxtacentromeric heterochromatin, showing that these compact region were not refractory to digestion and solubilization. With the CotI treatment the hybridization signals along the euchromatic part of chromosome arms were indistinguishable (Gilbert et al., 2004).

RNA quality of digested chromatin was checked on denaturating agarose/formaldehyde gel and then chromatin was sedimented through a step sucrose gradient (10%, 20% and 50%). The gradient was fractionated from top to bottom so that fraction contain chromatin fibers with progressively increased sedimentation rate. DNA content of each step was quantified and resolved on agarose gel electophoresis to detected the average length of fragments (Fig 3.1 B). Each fraction appeared like smears of the ethidium bromide signal: the peak corresponded to sequences that were packaged within fibers characteristic of the bulk genome; shorter and longer fragments consisted of sequences packaged in fiber that are respectively more or less compact than those of the bulk genome. RNA from the sample loaded on sucrose gradient were checked for quality on a denaturating agarose 1% gel.

The fractions containing the peak of ethidium bromide corresponding to the molecular weight of interest (10-20 kb) were collected and pulled together and added of RNase inhibitor (Fig 3.1 B, fraction 6 and 7). Chromatin was then biotinilated in both DNA fragment extremities and isolated from the unlabeled fraction on streptavidine magnetic beads. Several higher salt concentration washes were done in order to remove unspecific RNA bound to chromatin. High salt concentration affects chromatin structure: DNA dissociates from the intact core octamer at 2M. Comparing literature and our previous works, and because of the limited knowledge about the strength of the CARs interaction, we decided to do not use too stringent salt condition. This in order to minimal affect the structural organization of chromatin fibers but inducing the removal of weakly non specifically bound RNAs. Our knowledge brought us to use a maximum of 300mM salt concentration washes. At this concentration the two H2A-H2B histone dimers weakly interact with the tetramer but still remain bound to DNA (Martinson and True, 1979) so that chromatin structure is not affected and the conspicuous of the aspecific weakly bound RNAs is removed from the chromatin fiber.

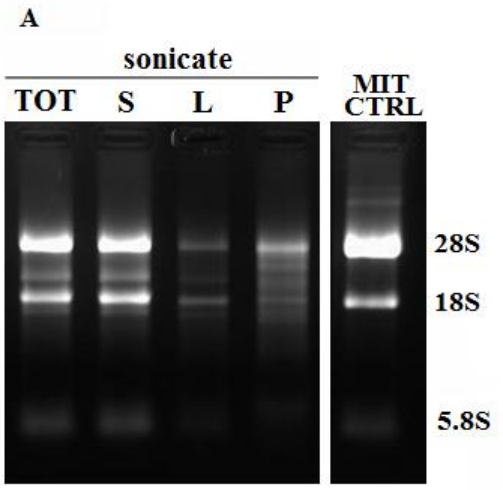
After the washes step RNAs remained bound to chromatin (CARs) were extract with tri-reagent, treated with DNase and then with ribominus Human/mouse module kit (invitrogen) to deplete samples from the abundant and aspecific ribosomal RNA and deep sequenced with the Illumina Solexa GAII**x** technology. The protocol we set up is in the section 5.1 of Mat. & Met. The name of the sample obtained in this way was HT1080-6-7.

Total RNA of interphasic HT1080 cells was extracted with tri-reagent, treated with DNase and then with ribominus Human/mouse module kit (invitrogen) in order to purify from the ribosomal RNAs because of their highly transcription rate that determine their abundance in cell extract. This sample was called HT1080-tot was deep sequenced. This RNA population was used like reference for the interphase6-7 sample. We than compared bioinformatically each of our CARs arising from the experimental sample to that of the total and concluded which of that CARs are enriched over the total population (§3.2.1).

3.1.2 THE MITOTIC CARs

To prepare the CARs tightly and specifically bound to the mitotic chromatin we optimized the protocol described in section 5.4 of Mat. & Met.

In summary human fibrosarcoma cell line (HT1080) was colcemid-blocked in mitosis (efficiency of 70%). Mitotic cells were harvested, nuclei from them were isolated and chromatin was fragmented by sonication (to have 10-20Kb molecular weight chromatin). The precipitated fraction of chromatin was positive for the detection of the H3Ser10P mitotic marker. It was washed with higher salt concentration step. We decided to collect two different samples, one after 160mM salt concentration washes (TEEP-160) and the other one at 80mM salt concentration (TEEP-80). After the washing step RNAs remained bound to chromatin (CARs) were extract with tri-reagent, treated with ribominus Human/mouse module kit (invitrogen) to deplete from the abundant ribosomal RNA and deep sequenced with the Illumina Solexa GAII**x** technology .



3.3 Mitotic CARs. RNA quality was checked on 1% agarose denaturing gel. TOT is total RNA from nuclei, S was after the sonication, L was RNA washed away, P corresponds to mitotic CARs.

Total RNA of HT1080 mitotic cells was specifically extracted with tri-reagent after 16hs of colcemid treatment of HT1080 cells and then DNAse treated and purified from the ribosomal RNA population with Ribominus Human/mouse module kit (invitrogen). This RNA was then deeply sequenced in order to use it like a reference for the bioinformatic analysis for the mitotic experimental samples (TEEP-80 and TEEP-160).

3.1.3 CENTROMERIC SPECIFIC CARs

To the purpose of isolating CARs specifically associated to centromeric chromatin we developed two different methodic (see § 5.3 of Mat. & Met. to see protocols). Both of them are based on the specific localization of the CENP-A protein to functional centromeres. Thus, the immunoprecipitation of the DNA bound to this protein allows the isolation of centromeric chromatin.

In the first method we optimized the ChIP protocol in order to purify RNA (instead of DNA) and deep sequencing it in a genome wide view. To do it we crosslinked with formaldehyde a population of interphasic HT1080 cell line, then nuclei were extracted and chromatin was sonicated randomly to obtain fragments with an average sharing of 500 bp (Fig 3.3). The following step is the immunoprecipitation with polyclonal antibody that are specifically directed against the human CENP-A protein (Trazzi et al 2009) or rabbit IgG as negative control. This enabled us to purify the centromeric chromatin. One sample was taken in order to represent the total chromatin before the immune selection (INPUT). Immunoprecipitated samples were purified on ProtA magnetic beads, crosslinking was reverted with proteinase K treatment and the RNA was extracted with tri-reagent and sent for sequencing.

To test the efficiency of the immunoprecipitation, the DNA alpha-sat enrichment of the CENP-A immunoprecipitated DNA over both the INPUT and IgG immunoprecipitated DNA was evaluated by Real Time PCR with specific a primer set (see mat met). The alpha satellite DNA was 1,9 times enriched in the CENP-A immunoprecipitated.

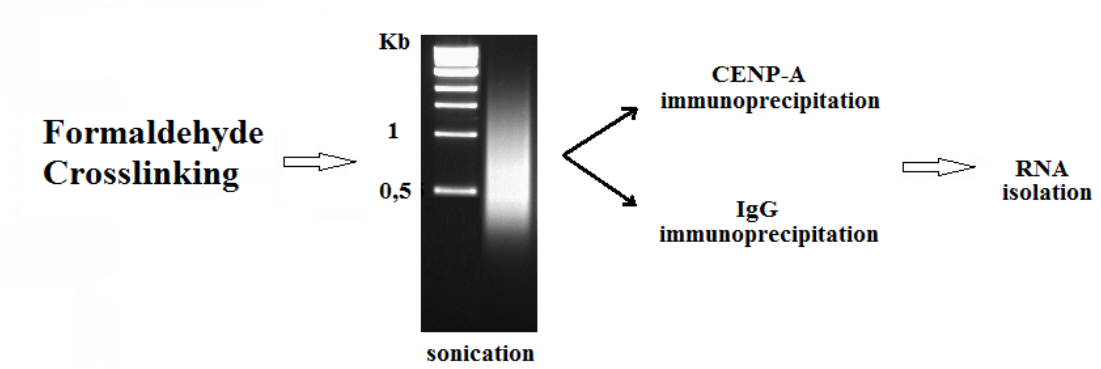


Fig 3.3 Scheme for the centromeric CARs isolation. HT1080 were crosslinked, the chromatin from nuclei sonicated to give an average bulk of 500bp, then follow the immunoprecipitation with anti-rabbit antibody against CENP-A or IgG as a negative control. The last step is the isolation of that RNA.

The second method (Mat. & Met. § 5.3.2) we developed in order to purify centromeric chromatin was without crosslinking cells. Nuclei of interphasic HT1080 cell line were extract and digested with the activated mDFF enzyme. Soluble chromatin was then sedimented trough a step sucrose gradient (10%-20%-50%). The gradient was fractionated from top to bottom and checked for the DNA length content on agarose gel in order to choose fractions of higher molecular weight (Fig 3.4) for then be immunoprecipitated with anti-CENPA antibody, that specifically bound centromeric region, and with rabbit IgG as a negative control. Immunoprecipitated samples were then isolated with protA magnetic beads and RNA was extracted with trireagent, DNase treated and sent to be sequence.
To test the efficiency of the immunoprecipitation, the DNA alpha-sat enrichment of the CENP-A immunoprecipitated DNA over both the INPUT and IgG immunoprecipitated DNA was estimated by Real Time PCR with a specific primer set.. The alpha satellite DNA was 1,2 times enriched in the CENP-A immunoprecipitated.

Unfortunately the two centromeric CARs samples here described have not been sequenced in time to be part of this thesis.

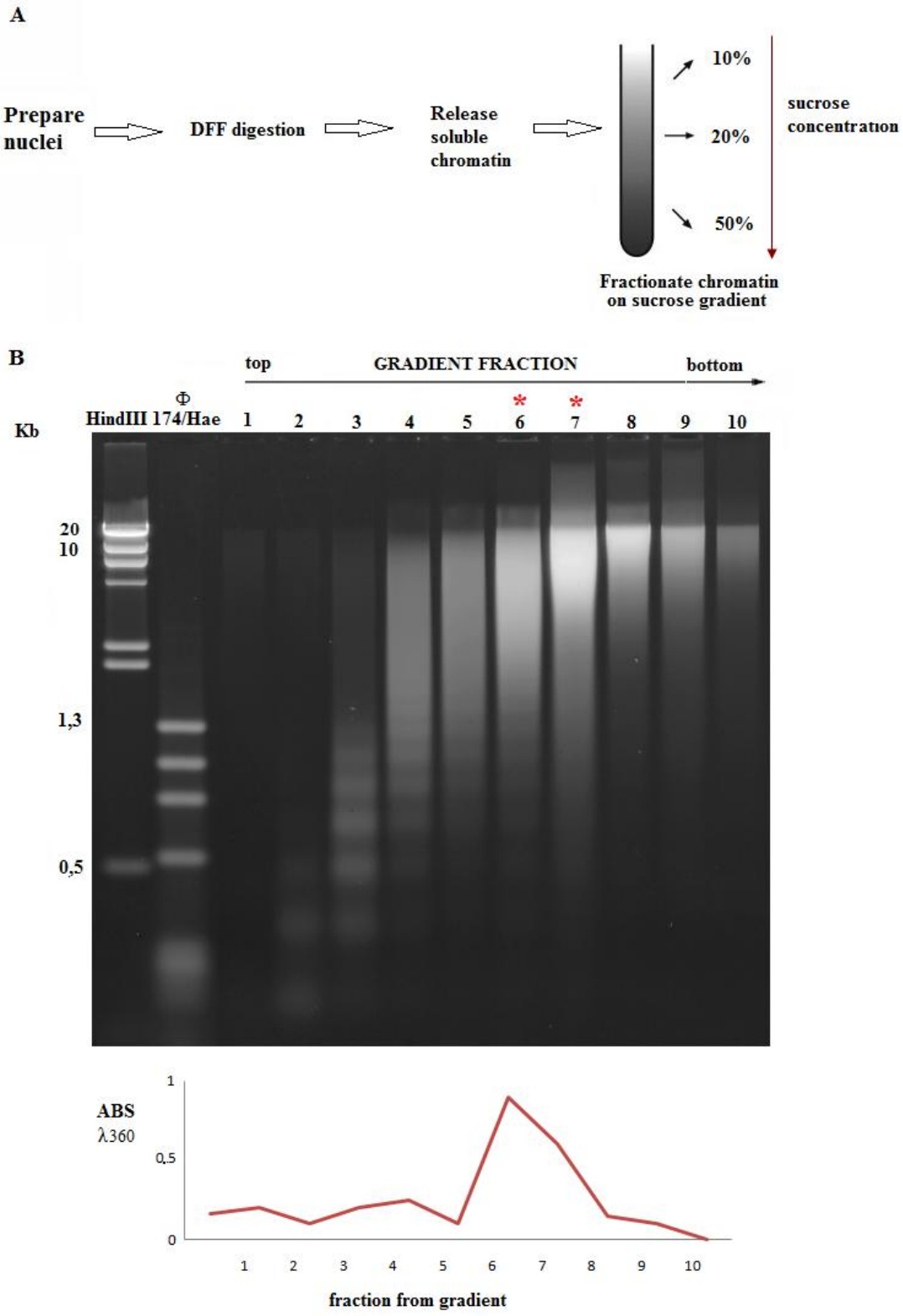


Fig. 3.4 Sucrose Gradient Fractionation of Human Chromatin (A) The DFF digested soluble chromatin was run on a 50-20-10% sucrose step gradient. (B) The gradient was fractionated from top to bottom and the DNA purified from each fraction examined by agarose gel electrophoresis. Fraction 6 and 7 were collected for the immunoprecipitation with anti-CENP-A or anti-rabbit IgG for CARs isolation. The graph shows the OD at λ=360nm to evaluate the DNA-RNA content for each fraction of the gradient.

3.2 HIGH-THROUGHPUT SEQUENCING OF RNA

The Solexa GAIIx is a second-generation sequencing technology. It's a so called Polymerase-based sequence-by-synthesis reaction which uses a small 'flow cell' to immobilize, amplify and sequence up to 250 million molecules at once. Single-end fragments were sequenced. See the section §5.5 of Material & methods to know more about Solexa GAIIx technology.

We sequenced 5 total CARs samples: HT1080-tot, HT1080-6-7 (both explained in § 3.1.1), Mitotic-tot, TEEP-80, TEEP-160 (see § 3.1.2). Each file resulting from a sequencing process contains over millions of 50bp reads that should be analyze and aligne over the reference genome. In our case the total number of hits from each sample was reported in the table 3.1. The genome taken as reference was the human genome assembled in Febrary 2009 (GRCh37/hg19).

Table 3.1

	Total reads
HT1080-tot	3751396
HT1080-6-7	2156950
Mitotic-tot	10752246
TEEP- 160	3477703
TEEP-80	1454782

To analyze the enrichment between a sample and its corresponding control (Mitotic-tot was used like a reference for the mitotic CARs while the HT1080-tot was used for the interphasic CARs), first the number of reads mapping to each known transcript of the human genome hg19 was counted with Bowtie. Reads mapped on the known rRNA genes where excluded. Many of the obtained reads where on non coding region of the human genome, so a deeply analysis of that regions was needed. A window of 80bp length (with a 20bp slide) was run across the non-coding regions of the human genome to find peaks of expression and concatenated neighboring expressed windows to define putative unknown transcriptional units.

Then the files containing mapped reads from RNA-seq data with the gene annotation of the human genome and the RPM (Reads Per Million mapped reads) of both the novel units and the Ensembl Units were upload on DEG-seq to work out the differential sequence enrichment of the regions and to calculate p-values. The RPM measure of read density reflects the molar concentration of a transcript in the starting sample by normalizing for the total read number in the measurement. This facilitates transparent comparison of transcript levels both within and between samples.

Finally we used Galaxy and UCSC genome browser to upload the bigWig files in order to display sequencing data aligned on human genome hg19(Feb2009) release.

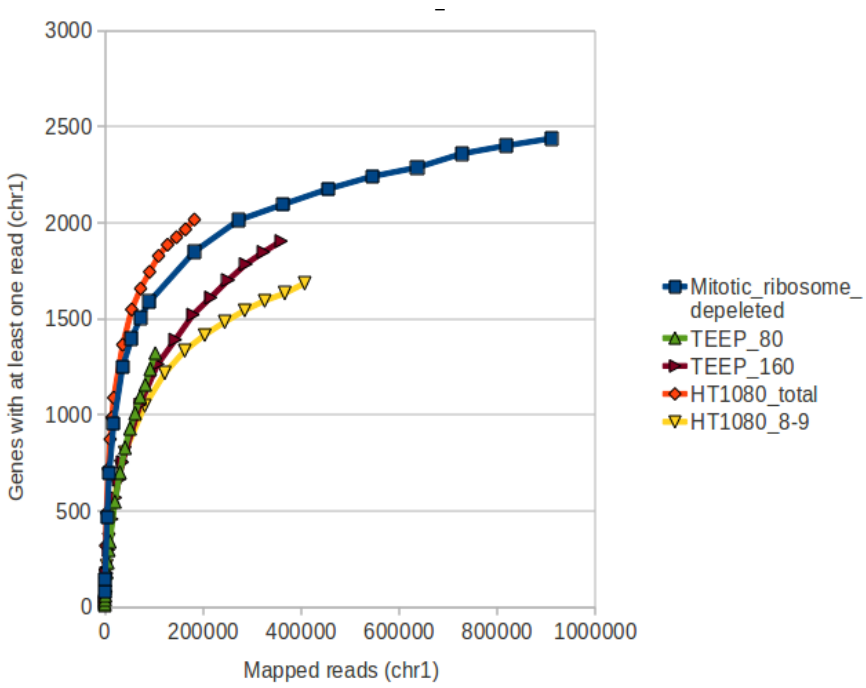


Figura 3.5 Coverage plot on chr1. It show the total number of sequence reads that mapped onto known genes encoded in chr1 at least one time.

A coverage plot of the mapped reads versus the known genes with at list one read in each sample were obtained for each chromosome. Each of 5 samples were analyzed in order to determine the quality of the sequencing. The TEEP80 sample (green line in the figure 3.5) does not give a good coverage of the genome and could not be useful for the detection of novel CARs but could be good in order to validate data coming out from TEEP160 sample. All the other tested samples gave us good coverage plot, so they have been further analyzed.

3.2.1 DATA ANALYSIS

The total RNA sequenced from both interphasic either mitotic HT1080 cells showed a good coverage of the human genome. The reads mostly mapped on exons of known genes or in silico predicted gene (the 80% for the mitotic and the 78% for the interphasic RNA). The percentage of exon's mapping reads conspicuously decrease in the CARs sample: 35 % for the interphasic HT1080-6-7 and 31% for the mitotic teep-160.

Highly expressed genes were chosen as an experimental control of the data quality. The data were uploaded on Galaxy and displayed by UCSC genome browser. Peaks correspond to enrichments for that sequence in the sample analyzed. Figure 3.6 show the coverage of the ACTB gene (chr7:5,533,305-5,536,758), here RPM (reads per million) of the sequenced reads

of each experimental samples was shown as a function of the genomic distribution. This gene encodes one of six different actin proteins. Actin B is a major constituent of the contractile apparatus and one of the two non muscle cytoskeletal actins. Notably in the totals RNA samples (both interphasic either mitotic, respectively graph in lanes 1 and 2 in figure 3.6) the expression is high. Most of the reads map spanning over the exons of the gene. This specific distribution represent the mature form of the RNA transcript of the ACTB gene. Despite the high expression of this genes, no reads map over ACTB were detected in the CARs samples (graph in lanes 3,4,5).

Similar consideration could be done for other 2 metabolic gene: GAPDH gene (glyceraldehyde-3-phosphate dehydrogenase; genomic position chr12:6,513,918-6,517,797) in the figure 3.7 and the glycolitic ALDOA gene (fructose-bisphosphate aldolase; chr16:29,983,101-29,989,236) showed in the figure 3.8. In both genes all the RPM peaks map over exons in the two controls (lane 1 and 3 of each figures) Any significantly read enrichment corresponding to the genes are found in none of the CARs samples.

This to confirm the fidelity of the technique: transcripts are not pulled down because of their abundance in nuclei but because of their specific association to the chromatin.

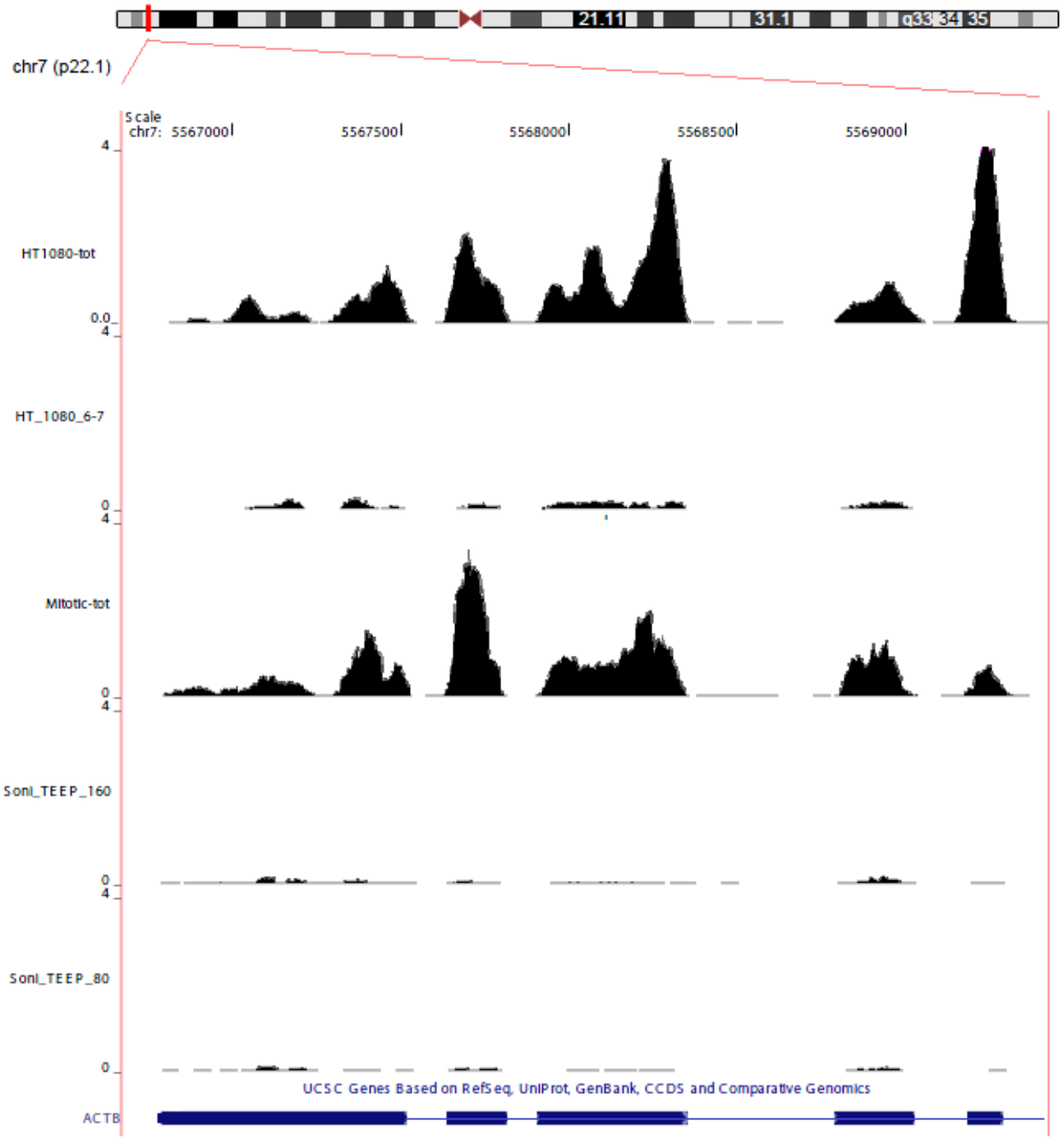


Fig. 3.6 Genomic distribution of the RNA reads for each sample sequenced. The gene displayed is ACTB (chr7:5,533,305-5,536,758). UCSC genome browser is used. The graphs show the RPM (reads per million) as a function of the human genomic coordinate. First lane shows the HT1080-tot sample that is the total RNA extracted from HT1080 interphase cells, this is the reference for interphasic CARs. Second lane is the HT1080-6-7 sample, the interphasic CARs. Third lane is the RNA distribution of the Mitotic-tot sample, the reference for the mitotic CARs. Fourth and fifth lanes show the two mitotic CARs samples, respectively TEEP-160 sample and TEEP-80. A schematic representation of the ACTB gene is rapresented below the graphs. In the two references (lanes 1 and 3) RNA of ACTB is highly represented, notably peaks map in exons (blue block of the gene scheme). ACTB RNA transcript is not represented in any of the CARs samples (lanes 2, 4,5).

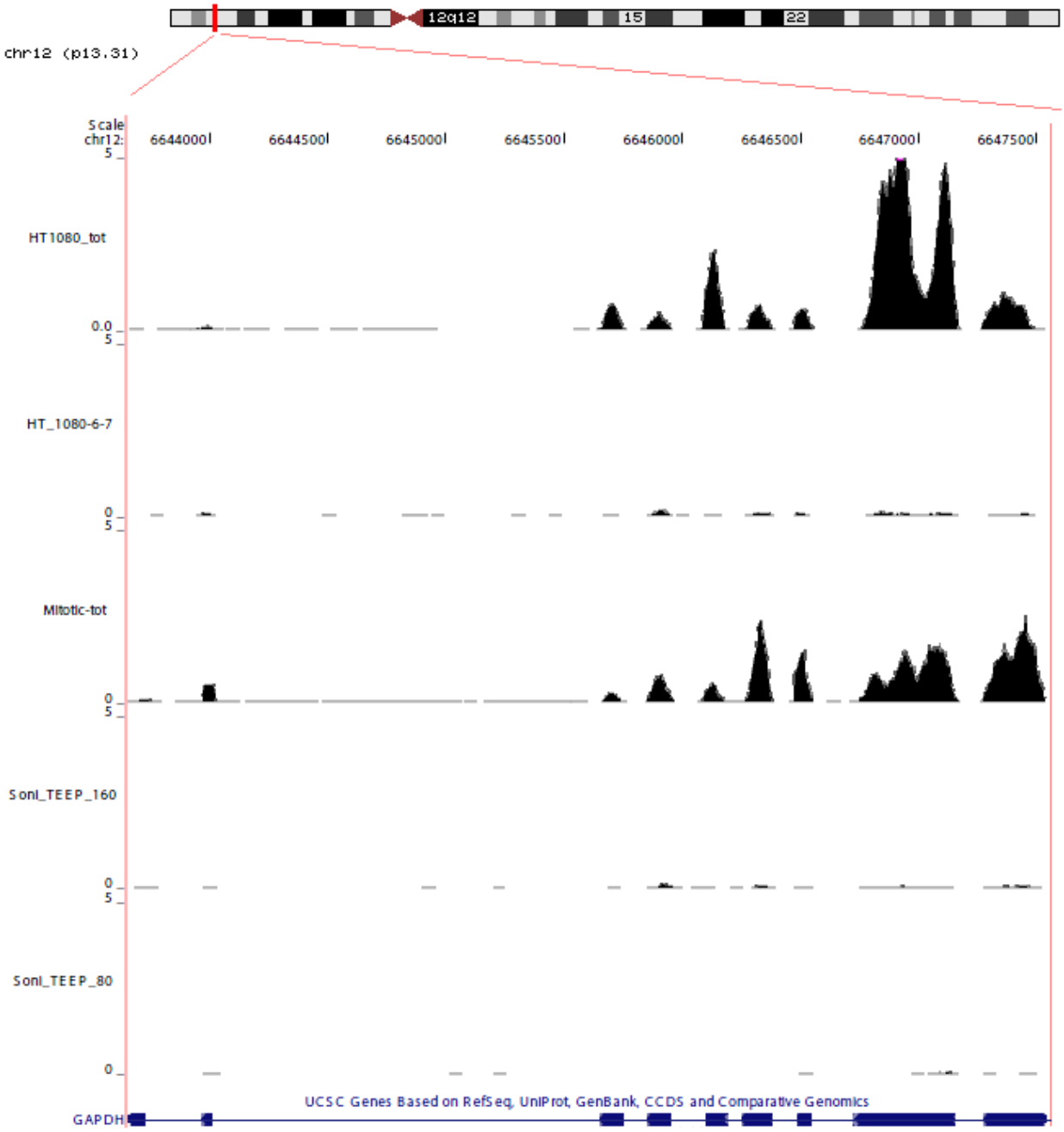


Fig. 3.7 Genomic distribution of the RNA reads for each sample sequenced. The gene displayed is GAPDH (chr12:6,513,918-6,517,797). UCSC genome browser is used. The graphs show the RPM (reads per million) as a function of the human genomic coordinate. First lane shows the HT1080-tot sample that is the total RNA extracted from HT1080 interphase cells, this is the reference for interphasic CARs. Second lane is the HT1080-6-7 sample, the interphasic CARs. Third lane is the RNA distribution of the Mitotic-tot sample, the reference for the mitotic CARs. Fourth and fifth lane show the two mitotic CARs samples, respectively TEEP-160 sample and TEEP-80. A schematic representation of the GAPDH gene is represented below the distribution graphs. In the two samples taken as references (lanes 1 and 3) RNA of GAPDH is highly represented, notably peaks map in exons (brown blocks of the schematic gene representation). GAPDH RNA transcript is not represented in any of the CARs samples (lanes 2, 4,5).

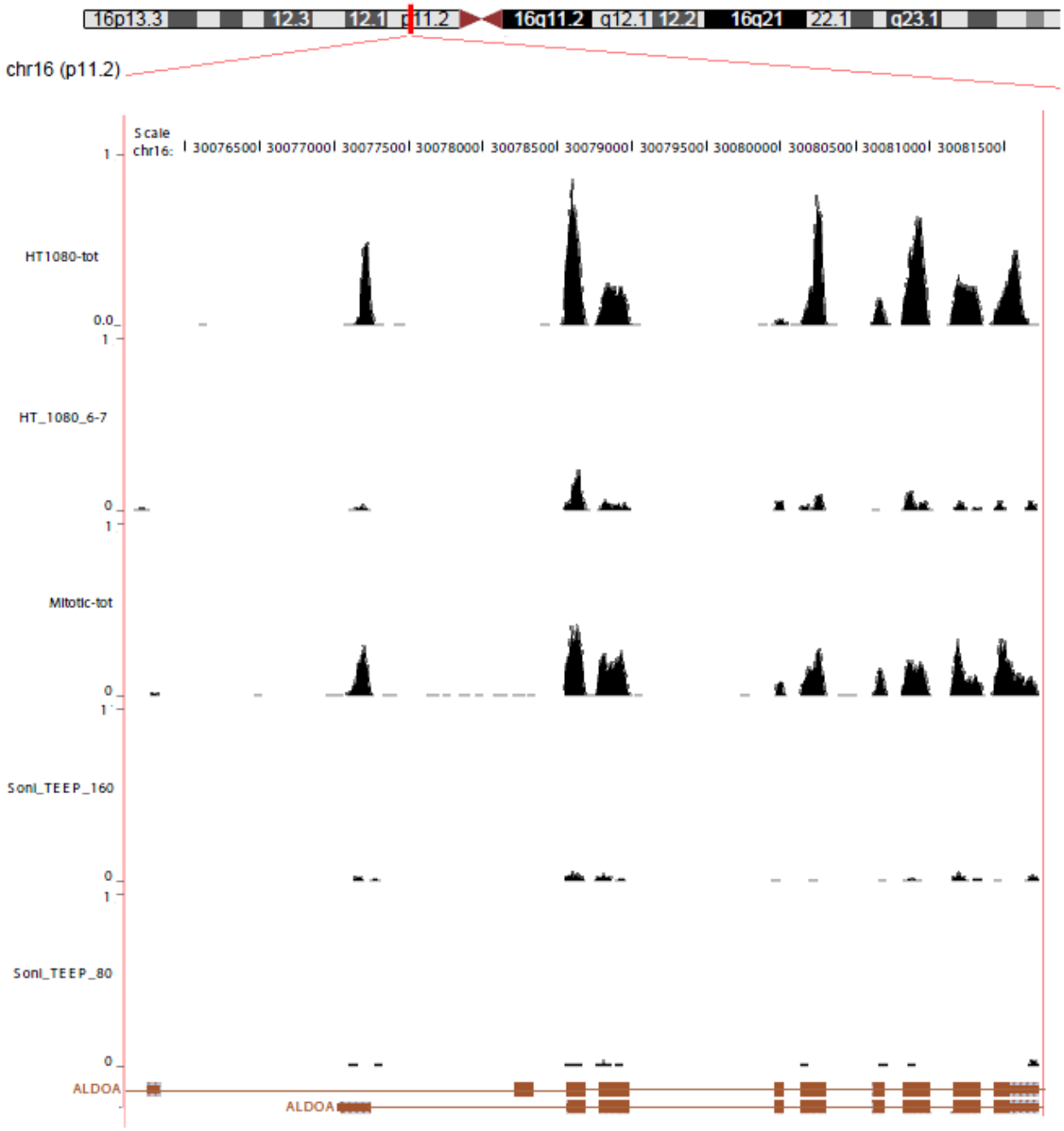


Fig. 3.8 Genomic distribution of the RNA reads for each sample sequenced. The gene displayed is ALDOA (chr16:29,983,101-29,989,236). UCSC genome browser is used. The graphs show the RPM (reads per million) as a function of the human genomic coordinate. First lane shows the HT1080-tot sample that is the total RNA extracted from HT1080 interphase cells, this is the reference for interphasic CARs. Second lane is the HT1080-6-7 sample, the interphasic CARs. Third lane is the RNA distribution of the Mitotic-tot sample, the reference for the mitotic CARs. Fourth and fifth lane show the two mitotic CARs samples, respectively TEEP-160 sample and TEEP-80. A schematic representation of the two transcription variants of ALDOA gene is represented below the graphs. In the two references (lanes 1 and 3) RNA of ALDOA is highly represented, notably peaks map in exons (brown blocks of the schematic gene representation). ALDOA RNA transcripts is not represented in any of the CARs samples (lanes 2, 4,5).

As just introduced the notably mark of all the three CARs samples, both the intherphasic (interphasic 6-7) either the mitotic (Teep-160 and Teep-80) was the conspicuous enrichment for intergenic regions of human genome and intronic portions of the protein coding genes.

When we first analyzed the data we focalized our interest in some long non coding RNA that were highly enriched in both interphasic and mitotic CARs when compared to the respective two biological controls of reference.

GAS5 (chr1:172,099,662-172,103,748) and SNGH1 (chr11:62,376,036-62,379,933) are two examples of RNA found to be significantly ($p < 0.001$) enriched in CARs samples if compared to the respective references. Notably both of them are classified as non-protein-coding RNA which hosts snoRNAs (fig. 3.9 and 3.10).

GAS5 was already known in literature (Kino et al., 2010; Mourtada-Maarabouni et al., 2008; Mourtada-Maarabouni et al., 2009). It's ubiquitously expressed in human and mouse in many alternate splicing but interestingly the putative ORF is small and poorly conserved during evolution. Literature suggests that any important biological activity must be mediated throught introns, which encode multiple small nucleolar RNAs (snoRNA). The overexpression of certain GAS5 trancripts induces growth arrest and apoptosis in several mammalian cell lines and slowing of the cell cycle with an increase in the proportion of cells in G1. Its expression is also downregulated in breast cancer cells and RNA interference experiments show that it is both necessary either sufficient for the normal growth arrest of T-cell line. It has been proposed as an oncogenic gene.

Interestingly, as shown in the figure 3.9 and 3.10 our CARs form peak maps over the introns of the differently spliced gene instead of inside the putative ORF. Moreover the enrichment, in both SNGH1 and GAS5, is in correspondence of the small snoRNA encoded from introns.

snoRNA hosted by the same gene are not equally enriched in the samples: focalizing the attention on the GAS5 gene we can see that SNORD76 and SNORD80 are more enriched in the mitotic sample (fig. 3.9; 3.10 lane 4 and 5) than in the interphasic one (lane 2), on the contrary SNORD44 is higher in the interphase then in mitosis. Probably each of the snoRNA is expressed like independently transcript differently regulated and with different rule. Their hypothesized chromatin association should be cell cycle dependent.

A further example of different enrichment for snoRNA hosted by the same gene is given by the RPS10 gene (Fig. 3.11). RPS10 gene is not enriched in our samples compared to the references, while some snoRNA hosted in its intron figure like CAR from our analysis. SNORA33 is found to be a mitotic CAR, while SNORD100 is an interphasic CAR. As you

can see from the picture in the control samples (lanes 1 and 3) the reads map not in introns but in exons.

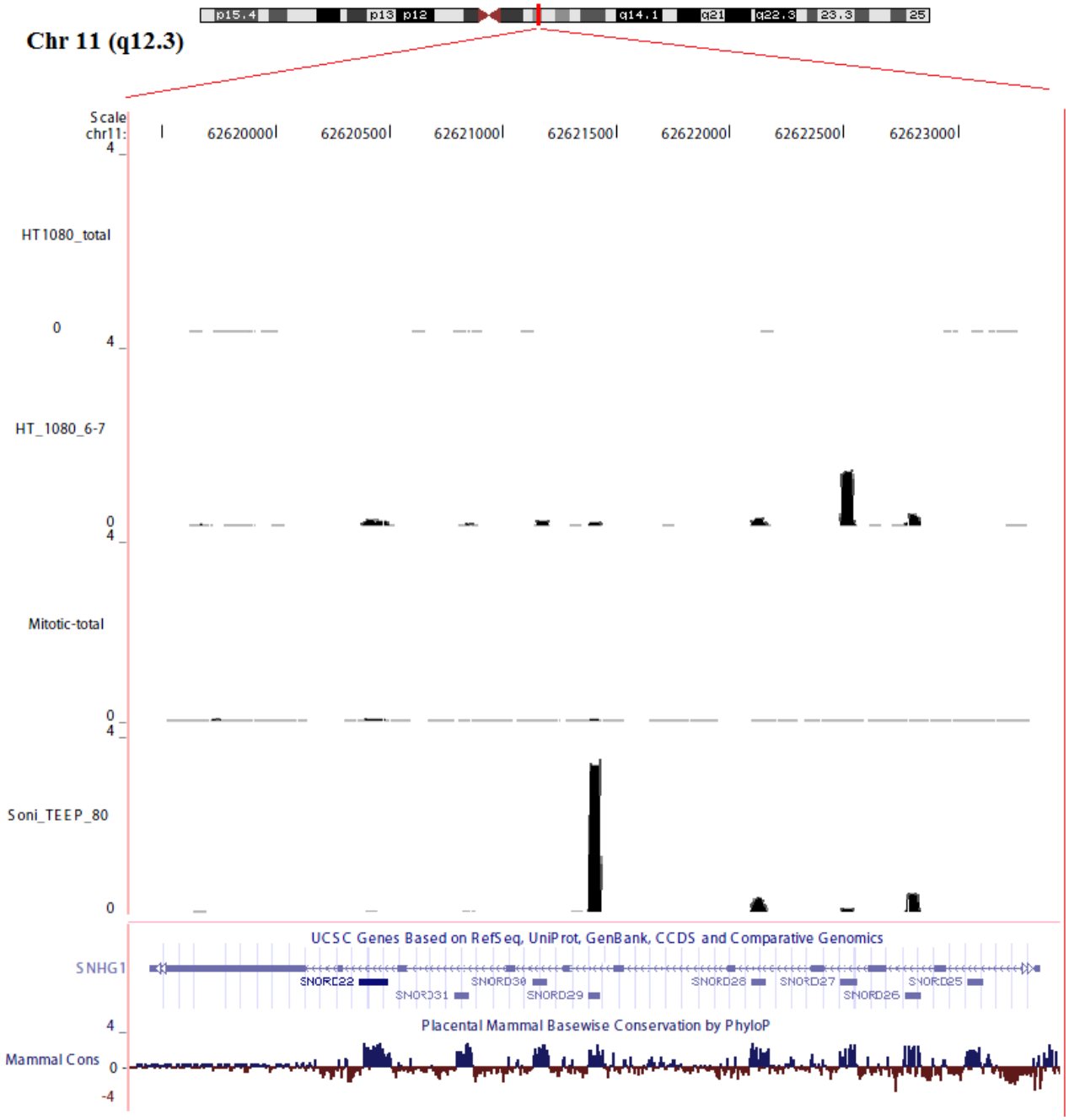


Fig 3.9 Genomic distribution of the RNA reads for each sample sequenced. The gene displayed is SNHG1 (chr11:62,376,036-62,379,933). UCSC genome browser is used. The graphs show the RPM (reads per million) as a function of the human genomic coordinate. First lane show the HT1080-tot sample that is the total RNA extract from HT1080 interphase cells, this is the reference for interphasic CARs. Second lane is the HT1080-6-7 sample, the interphasic CARs. Third lane is the RNA distribution of the Mitotic-tot sample, the reference for the mitotic CARs. Fourth lane shows the mitotic CARs sample TEEP-160. A schematic representation of the SNHG1 gene is represented below the graphs. In the two samples taken as reference (lanes 1 and 3) RNA transcript of SNHG1 is not represented. This gene appear to be enriched in the two CARs samples (lanes 2 and 4). Notably peaks map in snoRNA elements encoded within introns (blue blocks below the schematic gene representation).

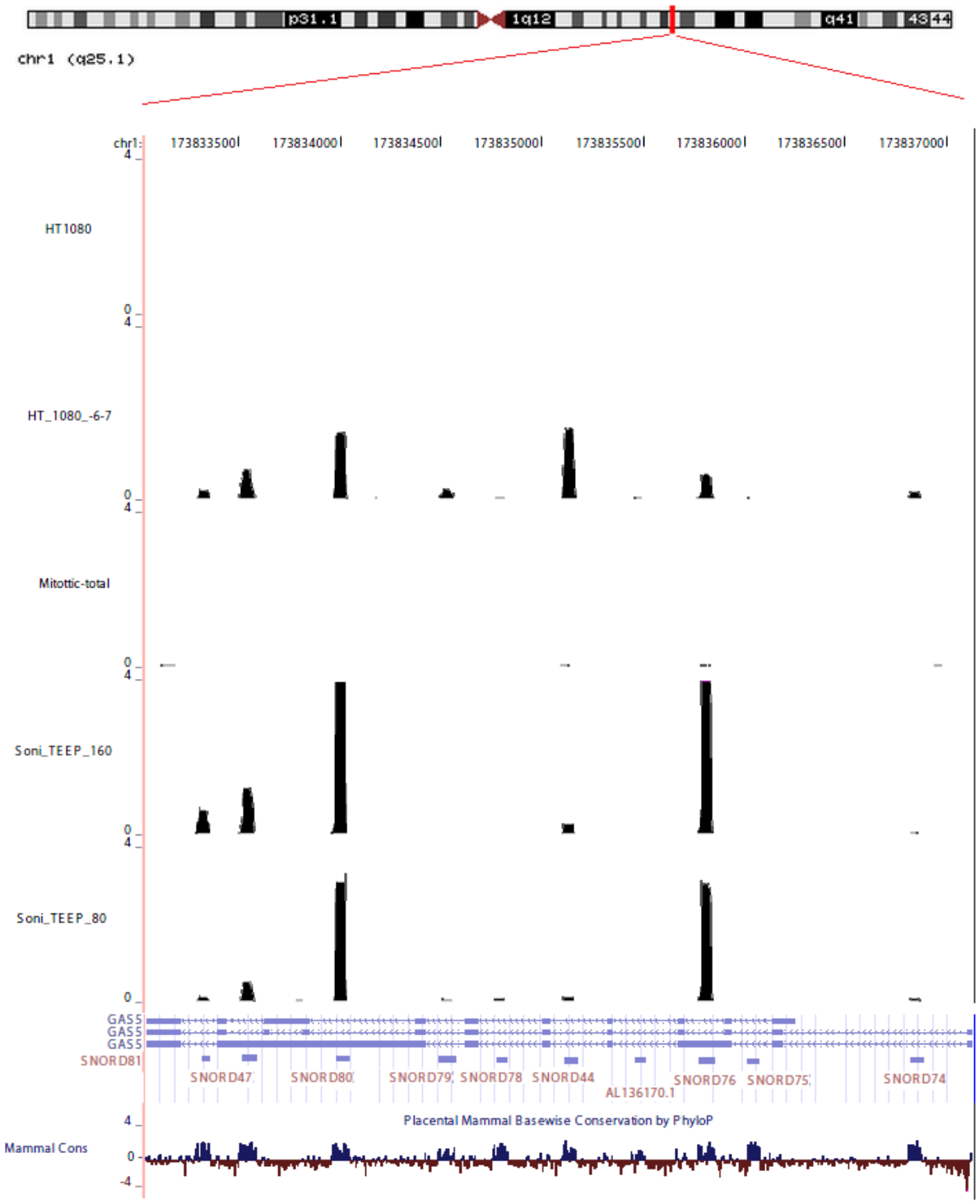


Fig 3.10 Genomic distribution of the RNA reads for each sample sequenced. The gene displayed is GAS5 (chr1:172,099,662-172,103,748). UCSC genome browser is used. The graphs show the RPM (reads per million) as a function of the human genomic coordinate. First lane show the HT1080-tot sample that is the total RNA extract from HT1080 interphase cells, this is the reference for interphasic CARs. Second lane is the HT1080-6-7 sample, the interphasic CARs. Third lane is the RNA distribution of the Mitotic-tot sample, the reference for the mitotic CARs. Fourth and fifth lane show the mitotic CARs samples TEEP-160 and TEEP-80. A schematic representation of the GAS5 gene is represented below the graphs. In the two samples taken as reference (lanes 1 and 3) RNA transcript of SNHG1 is not represented. This gene appear to be enriched in the three CARs samples (lanes 2,3 and 4). Notably peaks map in snoRNA elements encoded within introns (blue blocks below the schematic gene representation).

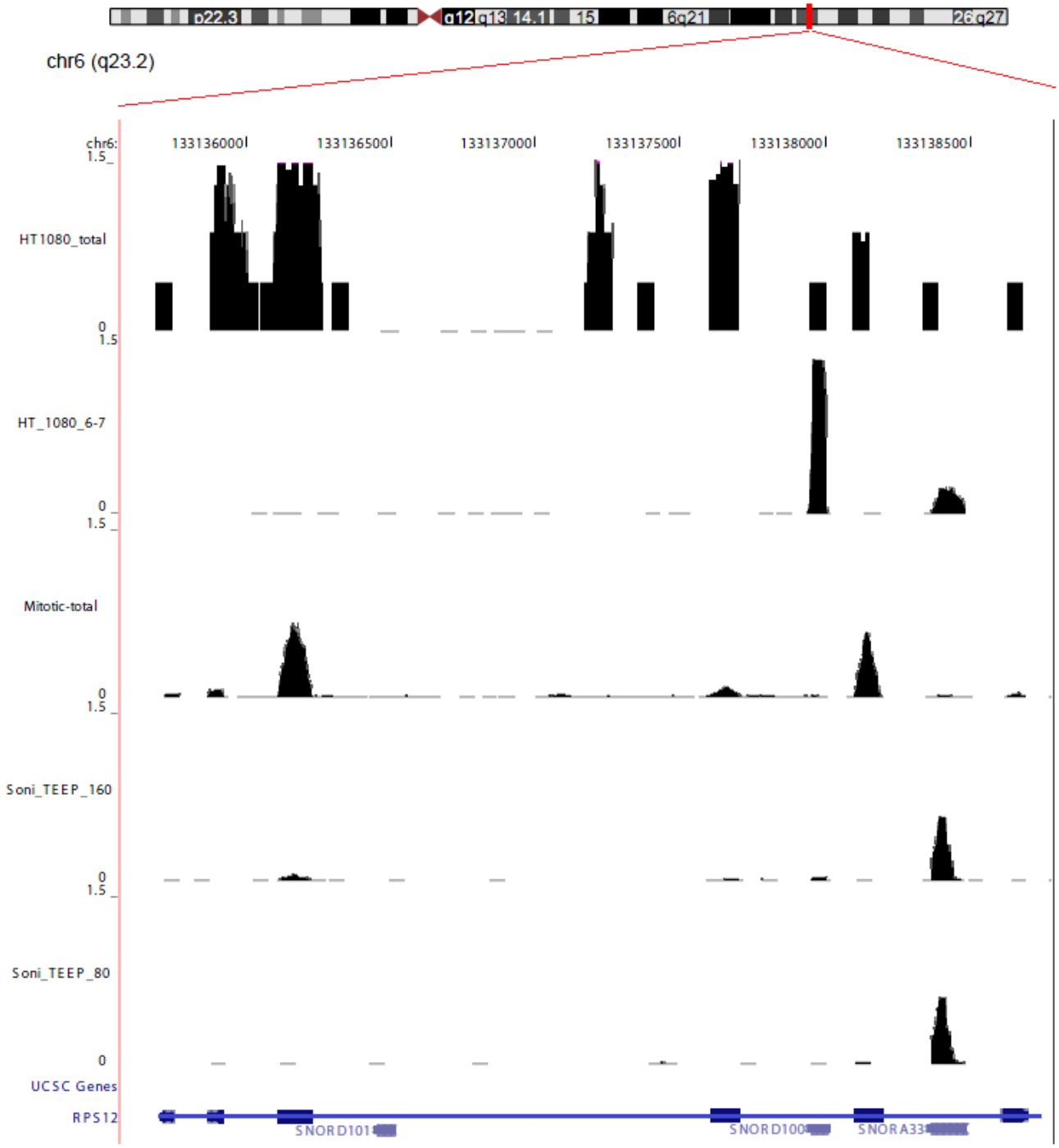


Figure 3.11 Genomic distribution of the RNA reads for each sample sequenced. The gene displayed is RPS12 (chr6:133,177,401-133,180,396). UCSC genome browser is used. The graphs show the RPM (reads per million) as a function of the human genomic coordinate. First lane shows the HT1080-tot sample that is the total RNA extracted from HT1080 interphase cells, this is the reference for interphasic CARs. Second lane is the HT1080-6-7 sample, the interphasic CARs. Third lane is the RNA distribution of the Mitotic-tot sample, the reference for the mitotic CARs. Fourth and fifth lane show the mitotic CARs sample TEEP-160 and TEEP-80. A schematic representation of the GAS5 gene is represented below the graphs. In the two samples taken as reference (lanes 1 and 3) RNA transcript of RPS12 maps in exon. In the three CARs samples (lanes 2,3 and 4) notably peaks map in some of the snoRNA elements encoded within introns (blue blocks below the schematic gene representation).

In light of this observation we found that remarkably a significant subset of the intronic and intergenic CARs belongs to the two major groups of small RNAs: snRNA and small nucleolar RNAs (snoRNAs).

369 and 347 of the 463 total known or predicted snoRNA had at least once map respectively in the mitotic-tot samples and in the HT1080-tot sample. Of that a subgroup of 24 small nucleolar RNAs appear to be highly enriched (more than 3 times) in the TEEP-160 mitotic sample if compared to its total reference ($p<0.001$) (table 4.2). This enriched subgroup increase if we consider the interphasic CARs: 99 snoRNA appear to be enriched when compared to the total. snoRNAs are a class of small RNA molecules (an average of 150 bp) that primarily guide chemical modifications of other RNAs. To date three main classes of snoRNA are known:

1) SNORD: because of the C/D box . Their documented rule is to guide the modification of other RNA;
2) SNORA: because of the H/ACA box. They are associated with pseudouridylation of other RNA;
3) SCARNA (small Cajal body-specific RNAs) with both C/D-H/ACA boxes and that guide the modification of RNA pol II transcribed splicesomal RNAs.

If we divided the snoRNAs found to be in CARs in the three main classes we have that half of the known SCARNA are enriched in the interphasic CARs if compared to its own total while no SCARNA are significantly enriched in the mitotic CARs. 49 SNORA are enriched in the intherphasic CARs while only 5 in the mitotic CARs and of the 250 known SNORD class, 21 and 19 are found to be enriched respectively in the interphasic and mitotic CARs. Results are summarized in table 3.2. The complete lists of CARs belonging to snoRNA class are shown in the section 5.6 of Material & Method.

	Mitotic Total	HT1080 total	Interphasic CAR	Mitotic CAR
SCARNA	22	22	12	0
SNORA	97	84	58	5
SNORD	250	241	31	19

Table 4.2 Subfamilies snoRNA element distribution of CARs between intherphase and mitosis.

Comparing the list of the snoRNAs enriched in mitosis with that of the interphasic sample we can see that, even if smaller, it is not simply a subgroup: 11 of the 19 mitotic SNORD snoRNA are not enriched in the interphasic sample but seems to be specific of the mitotic chromatin.

DISCUSSION

1. CHARACTERIZATION OF NEW HUMAN NEOCENTROMERES

Animal neocentromeres are defined as ectopic centromeres that have formed in non-centromeric locations and avoid some of the features, like the DNA satellite sequence, that normally characterize canonical centromeres. Despite this they are extremely stable functional centromeres inherited through generations. The only existence of neocentromeres provide convincing evidence that centromere specification is determined by epigenetic rather than sequence-specific mechanisms.

For all this reasons, we used them as simplified models to investigate the molecular mechanisms that underlay the formation and the maintenance of a functional centromere. Specifically, our attention focalized on four human lymphoblastoid cell lines each carrying a neocentromere located on chromosomes 2, 3, 6 or in a ring chromosome of about 12 Mb derived from chromosome 9 (HL-2887, HL-Portnoi, HL-neo6 and HL-neo9, respectively). These cell lines have been cytogenetically characterized by FISH experiments in Rocchi's laboratory, University of Bari. In each of these cell line, neocentromeres are in heterozygosity which means that they arose only in one of the two homologous chromosomes. HL-neo6 cells have a normal karyotype and derived from a patient who don't show any metabolic and behavioural deficits. HL-neo9 cells derived from a patient with mild mental retardation and are characterized by the 9q partial deletion and the formation of a small marker chromosome in 60% of the cells. As well HL-Portnoi cells derived from a patient of normal intelligence but with pigmentary cutaneous anomalies and are characterized by an acrocentric marker chromosome present in the 30% of cells. Finally HL-2887 cells derived from a patient with facial dysmorphism and moderate mental retardation and are characterized by the formation of a highly stable marker chromosome derived from chr2. Therefore, neo6 neocentromere originates from a genuine event of centromere relocation (Capozzi et al., 2009; Ventura et al., 2004) and is an examples of evolutionarily new centromeres (ENC) (Capozzi et al., 2008) while the neo9, 2887 and portnoi marker centromeres are examples of human clinical neocentromere (HCN) which emerged in ectopic chromosomal regions (Amor and Choo, 2002; Capozzi et al., 2008; Warburton, 2004), as a result of a chromosome rearrangement, as observed in other cases (Alonso et al., 2003; Chueh et al., 2005).

To define in detail the region associated to these neocentromeres at the DNA sequence level, we applied a recent technology that integrates Chromatin Immuno-Precipitation and DNA

microarrays (ChIP-on-chip) on these cell lines, that have previously been cytogenetically characterized.

ChIP-on-chip analyses have been performed in HL-neo9 and HL-neo6 using two rabbit polyclonal antibodies directed against CENP-A or CENP-C human centromeric proteins, produced in our laboratory (Trazzi et al., 2009). These DNA binding-proteins are required for kinetochore function and are exclusively targeted to functional centromeres (Carroll and Straight, 2006). Thus, the immunoprecipitation of DNA bound by these proteins allows the isolation of centromeric sequences, including those of the neocentromere. The analysis, validate also by real-time PCR, demonstrated that CENP-A and CENP-C co-localize in both neocentromeres, provides their exact position and defined the sequence they occupy with the highest resolution currently possible (100bp).

ChIP-on-chip analyses have also been performed on HL-portnoi and HL-2887. Native ChIP using rabbit polyclonal antibodies directed against human CENP-A protein was done. ChIP DNAs were then hybridized onto custom genomic tiling array from Nimblegen. Data analysis provided CENP-A binding domain exact position and defined the sequence with the highest resolution possible nowadays.

1.2 NEOCENTROMERE DO NOT DEPEND BY THE PRIMARY DNA SEQUENCE

The fine mapping of a number of human neocentromeres has allowed a precise sequence comparison among different seeding domains (Capozzi et al., 2008; Marshall et al., 2008). The analysis, however, did not disclose any shared critical sequence features that could predict this potentiality, with the only exception of a satellite DNA in human that corresponds to an evolutionary new centromere (Carbone et al., 2006). In our cases, in agreement with the literature, we did not find any common feature between the primary neocentromeric sequences.

Firstly the CENP-A binding domain in our model appear to be quite different even in the length of the region involved. Collectively they are on average smaller then neocentromeres already characterized, even if with low resolution, in literature (~100 Kb): the smallest one (~54 Kb) is the neo9 neocentromere, then follow 2887 (~77Kb) and portnoi (~77,96 Kb) that show similar length, and at last neo6 (~84Kb). The genomic region involved in neocentromeric formation in marker chromosome could depend from the dimension: in neo9

a region of 54 Kb in enough to give stability to a marker ring chromosome of 12Mb. Probably region more extended should be necessarily involved in neocentromerization when the fragment to rescue from lost is bigger. For the sequence analysis some neocentromeres in literature have been shown to have higher AT content if compared to the average of the genome and seem to be enriched in LINE1 element. In our model this is true for the neo9 and 2887 neocentromeres (respectively with 41,58% and 29,30% against the 20,42% of the average). Conversely this is not true for neo6 and portnoi neocentromeres (respectively show 6,44% and 11,84%). Neo6 and portnoi share also the LTR elements enrichment in their sequences (respectively 17,59% and 12%, against the genomic average of 8%). Neo6, but not the other neocentromeres, is also quite enriched for SINE (19,68) in particular *Alu* sequence (17,54%) if compare to the average percentage of the human genome (13% and 10% respectively). Moreover only in neo6 sequence we have noticed a massive clustering of tRNAs.

The failure of neocentromeric sequences studying in order to find common and significant deviations from the genome average, in terms of various centromere motifs or repetitive elements (Marshall et al., 2008) suggests that the composition of the chromatin and its conformation, and not the underlying DNA sequence, are important for specifying a functional centromere.

1.3 NEOCENTROMERE EVEN ARISE IN PROTEIN-ENCODING REGIONS. THEIR FORMATION DO NOT REPRESS THE GENES WITHIN

One of the more fascinating features of neocentromeres is their location within euchromatic, protein-encoding regions of the genome. This is particularly evident in the cases of neo6 and portnoi neocentromeres which have gene transcripts (two in the case of portnoi) spanning inside the CENP-A binding domain. Determining if such euchromatic genes could be expressed within kinetochore chromatin has been a question of considerable interest. Through comparison of the expression levels of genes within neocentromeric domain (BTN3A2 gene encode inside neo6 and CLDN16 and TMEM207 both encode inside portnoi) to that of other limphoblastoid cell lines that do not present the neocentromere formation on the same chromosomal domain, we could investigate this problem. BTN3A2 is the gene encoded within the CENP-A domain of neo6. Its expression, evaluated by reverse real-time PCR, does

not reveal any differences when compared to that of other two limphoblastoid cell line (L1 and L2). The gene is still expressed even following the neocentromere formation.

This is in agreement with the few case reported in literature: no changes on genes expression profile were detect following the neocentromere formation in gene encoded within or near the CENP-A binding domain ((Lam et al., 2006; Saffery et al., 2003).

Different is the case we have found in portnoi. Inside the portnoi neocentromere domain there are two gene (TMEM207 and CLDN16). Surprisingly they apparently behave differently between each other. Reverse realt-time expression analysis comparing portnoi with neo6, using the GUSB gene as an internal control, reveals that the expression of TMEM gene is higher following the neocentromere formation while CLDN16 appears to be repress. In literature, the only differences in gene expression detected after a neocentromerization process was the activation of two genes encoded inside the S/MAR domain (scaffold/matrix attachment region), a domain described to be overlapping but much larger than the CENP-A binding domain and that some tend to consider as the physical boundaries of the centomere and defines the primary constriction (Saffery et al., 2003).

The CENP-A N-terminal tail lacks a lysine amino acid at residue 4, preventing the methylation marks that active genes, and consequently it might be thought that centromeric chromatin was silent by default. It has also been shown that CENP-A forms tighter nucleosomes structure than that of H3 which might also form a barrier to transcription. Collectively our expression analysis data tend to suggest that the neocentromerization process do not remodel the involved chromatin in order to create barrier to gene transcription but still be permissive.

Thus despite the increased scaffold attachment sites and a corresponding tighter chromatin packaging gene, transcription can continue and can occasionally even be promoted.

2. THE CURIOUSE CASE OF HORSE ECA11 CENTROMERE

2.1 THE ECA11 NEOCENTROMERE SHIFTING ALONG THE HORSE CHROMOSOME 11

The ECA11 centromere is the only horse centromere lacking any hybridization signal in fluorescence in situ hybridization experiments probing with the two major horse satellite

sequences. The absence of satellite signals in the ECA11 centromere suggests that this ENC may not have yet "matured" to the point of being endowed with satellite DNA (Wade et al., 2009). The centromere of ECA11 resides in a large region of conserved synteny in many mammals, where the horse is the only species with a centromere present, strongly suggesting that this centromere is evolutionarily new. ECA11 was cytogenetically localized, we then decided to precisely map, at the sequence level, the centromeric function using ChIP-on–chip experiments on fibroblast cell line using a rabbit polyclonal antibody against the CENP-A protein, that specifically localize to functional centromere.

The HSF cell line, used firstly to characterized the ECA11 domain, showed an overlapping domain for CENP-A/C antibodies. Of notice the total region interested in the CENP-A/C domain was of 400Kb, a huge domain if compared to that of other neocentromeres already known in literature (with average length domain of ~100kb). The analysis, showed two clear peaks of hybridization spanning about 136 kb and 99 kb, respectively, separated by a region of about 165 kb. In the sequence analysis of this region we found only five sequence gaps [none >200 base pairs (bp)], no protein coding sequences, normal levels of non-coding conserved elements, and typical levels of interspersed repetitive sequences, but no satellite tandem repeated sequences. We also found no evidence of accumulation of L1 transposons or KERV-1 elements, which were previously hypothesized to influence ENC formation.

Because of the original nature of HSF CENP-A/C domain we decided to extend the mapping of the ECA11 centromere to other horse fibroblast cell lines. Collectively it has been shown that ECA11 centromere differs among horse cell lines belonging to the same species. Surprisingly neither the region involved nor the length of the CENP-A binding domain appear to be conserved.

2.2 THE FIRST CASE OF STRUCTURAL POLYMORFISM

It's abundantly demostrated in literature that centromeres are functional structure extremely dynamics during evolution that could reposition along chromosomes in ectopic loci without altering the gene order (Montefalcone et al., 1999). In our horse study emerged a singular discovering never seen before: the locus where the neocentromere is established differ, at the basepair level, even within the same species its position is not fixed. Centromere could migrate along the chromosome. The plasticity of the centromere here showed is extremely interesting if we consider the importance of centromere and its conserved rule in all eucariotic organisms. Moreover this "shift" of the centromere along chromosome put in

evidence the epigenetic nature of the mechanism underlying the specification of the centromere identity. All together our results confirm that the centromerization phenomenon is not closely/exclusively due to the primary DNA sequence.

In HSF cell line, as well as in other horse fibroblast cell lines analyzed in this work (HSF-B, HSF-C, HSF-G), ECA11 centromere exhibit an unexpected double CENP-A domain. The position of these CENP-A binding domain and even the length of these differ between each other: the CENP-A binding domain do not overlap between different cell lines. We deeply analyzed HSF and HSF-G cell lines. In HSF the first region of binding is ~136 kb and the second ~99 kb while in HSF-G the first is ~91 and the second ~64. These lengths are quite in agreement with the average lengths of other neocentromeres before characterized in literature (~100 kb). The horse chr11 is mitotic stable (Wade et al., 2009)), thence the double domain could not be explained like a dicentric chromosome. Moreover, because CENP-A localized exclusively on active centromeres both the CENP-A binding domain seen in that studies must be distinct functional region.

Following these considerations, we had hypothesized that the two binding domain could identify the two different homologous chromosomes. This mean that each binding site localize only in one of the two homologous chromosome that belongs to the cell.

In order to verify this hypothesis, SNPs (*single nucleotide polymorphism*) in heterozygosis localized inside each of HSF and HSF-G ECA11 domains were separately identified. Through the sequencing of the immunoprecipitated CENP-A ChIP DNA along with the INPUT genomic control for both cell lines, emerged that in ChIP products only one of the two possible base of the corresponding SNP were present. This means that immunoprecipitating with α-CENP-A antibody we were able to isolate just one of the two homologous chromosome. Thus on the not immunoprecipitate chromosome the centromeric region is localized somewhere else in the genome. This confirm the hypothesis that centromere localizes on different genomic region between homologous chromosomes implying that the double binding domains seen in that horses ChIP-on-chip experiment are not localized on the same chromosome.

Similar SNPs analysis were done for a third cell line that have showed a unique peak of CENP-A binding site of about 185,5 Kb: HSF-D. Of notice the lengths of the domain involved in neocentromere formation is the biggest seen from our work. SNP in heterozygosis both from the boundaries and inside the central-core of the CENP-A binding site were chosen. Surprisingly the ChIP product sequencing become homozygotic on the boundaries but still remain heterozigotic in central domain. In agreement with other SNPs results, the huge

unique CENP-A domain seen in HSF-D cell line is at least the results of the partial overlapping of slightly different CENP-A binding site characterizing the two ECA11 homolog chromosomes.

All together these interesting results support the idea that ECA 11 is not a particular case of bipartitic centromere, but ECA11 should represent the first case of a structural centromeric polimorfism. Nothing similar was documented before in literature.

More experiments should be do in future in order to unequivocal confirm that the two CENP-A domain are positioned differently between the two homolog chromosomes of the same cell. For example this aim could be argue with the setting up of immunofluorescence coupled to FISH experiments on horse fibroblast chromatin fiber using specific BAC differently labeled distributed before and after the specific CENP-A binding sites immunodected with CENP-A antibody. Using the differently color distribution of fluorescence signals (between that of BACs and this of immunofluorescence) we should be able to rebuild the structural organization of neocentromere distinguishing between the two homologous.

2.3 CENTROMERE REPOSITIONING: A KEY EVENT IN EVOLUTION

Evolutionary studies on primate, marsupials, birds and rice have disclosed the unprecedented centromere repositioning (CR) phenomenon, that is the displacement of the centromere along the chromosome without disruption of the gene order: the old centromere is inactivated and a new one produced, which then becomes fixed in the population. The repeat-free region where the neocentromere initially form, progressively acquires extended arrays of satellite tandem repeats that may contribute to its functional stability (Ventura et al., 2004). Recently two human clinical neocentromeres on the chr15 map to duplicons that flanked an ancestral inactivated centromere. Furthermore, a neocentromere on chr3 has been found to map to the genomic place corresponding to the normal centromere in the Old World monkey. This scenario deeply affects our understanding of karyotype changes during evolution and strongly suggests that the present-day neocentromeres (like also human HCN) are better understood if viewed in an evolutionary frame.

Eight CR events have occurred in the last 3 MY in the genus *Equus* (Carbone et al., 2006). Surprisingly, at least five of these events appear to have arisen in the donkey after its divergence from the zebra, which took place approximately 1 MYA. It appears, therefore, that in some lineages, the CR phenomenon can be very frequent. The plasticity of the genome

(even the CR phenomenon) have important implications in the mechanisms of speciation. The Equus genome seems to be very dynamic. ECA11 centromere seems to be repositioned in horse cell line not far in evolution and had not had enough time to neither fix the position onto chromosome arm nor acquire satellite DNA.

Our work contribute to this evolutionary study and put new insights in the extremely plasticity of the centromere position. Centromere is not a cis-acting static structure. Its evolution is indipendent from the flanking marker gene and is not strictly depending from the primary DNA sequence. The formation of a neocentromere is a phenomenon of chromosomal rearrangement. The centromere repositioning plays a fundamental role in the evolution of the karyotype. The key question that fascinates biologists is understanding how the centromere plasticity could be combined to the stability and maintenance of centromeric function. The study of the mechanisms that regulate the location of the centromere and the formation of the kinetochore are all aimed to the understanding of the processes that regulate the transmission of genetic material. This information may allow the engineering of artificial chromosomes, which are used as important therapeutic carriers. The work presented is part of this research.

3. CHROMATIN ASSOCIATED RNA (CARs) IN MITOTIC AND INTERPHASE CELL IDENTIFY MANY INTRONIC AND INTERGENIC TRANSCRIPTS

Many observations suggest that transcription of centromeric DNA or of other non-coding RNAs could affect centromere formation: transcripts homologous to centromeric DNA have been detected in mammals (Bouzinba-Segard et al., 2006) and plants and are processed into siRNAs in plants (May et al., 2005; Neumann et al., 2007). Active protein-coding genes can reside within centromeric chromatin and being transcribed (Saffery et al., 2003; Yan et al., 2006) as we also have seen in two of our neocentromeres model. Centromere chromatin is not a compact place where transcription is avoid. Strikingly, RNAs derived from centromeric retrotransposons and CentC centromeric repeats are enriched in CENP-A chromatin immunoprecipitates in maize (Topp et al., 2004). Moreover, transcripts from human alpha-

satellite DNA are associated with CENP-C and the inner centromere protein INCENP, and the addition of recombinant alpha satellite RNA to permeabilized human cells is required to target exogenous CENP-C to centromeres (Wong et al., 2007)). Finally, a LINE1 retrotransposon of a human chr10 neocentromere has been shown to be transcribed and subsequently incorporated as a functional epigenetic component into the core neocentromeric chromatin (Chueh et al., 2009). Thus, it is possible that transcription of centromeric DNA promotes CENP-A deposition or, alternatively, centromeric RNAs might help to localize factors or protein complexes required for CENP-A deposition. Our study on RNA started with the aim to put new insight about RNA that could have functional role in chromatin compaction and centromerization. RNA could have an important role in centromeres because of their characteristics. They rapresent the best candidate to act as a flexible bridge between the very highly conserved protein component of functional centromere and the highly divergent DNA component.

Before focusing our attention on RNA that could have specific centromeric localization, was important to give a global view of RNA that could interact tightly with chromatin (CARs).

To date, there has been no thorough investigation addressing the identity of the chromatin-associated RNAs (CARs) on a global scale. This prompted us to develop a technique with the aim to identify CARs in a genome-wide approach using high-throughput genomic platforms.

Recent high-throughput transcriptomic analyses have revealed widespread transcription of the human genome (Cheng et al., 2005); the ENCODE Project Consortium 2007). A small portion of these transcripts code for proteins, while the rest are non-coding RNAs (ncRNAs). Only a limited number of ncRNAs have been assigned biological and molecular functions.

In this study, CARs were purified from interphase or mitotic human fibrosarcoma cell lines (HT1080) to be high-throughput sequenced on the Illumina platform. Collectively the sequencing data of CARs revealed some non-protein-coding RNA which hosts snoRNAs. Moreover we identify the association of many intronic and intergenic transcripts with chromatin, indicating that they may have structural and functional roles in chromatin organization through direct or indirect interactions with chromatin. Many of the isolated CARs belong to the family of snoRNA. These are a class of small RNA molecules that primarily guide chemical modifications of other RNAs and in which could be distinguished three subgroup: SNORA, SNORD and SCARNA. Between mitotic and interphase CARs we have identify differences within each family. Mitotic CARs are not simply a subgroup of that interphasic: many of SNORDA are specifically enriched in mitotic CARs when compared to the interphasic.

Because to date their documented function is to guide the modification of other very abundant RNA, and despite our technical precaution, a criticism could be done to this work is about nucleolar contamination. We could not totally rule out the possibility of small RNP complexes copurification in the sucrose gradient. Ono M. et al in 2010 have published a list of RNA they had isolated and sequenced from nucleolus, but the many part of the CARs we have found in our work is not belonging to that list (Ono et al., 2010). Moreover nucleolus fall apart in mitotis so our mitotic CARs should be avoid of any possible nucleolar component. Because the specificity of the enrichment for some snoRNA in mitosis we in future could test by RNA-ChIP experiment their specific association with protein that have specific role in mitosis or in chromosome condensation. An interesting candidate could be cohesin, the protein complex that regulates the separation of sister chromatids during cell division. One more suitable candidate could be condensin, proteins complex that play a central role in chromosome assembly and segregation in eukaryotic cells. Another interesting results we obtained only from our mitotic CARs analysis was the massively enrichment for RNA transcripts derived from alpha-satellite belonging to all chromosomes. These results were in agreement with the documented transcription of centromeric region and their findings in our sample mean their chromatin-association. They could be involved in centromere formation. Hopefully the centromere-CARs we already isolated and that should be sequence could put new insights in this founding.

We cannot determine, from the sucrose-gradient purified CAR, whether the chromatin association of these RNAs is stable or transient in nature but our experiment suggest a possible role for small RNAs in chromatin organization. The biological relevance of chromatin interaction of all different CARs is not clear. Since the intergenic and intronic ncRNAs identified in this study are based on their chromatin-association property, it is more likely that many CARs may mediate their actions through shaping the chromatin structure. It is possible that some ncRNAs simply act as a scaffold for protein factors, which, in turn, could guide them to the target genes in *cis* or *trans* through protein–protein interactions, thereby modifying the associated chromatin to regulate transcription either positively or negatively. Previously, a handful of intergenenic transcripts such as Xist, Hotair and H19 have been extensively investigated for their role in gene regulation. Identification of several intergenic CARs in this study provides the possibility of studying their actions in diverse biological processes. The most challenging task ahead is to identify the target regions for the intronic and intergenic CARs, and the modes of action by which these CARs influence their target genes.

This study is just the first step of a huge work that could be done: the characterization of the functions of these intergenic and intronic CARs would contribute to a greater understanding of the largely unexplored world of RNA-mediated functions. Moreover the determination of the RNA specificly associated to centromere could help us in the understanding more the unknown centromerization process. The same experimental procedures we set-up to isolate CARs will be applied to cells carrying a neocentromere in order to established if RNAs tightly associated to the normal centromere are also part of the heterochromatin-associated complex in neocentromeres, or if they derive from sequences near or within the CENP-A/-C domain.

MATERIAL & METHODS

1. CELL CULTURE

The neocentromeres of horse were studied on different horses fibroblast cell lines (HSF, HSF-B, HSF-C, HSF-D, HSF-G) that Prof. Elena Giulotto, University of Pavia provided us. These primary cell lines grow in high glucose DMEM (EuroClone) supplemented with 10% North American FBS (Fetal Bovin Serum), 2mM L-glutammine and 1% streptomycin/ampicillin at 37°C with 5% CO_2. Cells grows in monolayer and contact inhibition is consistently detected.
For the human neocentromeres studies lymphoblastoid cell lines carried neocentromeres in only one of the two homologues chromosomes were used (HL-Neo-6, HL-neo9, HL-portnoi, HL-2887).
These cell lines grow in RPMI-1640 (EuroClone) supplemented with 15% North American FBS (Fetal Bovin Serum), 2mM L-glutammine and 1% streptomycin/ampicillin at 37°C with 5% CO_2. Molecular cytogenetic analysis on Neo-3 and Neo-6 cell lines show stable and hereditable neocentromere in 3q24 and 6p.22.1 position respectivly. Neo cell line grow in suspension and form aggregates.
The RNA studies were carried in interphasic and mitotic chromosome of human sarcoma cell line (HT-1080). Them derived from a fibrosarcoma tissue of a 35 years old Caucasioan man. Cells grows in DMEM (Gibco) with 10% FBS (Fetal Bovin Serum), 2mM L-glutammine and 1% streptomycin/ampicillin at 37°C with 5% CO_2. Cytologically them consist of uniform population of undifferentiated tumor cells that multiplied rapidly with loss of contact inhibition.
1x Trypsin-EDTA (Euroclone) was used to harvest each adhesion cell types here discussed.

2. CHROMATIN IMMUNOPRECIPITATION (CHIP)

Chromatin Immunoprecipitation (ChIP) has become a popular method to detect the in vivo binding of proteins to DNA. Native-Chromatin Immunoprecipitation (N-ChIP) is used for protein tightly bound to DNA, like histones. Cross-linking Chromatin Immunoprecipitation (X-ChIP) is the technique used for other proteins bound to DNA. In this technique the protein/DNA complex is fixed with formaldehyde, a cross-linking agent that, because of its

short spacer arm (about 2 A), generates reversible covalent links mainly between proteins and DNA. The following are the protocol we set up.

2.1 NATIVE-CHROMATIN IMMUNOPRECIPITATION (N-CHIP)

- Harvest 2 x 10^7 cells and collect in PBS+10% FbS in 15 ml tubes, spin 4' at 1200 rpm at room temperature.
- Resuspend cells in each tube in 5 ml NB-A (85mM KCl, 5.5% saccarosio, 10mM Tris pH 7.6, 0.5mM spermidina, 0.2mM EDTA, 250µM PMSF) and add 5ml of NB-B (NB-A supplemented of NP40 to a final concentration that depend by the cellular type), mix by inversion and leave 3' on ice.
- Spin nuclei 4' at 2000 rpm, wash the pellet in each tube in 10 ml of NB-R (85mM KCl, 5.5% saccarosio, 10mM Tris pH 7.6, 1.5mM $CaCl_2$, 3mM $MgCl_2$, 250µM PMSF)
- Spin nuclei 4' at 2000 rpm and resuspend combined pellets in 0.5ml NB-R
- To quantify the nuclei:
  - Take a 5 µl aliquot of nuclei and dilute in 15 µl NB-R. Add 1 µl DNaseI
  - Incubate 5' at room temperature, then add 80ul of Urea Buffer (5 M Urea, 2 M NaCl)
  - Misure the A at 260nm and adjust the nuclei concentration in NB-R to A_{260}= 10
- Split nuclei in 0,5ml aliquote and add 1,5ng RNaseA each
- Digest the chromatin adding MNase at 60 units/ml, leave room temperature for 20'
- Add EDTA to 10 mM to stop digestion
- Spin down at 5000 rpm 30 sec in microfuge, transfer the surnatant (sn1, countains mononucleosomes)
- Resuspend the pellet in 225 µl $TEEP_5N$ (10mM Tris pH7.5, 0.5mM EDTA, 0.5mM EGTA, 250µM PMSF, 0.05% NP-40, 5mM NaCl) and leaving overnight the chromatin release at 4°C , with occasionally and gentle mixing
- Next day, spin the nuclei hard, 13000 for 10' and take the supernatant (sn2, containing the high molecular weight chromatin).
- For each IP we normally take 10ug of total chromatin (5ug of sn1 and 5 ug of sn2). Decide how many IP do, collect the chromatin needed and pull to 1ml volume in TEEP50N (10mM Tris pH7.5, 0.5mM EDTA, 0.5mM EGTA, 250µM PMSF, 0.05% NP-40, 50mM NaCl).

- Add 5ug of salmon sperm to each IP sample, leave 5' at room temperature and then add 100ul of proteinA-trisacryl slurry beads (Pierce) prewashed 2 times in TEEP50N.
- Leave 1h at 4°C rotating on wheel. This step is for preclear the chromatin from aspecifity binding to the beads.
- Spin at 4000rpm for 2' and collect the surnatant. Don't reuse the beads.
- Split the chromatin in 10ug aliquots, take one to use like a control (INPUT) and keep at 4°C
- Add the antibody of interest. For the CENP-A IP use 10 ul of the anti-CENPA 387 antibody (Tazzi et al 2009), leave at 4°C on rotating wheel for 20h.
- Add to each IP sample 50µl of ProteinA-trisacryl 50% *slurry* (Pierce) pre-washed 2 times in TEEP50N and leave the incubation proceed on wheel at 4°C for 4h
- Spin 1500 rpm to pellet the antigen/antibody/beads complex and wash it with 10 ml of TEEP140N (10mM Tris pH7.5, 0.5mM EDTA, 0.5mM EGTA, 250µM PMSF, 0.05% NP-40, 140mM NaCl). Leave to rotate 10 min at room temperature.
- Spin 1500 rpm for 2' to pellet the complex and wash it with 10 ml of TEEP200N (10mM Tris pH7.5, 0.5mM EDTA, 0.5mM EGTA, 250µM PMSF, 0.05% NP-40, 200mM NaCl). Leave to rotate 10 min at room temperature.
- Elute the immune precipitated protein of interest from the beads with 400 µl of Elution Buffer (TEEP50N + 1%SDS) each IP, leave 30' at room temperature on wheel.
- Spin at 4000 rpm for 2' and collect surnatant in fresh tubes.
- Purify the immunoprecipitated DNA on QIAquick PCR purification column (QIAGEN) following the kit instruction. Use 100µl di buffer EB (10mM Tris-Cl, pH 8.5) to elute DNA from the column.

➢ Test the DNA centromeric enrichment by Real-Time PCR using primer onto centromeric-satellite sequence (that of caballus or those of human alpha-satellite, see below) and proceed with the Whole Genome Amplification (Sigma) following the instruction in order to achieve the 4µg necessary for the microarray hybridization.

2.2 CROSS-LINKED CHROMATIN IMMUNOPRECIPITATION (X-CHIP)

- Harvest 20.000 cells and dilute them in 20ml of media in 50ml falcon tubes
- Add 540µl of 37% formaldehyde (to 1% final concentration) and mix immediately. Incubate samples on a rotating wheel for 10 minutes at RT

- Add 1ml Glycine from a 2.5M stock solution and mix immediately. . Incubate samples on a rotating wheel for 10 minutes at RT
- Centrifuge samples at 1000 rpm for 10' at room temperature, then keep samples on ice;
- Wash the cell pellet in 5ml of PBS 1%PMSF for 3 times
- Remove supernatant and resuspend pellet in 500µl ice-cold Cell Lysis Buffer (Pipes pH8 5mM, KCl 85mM, NP40 0,5%, PMSF 1:100, protease inhibitor cocktail) Pipette up and down 10-20 times, then incubate on ice for 10'.
- Spin 3000 rpm at 4°C for 4 min.
- Remove supernatant and resuspend gently the nuclei pellet in 600µl ice-cold Nuclei Lysis Buffer (TrisHCl ph8.0 50mM, EDTA 10mM, SDS 0,5%, PMSF 1:100, protease inhibitor cocktail). Leave on ice for 10' at least.
- Sonication of cross-linked cells is performed with the Diogene Bioruptor for 1h at a full power in a tank filled with ice/water in order to keep cell samples at low temperature during sonication
- Centrifuge samples at 14000 rpm for 15 minutes at 4°C
- Transfer supernatant to a fresh tube and pre-clear lysate by incubating it with 75µl of Immobilized Protein A (Pierce) for 15 minutes in the cold room at constant rotation; then spin down the beads at 4800rpm and keep the surnatant.
- Take a volume that correspond to 2.000.000 cells to use like Input DNA and keep at 4°C.
- A volume that correspond to 10.000.000 cells and 5µg of antibody of interest is used each IPs. Rotate the sample O/N in the cold room
- Prepare the Immobilized Protein A (Pierce) as follow:
  - Wash beads in 1ml Ripa Buffer (150mM NaCl, 1% NP40, 0,5% NaDoc, 0.1% SDS, 50mM TrisHCl pH 8, 1mM PMSF)
  - spin at 2000 rpm for 2';
  - incubate beads in 1ml of Ripa Buffer with 1% BSA and 5ug of Salmon Sperm in rotate sample for 4h at 4°C
  - Wash the beads in 1ml of Ripa Buffer 2 times and then resuspend in Ripa Buffer to make a slurry solution.
- Add 50µl of Immobilized Protein A (Pierce) pre-washed, and incubate by constant rotation at room temperature for 30'

- Centrifuge the sample at 4000 rpm for 5 minutes at room temperature
- Remove the supernatant and proceed to wash the beads. For each wash, incubate the sample by constant rotation for 3 minutes at room temperature and the centrifuge at 4000 rpm for 2 minutes at RT
  - Wash 4 times with 1ml of Ripa Buffer added of Complete protease inhibitors (Roche)
  - Wash 4 times with 1ml Washing Buffer (Tris HCl pH8 100mM, LiCl 500mM, NP40 1%, NaDoc 1%, PMSF 1mM, Complete protease inhibitors (Roche));
  - Wash 2 times with 1ml TE buffer + Complete protease inhibitors (Roche);
- Remove the supernatant and add 70 µl TE buffer to the beads. Add 10µg RNAse A and incubate at 37°C for 30 minutes.
- Add 50µl Proteinase K Buffer 5X and 6µl Proteinase K (19mg/ml). Then, incubate at 65°C in a shaker at 950 rpm for 6 hours
- Centrifuge at 14000 rpm for 10 minutes at 4°C, then transfer the supernatant (250µl) to a new tube
- Purify the immunoprecipitated DNA on QIAquick PCR purification column (QIAGEN) following the kit instruction. Use 100µl di buffer EB (10mM Tris-Cl, pH 8.5) to elute DNA from the column.

3. PRIMER DESIGN AND PCR REACTION

The genomic sequence of the different organism studied were taken from the http://genome.ucsc.edu bioinformatic database and specific primer were designed using Primer3 on-line source. Primers were chosen to satisfy some requirements: the high specifity for the region of interest; the melting temperature should be between 55°C and 61°C; the GC content should not be more then 60% and region of complementarity inside the primer sequence were avoid to decrease the secondary structure and primer dimer formation. Each primer were blast over the entire genome of interest to check for the specificity. Different primers were designed for different purpose.

Primers were tested on genomic DNA extract from horse cells with the Kit Blood & Cell Culture DNA midi (QIAGEN).

Routine PCR reaction were done using the Herculase II (STRATAGENE) in which the processivity of high-fidelity PCR is increased by fusing the Pfu-based DNA polymerase with a high affinity double-stranded DNA binding domain. The reaction contained distilled water

(dH2O) to a final volume of 25 or 50µl , 5× Herculase II reaction buffer, dNTPs (25 mM each), DNA template (150ng), Primer mix of interest (10 µM) and 0,25 µl of Herculase II DNA polymerase. The reaction consist of an initial 2' step at 95°C to activate the polymerase, follow by 40 cycles of 20'' at 95°C, 20'' at 55°C and 30'' at 72°C, the last step is an elongation time of 3' at 72°C.

Each PCR reaction was purified by QIAquick PCR purification column (QIAGEN) following the kit instruction. 1 µl of the DNA eluted were quantify with the NanoDrop thermo scientific micro-volume spectrophotometer and 500ng were run on an agarose 1% gel in electrophoretic cumber.

3.1 PRIMERS FOR *HORSE* NEOCENTROMERES

To test by Real-Time PCR the immunoprecipitated enrichment of the ECA11 centromeric domain over the input control we used the following primers. PCR products have similar length (90bp ≤ product length ≤ 110bp):

1picco	chr11:27770997-27771100	Forward:	ctccttctcatgggttgcat
		Reverse:	ctggtcctcctcagcatctc
1picco2	chr11: 27687704-27687797	Forward:	caaagcctgggaaaacactc
		Reverse:	cacgtgcccctgttttactt
1picco3	chr11: 27739831-27739923	Forward:	gctttggagacaagcagacc
		Reverse:	atgctttggtgggagttcac
2picco	chr11: 27990583-27990679	Forward:	ctttgcgcatgtctctcaaa
		Reverse:	gctgcacacaaaacgaaaga
2picco2	chr11: 27966050-27966138	Forward:	cataacccctggcatcctta
		Reverse:	tgccccagggataaatcata
2picco3	chr11: 27985955-27986054	Forward:	tccactttcgacaacactgc
		Reverse:	acggacataccgttgcctac
noup	chr11: 27569560-27569649	Forward:	atgccctggactgtaaaacg
		Reverse:	atcctcaaagctgagccaaa
nointra	chr11: 27934666-27934768	Forward:	gcttctcgcccatatgaaag
		Reverse:	atgtcccaacgctgaaaaac
nodown	chr11: 28227839-28227938	Forward:	ttgccctgatagcgagaaat
		Reverse:	ctattccttggggcttctcc
sat caballus	canonical centromere	Forward:	cttcccaaagagctggaagc
		Reverse:	tttgccctagagctgaaagg

3.2 PRIMER FOR THE SEQUENCING OF CENTROMERIC SNPS (*SINGLE NUCLEOTIDE POLYMORFISM*) IN HORSE

The http://www.broadinsitute.org/mammals/horse/snp web site were used to determine the ECA11 centromeric region SNPs (*Single Nucleotide Polymorfism*) of the horse population. Specific primers over genomic region quite enriched for SNPs inside the CENP-A binding domain of the HSF cell line were designed. The average size of the products is 600bp.

523-1126	chr11:27648523-27649126	Forward	tccgagcaccatctatactt
		Reverse	agagcaatctgtcggtaaa
446-1190	chr11:27732446-27733190	Forward	ccaaaccaaacagaagaaag
		Reverse	atgcagatatccagttgctc
323-1020	chr11:27743323-27744020	Forward	gctaagcctatgtcttggaa
		Reverse	atgcctctacctatgtttgc
206-963	chr11:27769206-27769963	Forward	ctggttgcagctaagagtatt
		Reverse	acatactcagaaactggcaga
66-613	chr11:27728066-27728613	Forward	ccacagaaagtcatttttcc
		Reverse	agttgcagaaagtggctaga
101-813	chr11:27744101-27744813	Forward	gccagagggaatataattcag
		Reverse	ctatgtcaggatccttcctc
426-111	chr11:27966426-27967111	Forward	cagcacagttatggattcct
		Reverse	ttacgtttgatgctgtcttg
990-634	chr11:27985990-27986634	Forward	caaaaggtcattgaagaagg
		Reverse	tgctacatgtttgtttgcag
960-80	chr11:28026960-28027080	Forward	gatctgatgggaggtcatta
		Reverse	tgatatgtcctggtttacagtg
169-682	chr11:28045169-28045682	Forward	gaatgcttccttctccact
		Reverse	ggttgatggttgtcaagaaa

Heterozygotic SNPs identified inside the CENP-A binding domains (first domain 27.643.412-27.779.345 bp and second domain 27.950.821-28.049.577 bp) in the HSF genome were analyzed in CENP-A immunochipped DNA. Primers for the HSF cell line were:

SNP 1PICCO	chr11: 27648683-27648795	Forward	agcagctattatgtggtggt
		Reverse	gcaaaatcctgtgcaatc
SNP 1PICCO2	chr11:27728393-27728492	Forward	aatggccagtgatggttac
		Reverse	agagttgggagctgttcttt
SNP 1PICCO3	chr11: 27732643-27732782	Forward	catgacacatcgtagaaaca
		Reverse	tttcttctactggcagcttt
SNP 2PICCO	chr11: 27966473-27966577	Forward	caatgtaatattgttaaggagca
		Reverse	cctagcaactaggcaagatt
SNP 2PICCO2	chr11: 27966569-27966679	Forward	tcttgcctagttgctaggag
		Reverse	catcattgactgaaatgtcg
SNP 2PICCO3	chr11: 27966992-27967111	Forward	acaaggaaccaattacctga
		Reverse	ttacgtttgatgctgtcttg

The same process has been followed for the HSF-D and HSF-G centromeric domain. The following are the SNPs genomic primers used:

HSF-D GenA	chr11:27691217-27691671	Forward	catgtatgacctggaaggat
		Reverse	tccacgacagtcaagataca
HSF-D GenB	chr11:27713107-27713632	Forward	gtgatggaggactcttgagaa
		Reverse	tgggatgaactctagtttcc
HSF-G genA	chr11:27737024-27737494	Forward	gcagagcttgttaggaaaaa
		Reverse	ccagagtgtaaaaagcttgg

Primers used on the immunochipped material of HSF-D and HSF-G cell lines were:

HSF-D SNPsA	chr11:27643776-27643907	Forward	gtgtttcaagaggaagcagt
		Reverse	gagtgctcacactgactctg
HSF-D 452	chr11:27691452-27691564	Forward	ctcaagagagttatgctggatg
		Reverse	tgctaatctgttcttcatctcc
HSF-D 623	chr11:27691623-27691724	Forward	catgagggatctgtgtgata
		Reverse	atcacaaccaggaaattgac
HSF-G SNPsA	chr11:27743329-27743434	Forward	cctatgtcttggaagcactc
		Reverse	ctatccccgggttaaataat

HSF-G SNPsB	cher11:27937342-27937443	Forward	ttaaagtggccaatcctcca
		Reverse	ccctcgagatcacaggaagt
HSF-G SNPsC	chr11:27937367-27937509	Forward	ttagatccccaaataactgc
		Reverse	tcaggtgatgattgttgcta

3.3 HUMAN NEOCENTROMERES

To test by Real-Time PCR the immunoprecipitated enrichment of neo6, neo9, 2887, portnoi neocentromeric domain over the input control we used the following primers:

2887-1	chr2:41765941-41766040	Forward	tctatggcattggtgttcca
		Reverse	ccctcattcacaaggttgct
2887-2	chr2:41795300-41795495	Forward	gggcacaggtttgtagcatt
		Reverse	ttccccagactctcatggac
2887-no	chr2:8076782-8076884	Forward	gcaaggtggctaaattacgg
		Reverse	ggatttgaacccttctgcaa
portnoi1	chr3:190102319-190102409	Forward	tagtgcaagcactgggtgag
		Reverse	tccagggctattgattttgc
portnoi2	chr3: 190146710-190146826	Forward	tttcttttctccccgtgttg
		Reverse	agaggttttatgccccaacc
neo6-no	chr6: 25351300-25351392	Forward	tttcctgctgttcctgctct
		Reverse	tcactggaggactcaccaca
neo6-1	chr6: 26418586-26418685	Forward	acaaacggcagaggctctaa
		Reverse	ttaaagcgtcccatctgctt
neo6-2	chr6: 26454446-26454548	Forward	cctatgcccaaactcagca
		Reverse	gggcaacaaattccctttt
neo9-no	chr9: 121403401-121403422	Forward	aaattccccccggagtacaca
		Reverse	gggcactgaagactgaatctt
neo9-1	chr9: 121295048- 121295149	Forward	cgaagctgcttcaagtcacct
		Reverse	tccctggcaaagaacagaaag
neo9-2	chr9: 121311694-121311792	Forward	gcccagagtgcaatcgtgac
		Reverse	caggagctgagctggctttatt

a-sat	canonical centromeric satellite	Forward	aaacttctttgtgatgtgtg
		Reverse	aaagcggtccaaatatcc

For the expression profile analysis of the gene encoded within the CENP-A binding site of neo6 and portnoi the following primers on the cDNA of transcripts were used:

CLDN16	Forward	aatgcttttgatgggattcg
	Reverse	catcaacgctcgagttacca
TMEM	Forward	catggcagtttttgctgttg
	Reverse	ggtcaggggtttgagtttga
BTN3A2_A	Forward	aagacagccagcatttccat
	Reverse	gagaagcagcagcaagatagg
BTN3A2_A	Forward	gcaacagagcgggaaataag
	Reverse	acgaagactcctctccacga

3.4 REAL-TIME PCR

Real-time polymerase chain reaction (RT-PCR) is used to amplify and simultaneously quantify a targeted DNA molecule. It enables both detection and quantification (as absolute number of copies or relative amount when normalized to DNA input or additional normalizing genes) of one or more specific sequences in a DNA sample.

Is possible to following the amplify reaction because of the presence of SYBR Green in the reaction. SYBER Green binds to double-stranded DNA. The resulting DNA-dye-complex absorbs blue light (λ_{max}=488 nm) and emits green light (λ_{max}=522 nm). The intensity of the signal is directly associated to the quantity of DNA duplex form during the reaction.

The Bio-Rad SYBER green is a mix of 100mM KCl, 40mM Tris-HCl pH8.4, 0.4mM each NTPs, Taq DNA Polimerase 50u/ml, 6mM $MgCl_2$ Sybr green I, 20nM fluorescin and stabilizators. 2 ng of each DNA template is added of SYBER green 2X (Bio-rad), 800nM of each primer and distilled water to a 20ul final volume.

The standard protocol consist of: 3’ at 95°C, 40 cycles of amplification as follow 30’’ at 95°C, 15’’ at 60°C and 30’’ at 72°C.

The datas were analyzed with the ΔΔCt method.

4. MICROARRAY

For each ChIP-on-chip experiment 4µg of both input and CENP-A\C immunoprecipitated DNA were sent to Roche Nimblegen to be hybridized onto microarray. Samples immunoprecipitated were stained with Cy5 fluorophore (red), while the total chromatin (input) was labeled with Cy3 fluorophore (green). Differently labeled the sample were cohybridized onto Customs NimbleGen Tiling Array 385K. Fluorescence signals were detected by the NimbleScanTM scanner and processed by sophisticated software to give as output a log2 ratio between immunoprecipitated signal against the control signal. Data files (.gff) are displayed using the SignalMap software.

The custom array we design had an average resolution of 150bp. The probes used were oligomers of 50bp (50-mer) that map in the genome of interest just one time. This mean that repetitive sequence are avoid from the array. A total of 385,000 total probes could be spotted per array. The region coverage by our custom design were the following (from UCSC database) :

Organism	Build number	Chr	start/stop coordinates
Homo sapiens	HG17	3	146,500,000-151,500,000
Homo sapiens	HG17	3	88,000,000-97,000,000
Homo sapiens	HG18	3	189,000,000-194,000,000
Homo sapiens	HG18	6	24,000,000-29,000,000
Homo sapiens	HG18	6	57,000,000-64,000,000
Homo sapiens	HG18	9	118,500,000-123,500,000
Homo sapiens	HG18	9	45,000,000-67,000,000
Homo sapiens	HG18	1	81,500,000-86,000,000
Homo sapiens	HG18	1	117,500,000-142,500,000
Homo sapiens	HG18	2	39,000-000-44,000,000
Homo sapiens	HG18	2	88,000,000-97,000,000
Equus caballus	sep 2007/EquCab2	11	25.600.001-28.800.000
Equus caballus	sep 2007/EquCab2	11	44.516.043-50.918.798

4.1 STATISTICAL ANALYSIS OF BINDING PEAKS

The raw Nimblegen data were analyzed by a statistical on-line server in order to detect peaks of signals that correspond to the binding sites of the protein onto the genome to finely locate

the boundary. The name of the server we used was TAMALPAIS. http://chipanalysis.genomecenter.ucdavis.edu/cgi-bin/tamalpais.cgi (Bieda et al 2006).

The strategy is to consider the binding site in the data set as runs of consecutive points (each point representing a 50-mer) with enhanced amplitude. To overcome the problem of the threshold setting (that should be a function of the appropriate combination of amplitude and width for each array) this algorithm use a percentile for each array (95th and 98th percentile) of log2 oligomer ratios. Use of this percentile "normalizes" the threshold values for each array to reflect both the amplitudes and distribution of signal in the arrays and, furthermore, presents a consistent, non arbitrary way to set thresholds for different arrays. For each threshold (95th percentile and 98th percentile) the server use $P < 0.0001$ for a very stringent *P*-value (which requires six consecutive points above the 98th percentile or eight consecutive points above the 95th percentile) and $P < 0.05$ for a less stringent *P*-value cutoff (which requires four consecutive points above the 98th percentile or five consecutive points above the 95th percentile). They give as output four conditions, in decreasing stringency: 98th percentile threshold and $P < 0.0001$; 95^{th} percentile threshold and $P < 0.0001$; 98th percentile threshold and $P < 0.05$; 95th percentile and $P < 0.05$. As stringency is decreased from L1 to L3, we see a small increase in the number of detected peaks and in the apparent size of the peaks.

Because lowering the stringency results in an increase in false positives, for our peaks detection we decided to choose the higher stringency analysis : 98th percentile threshold and $P < 0.0001$.

5. CARs ISOLATION

5.1 CARs EXTRACTION FROM INTERPHASE CELLS

This is the protocol we set up and that was used to prepare the sample called *interphasic6-7*

- Four HT1080 F175 flask were treated with 50µl of [5-3H] uridine and 50µl of [METHYL-14C] thymidine and 10 µl of each of 37mM dA, dG, dC. Leave at 37°C for 4h
- To each was added 150 ug/ ml of alpha-amanitin

- Harvest cells using trypsin, wash in PBS 10% FBS, resuspend pellet in 5ml NBA (85 mM KCl, 5,5% sucrose, 0,5 mM spermidine, 10mM Tris Hcl PH 7.5, 250uM PMSF)
- Add each 5ml NBB (NBA plus 0,8% NP40) and leave 3 min at 4°C
- Collect nuclei spinning at 2000 rpm at 4 C for 4 min. Resuspend pellet in a total of 10ml NBR (85 mM KCl, 5,5% sucrose, 10mM Tris pH7.6, 1,5mM CaCl2, 8mM MgCl2, 250 uM PMSF)
- centrifuge 4000 rpm for 4 min at 4°C and resuspend each pellet in 1ml of NBR.
- Quantify the OD of the solution: dilute 5ul of the solution in 95ul of water, add 1ul of DNase, leave 5 min at room temperature, stop reaction with 400 Sonication Buffer (5M urea, 2M NaCl) and determine the wave length at 260 at spectrophotometer. Adjust the OD at λ=260nm to 20.
- Take 1 ml of nuclei at 20 OD and digest 10 min at room temperature with activated mDFF. To activate the mDFFT, take mDFFT and dilute 1:2 in water, add 1ul of AcTEV Protease 10U/ul (Invitrogen) any 50ul reaction and leave 30°C for 1 h. Then keep at room temperature. For our experiment we had used 20ul and 40ul of active DFF.
- Stop each reaction with 8mM EDTA, Spin at 5000 rpm for 1 min and resuspend in 800µl of TEEP20 (10mM Tris PH 7.5, 0.5mM EDTA, 0.5 mM EGTA, 250uM PMSF, 0,1% NP40, 20mM NaCl) added with 2ul Protector RNase inhibitor (Roche);
- Leave the releasing of digested chromatin from permeabilezed nuclei over night at 4°C
- Spin at max speed at 4°C, keep the surnatant and add 2µl of Protector Rnase inhibitor. This is the sample to load on the top of step sucrose gradient
- Sucrose gradient is made in Beckman 50 Ultra Clear tubes (13X51). Each step of the sucrose gradient is in made in TEEP80 (10mM Tris PH 7.5, 0.5mM EDTA, 0.5 mM EGTA, 250uM PMSF, 0,1% NP40, 80mM NaCl) supplemented of sucrose to achieve the percentage needed. Our step was with 1,5ml of 50% sucrose on the bottom, 2ml of 20% sucrose in the middle and 10% sucrose the upper step. Onto that gently add, drop to drop, each experimental sample
- Spin at 50000rpm in the MLS50 rotor in the Optima MAX-XP ultracentrifuge for 1,50
- Use the 60% sucrose in TEEP80 to pump out the gradient from tubes and collect 800ul of each of a total of 10 fractions. Take aliquots of each fractions to check the quality of DNA and RNA.
- Keep the fraction at 4degrees until use, add 5ul of Protector Rnase inhibitor (Roche);

- Pulled together fractions corresponding to the MW of interest (in our case was the fractions 6 and 7)
- In order to attach the biotin at each DNA extremity, add to the sample 10 µl of RNAse Inhibitor, 10ul of the 100mM CoCl2, 2,5 ul of the 200uM ddT, 25ul of the 1mM Biotin-dUTP and 12ul of TdT.
- Incubate 8h at 37°C and then purify onto exclusion gel column (Stratagene) from the un-biotinilated fraction. Check out the efficiency of the byotinilation
- Mix sample with 0.5 ml beads (4 mg/ml; pre-blocked for 5 min with BSA and ssDNA)
- Mix the chromatin-biotin labeled fraction eluted from spin column, with 500ul (that correspond to 2mg) of streptavidyn-magnetic beads (Invitrogen) pre-blocked for 5 min with BSA and ssDNA for 1h at room temperature. Incubate 2 hours at 4°C rotating on wheel.
- Washed the chromatin-beads complex with TEEP80 for four times. Then resuspended them into 500ul of TEEP80.
- Washed the chromatin-beads complex with TEEP160 (where 160 is the mM of NaCl) for two times and then one time with TEEP320 (where 320 is the mM of NaCl) and at least resuspend in 250ul of TEEP80.
- Added to the sample 750ul of TRI-reagent LD. See the 5.2 for CARs purification protocol followed.

➢ For the DNA-RNA content specific quantification we had use the Scintillation counter to measure ionizing radiation. Samples corresponding to 5% of the total where taken for each critical step.

The HT1080-TOT used as referee for *interphasic6-7* sample was made starting from a T75 flask of HT1080 at confluence. 7,5ml of TRI-reagent were added and RNA was extracted following the normal protocol of the kit. The section 5.2 explain what we following did. The RNAs were at least depleted from the massively Ribosomal RNA component by applying the RiboMinus Mouse/Human kit (Invitrogen).

5.2 CARs PURIFICATION

- The TRI-REAGENT LD protocol was followed in order to purify RNA.
- Pellet of RNA was resuspended in 25ul of RNase free water and quantified by Nanodrop spectrophotometer.
- Qiagen RNeasy Kit min-elute column kit protocol was used.
- Turbo DNase from Ambion in its buffer was used on column to degradate DNA
- RNA was eluted in 100ul of RNase free water and was left at -20°C in 1 ml of isopropanol added of 1ul of glycogen until the RNA sequencing.

5.3 CARs FROM CENTROMERIC CHROMATIN

The recepies of solutions, if not differently indicated, were the same of the 5.3 section.

5.3.1 PROTOCOL 1

- Four HT1080 F175 flask were trypsinized and collected in falcon tubes.
- Crosslinking:
  - Addiction of 37% formaldehyde to 1% final concentration;
  - Incubation 10 min, at room temperature, in agitation;
  - Adding of 1,25 M Glycine to 0,125 M final concentration;
  - incubation 5 min, room temperature, in agitation.
- Samples were spun at 1200 rpm, 5 min, 4°C and washed in PBS
- Pellet was resuspend in 5 ml of NBA, 5ml of NBB and incubate 3min on ice
- Samples were spun at 2000 rpm, 4 min, 4°C and then washed in 10 ml of NBA and finally resuspended each in 500ul of RIPA (150mM NaCl, 0.1% SDS, 0.5% Na deoxycholate, 1% NP40, 50mM Tris pH7.5; 1mM PMSF, protease inhibitor cocktail) supplemented with RNAse inhibitor to 0,2 U/ml final concentration
- Chromatin was sonicated keeping the sample on ice: 8x 20sec at 2 Amplitude
- Spun samples at 14000 rpm, 15min, 4°C.
- Supernatant was transferred in new tubes. 50 µl was taken as our Input. Store the Input at 4°C and remember to de-crosslink the RNA before the usage.

- Samples were precleared as following:
  - samples were diluted by adding 4 volumes of RIPA;
  - add 50 µl protein A magnetic beads previously associated to 25ug of rabbit-IgG and washed 3 times in RIPA;
  - incubate on rotating wheel, 1h, 4°C
- Removed beads with aspecificies and aliquot supernatant (500 µl for each IP). One sample was kept on ice, it was our INPUT. It was decross-linked and DNA was extracted to real time PCR analysis.
- To each IP 20ug of antibody of interest was add. We used anti rabbit-IgG antibody (Santacruz) and the polyclonal anti-CENP-A antibody, produced by our own lab
- Incubate on rotating wheel, 4°C, over night
- To each IP sample 50 µl of pro-A magnetic beads (previously blocked with ssDNA and washed 2 times in RIPA) were added.
- Samples were incubate on rotating wheel, 4h, 4°C. RNAse inhibitor 0,1U/ml was added.
- Immunocomplexes were washed 3 times in RIPA
- Immunocomplexes elution:
  - add 250 µl of Elution buffer (1%SDS, 250uM PMSF, 0,1 U/ml RNAse inhibitor, prepared in TEEP80 solution) to the each IP beads;
  - incubation of 30 min, at 37°C, shaking at 800 rpm;
  - supernatant was transferred in a new tube;
- Take a 25 µl aliquot from each IP sample for protein extract
  - add 1 volume of 2X SDS-LB;
  - incubate at 100°C, 5 min;
  - spin 15min, 14000 rpm;
  - ready to use: store at -20°C
- Reverse the crosslink: samples (each IPs and INPUT) were incubated 1h, 65°C with Proteinase K Buffer 5X and 6µl Proteinase K (19mg/ml)
- One IP sample for both antibodies used were taken in order to extract DNA to do Real Time PCR analysis. Human alpha-sat and nodown primers were used. Real time data analysis shown a centromeric enrichment in CENP-A IPs of 1,5 times.
- Add 750 µl of Tri-Reagent LS to each sample (1ml tot) to extract RNA as done in §5.2

5.3.2 PROTOCOL 2

- Four HT1080 F175 flask of cells were harvested using trypsin.
- The § 5.1 protocol were then followed until the sucrose gradient step.
- Fraction from sucrose gradient where collected and evaluation of DNA (run on 1% agarose gel), RNA (run on 1% denaturating agarose gel) and protein (western blot using anti-CENP-A antibody and anti-histoneH3, Santacruz, were done) content for each fraction were done.
- The fractions of interest were added each of 5ul of Protector Rnase inhibitor (Roche) and they were pulled together. Fraction of interest must show the presence of CENP-A and had high MW of chromatin).
- Fractions were dilute in TEEP80 3 times and kept on ice until use
- Precleared the chromatine with Protein A Magnetic beads #S 14255. New England Biolabs
- Wash 100ul beads in TEEP80, 2 times. Beads were resuspended in 500ul of TEEP80
- add 4ug rabbit IgG, 100ug salmon sperm, left rotating on wheel for 30min at room temperature. Beads were then washed 2 times in TEEP80.
- Washed beads were added to chromatin and left 1 h on wheel at room temperature
- Took the beads off and do not reuse them.
- Protein A Magnetic beads were left to associate to each antibody of interest for the following use in immunoprecipitation. We used anti rabbit-IgG antibody (Santacruz) and the polyclonal anti-CENP-A antibody, produced by our own lab
- Beads (50ul for each IP) were washed with TEEP80 for 2 times.
- 100ug of the antibody of interest was added in a 1ml total TEEP80 volume
- Left at room temperature for 1 h on wheel
- The precleared chromatin was split into the IP sample and 50ul of the beads-antibody complex was added to each. 100ul of the precleared chromatin was taken as input (From input we had extract the DNA in order to use it in Real Time PCR to see if ChIP worked).
- Each IP was boost to 1ml in TEEP 80
- IP samples were left 4h on wheel at room temperature
- Washed IP sample 2 times in TEEP80, 15 min on wheel at room temperature. Washed then 2 times in TEEP160 and left 15 min on wheel at room temperature

- Wash 2 times in TEEP160, 15 min on wheel
- To extract RNA TRI-reagent was put on beads
- The § 5.2 protocol were then followed

➢ Input sample and one of both IgG-IP and CENP-A-IP samples were taken apart in order to extract DNA. Qiagen PCR column protocol were used and DNA was eluted in 50ul of water. 1,5ul was used for Real Time PCR. Primer used was: human alpha-sat and nodown primer (on human chr6). Real time data analysis shown a centromeric enrichment in CENP-A IPs of 1,9 times.

5.4 CARs FROM MITOTIC CHROMOSOME

This is the protocol we set up to obtain the samples called TEEP-80 and TEEP-160

- Four F175 flask with HT1080 near confluence were treated with COLCEMID (0.1 ug/ml) over night
- Mitotic cells were harvested and washed in PBS 10% FBS for 2 times.
- Pellet of cells were washed in NBR3 (5,5% sucrose, 10mM tris ph 7.5, 3mM MgCl2, PMSF, added of phosphatases inhibitor and RNases A/T inhibitor)
- Samples were kept at room temperature for 15 min and then sonicated 3 times at amplitude of 2 for 15 sec.
- The shared chromatin was spun down at 5000 rpm, pellet was resuspend in 500ul of TEEP20 and left on ice for 1h.
- Samples were spun down at high speed and the pellet was washed in TEEP80.
- From that pellet we obtained the sample called TEEP-80 adding the TRI-reagent and following its normal protocol. The 5.2 section was the protocol used to purify
- To prepare the sample TEEP160 samples were washed further 2 times in TEEP-160 and then TRI-REAGENT was added.
- The protocol in 5.2 section was followed to purify CARs.
- After that, RNA was depleted from the massively Ribosomal RNA component by applying the RiboMinus Mouse/Human kit (Invitrogen).
- Sample, kept in isopropanol at -20, were then sent to be sequence.
- Aliquote of each step were taken and used to check DNA and RNA quality. Even protein were extract in order to detect by Western Blot the presence of the H3ser10 phosporilation (marker of mitotic chromosome) in fraction used for CARs isolation.

The MITOTIC-TOT used as referee for TEEP80 and TEEP160 samples sample was made starting from a T75 flask of HT1080 treated over night with COLCEMID. Mitotic cells were collected in falcon tube and 7,5ml of TRI-reagent was added. RNA was extracted following the normal protocol of the kit. The section 5.2 explain what we following did. The RNAs were at least depleted from the massively Ribosomal RNA component by applying the RiboMinus Mouse/Human kit (Invitrogen).

5.5 SOLEXA TECHNOLOGY

The Solexa GAIIx is a second-generation sequencing technology. It's a so called Polymerase-based sequence-by-synthesis reaction which uses a small 'flow cell' to immobilize, amplify and sequence up to 250 million molecules at once. Single-end fragments were sequenced.
RNA was first fragmented and then retro-transcribed in cDNA; through adapters, single molecules of cDNA template were immobilized on support and then amplified to create Illumina libraries. Single-stranded, adapter-ligated fragments are bound to the surface of the flow cell exposed to reagents for polyermase-based extension. Priming occurs as the free/distal end of a ligated fragment "bridges" to a complementary oligo on the surface. Repeated denaturation and extension result in localized amplification of single molecules in millions of unique locations across the flow cell surface. A flow cell containing millions of unique clusters was loaded into the sequencer for automated cycles of extension and imaging.
The first cycle of sequencing consists first of the incorporation of a single fluorescent nucleotide, followed by high resolution imaging of the entire flow cell. These images represent the data collected for the first base. Any signal above background identifies thephysical location of a cluster, and the fluorescent emission identifies which of the four bases was incorporated at that position . This cycle is repeated, one base at a time, generating a series of images each representing a single base extension at a specific cluster. Base calls are derived with an algorithm that identifies the emission color over time.

5.6 TABLES

5.6.1 INTHERPHASIC CARs BELONGING TO snoRNA

	Genomic position				
Chr	**Start**	**End**	**Description**	**Lenght**	**ENSEMBL Transcript ID**
1	28835071	28835274	SNORA73B	204	ENSG00000200087
1	28906276	28906405	SNORA61	130	ENSG00000207311
1	28906893	28907024	SNORA44	132	ENSG00000207314
1	28907432	28907566	SNORA16A	135	ENSG00000207070
1	40033046	40033180	SNORA55	135	ENSG00000201457
1	45242162	45242265	SNORD46	104	ENSG00000200913
1	76252757	76252835	SNORD45C	79	ENSG00000206620
1	76253574	76253657	SNORD45A	84	ENSG00000207241
1	76255162	76255233	SNORD45B	72	ENSG00000201487
1	93306276	93306408	SNORA66	133	ENSG00000207523
1	155889700	155889836	SNORA42	137	ENSG00000207475
1	155895749	155895877	SCARNA4	129	ENSG00000252808
1	173833507	173833583	SNORD47	77	ENSG00000202394
1	173833966	173834043	SNORD80	78	ENSG00000201692
1	173835106	173835166	SNORD44	61	ENSG00000206607
1	173835772	173835853	SNORD76	82	ENSG00000200016
1	175937534	175937676	SCARNA3	143	ENSG00000252906
1	235291118	235291252	SNORA14B	135	ENSG00000207181
2	10586840	10586975	SNORA80B	136	ENSG00000206633
2	86362993	86363129	SNORD94	137	ENSG00000208772
2	101889398	101889511	SNORD89	114	ENSG00000212283
2	207026952	207027083	SNORA41	132	ENSG00000207406
2	232320511	232320647	SNORA75	137	ENSG00000206885
2	234184373	234184648	SCARNA5	276	ENSG00000252010
2	234197322	234197586	SCARNA6	265	ENSG00000251791
3	12881811	12881949	SNORA7A	139	ENSG00000207496
3	39449880	39450030	SNORA6	151	ENSG00000206760
3	93465527	93465665	SNORA8	178	ENSG00000207304
3	129116053	129116191	SNORA7B	139	ENSG00000207088
3	160232695	160233024	SCARNA7	330	ENSG00000238741
3	186504464	186504641	SNORA81	178	ENSG00000221420
4	53579416	53579537	SNORA26	122	ENSG00000212588
4	119200345	119200475	SNORA24	131	ENSG00000206823
5	82360023	82360156	SCARNA18	134	ENSG00000238835
5	111497182	111497314	SNORA13	133	ENSG00000238363
5	138614470	138614667	SNORA74A	198	ENSG00000200959
5	172447731	172447931	SNORA74B	201	ENSG00000212402

6	31590856	31590987	SNORA38	132	ENSG00000200816
6	133137941	133138016	SNORD100	76	ENSG00000221500
6	133138358	133138487	SNORA33	130	ENSG00000200534
6	160201282	160201413	SNORA20	132	ENSG00000207392
7	45024977	45025109	SNORA9	133	ENSG00000206942
7	45143948	45144081	SNORA5A	134	ENSG00000206838
7	45144505	45144641	SNORA5C	137	ENSG00000201772
8	33370993	33371096	SNORD13	104	ENSG00000239039
8	99054314	99054445	SNORA72	132	ENSG00000207067
9	19063654	19063784	SCARNA8	131	ENSG00000251733
9	95054743	95054875	SNORA84	133	ENSG00000239183
9	130210780	130210909	SNORA65	130	ENSG00000201302
9	136216251	136216325	SNORD24	75	ENSG00000206611
9	139620556	139620691	SNORA43	136	ENSG00000199437
11	811681	811814	SNORA52	134	ENSG00000199785
11	2985001	2985123	SNORA54	123	ENSG00000207008
11	8705774	8705903	SNORA3	130	ENSG00000200983
11	8706986	8707116	SNORA45	131	ENSG00000212607
11	9450320	9450501	SNORA23	182	ENSG00000201998
11	62432894	62433042	SNORA57	149	ENSG00000206597
11	62622484	62622555	SNORD27	72	ENSG00000200851
11	93463679	93463812	SNORA25	134	ENSG00000207112
11	93465170	93465299	SNORA1	130	ENSG00000206834
11	93466632	93466763	SNORA18	132	ENSG00000207145
11	93468277	93468402	SNORA40	126	ENSG00000210825
11	122929617	122929703	SNORD14D	87	ENSG00000207118
11	122930043	122930130	SNORD14C	88	ENSG00000202252
12	6619388	6619717	SCARNA10	330	ENSG00000239002
12	7076500	7076769	SCARNA12	270	ENSG00000238795
12	49048165	49048301	SNORA34	137	ENSG00000221491
12	49050431	49050565	SNORA2A	135	ENSG00000206612
13	27829538	27829663	SNORA27	126	ENSG00000207051
13	45911615	45911744	SNORA31	130	ENSG00000199477
14	95999692	95999966	SCARNA13	275	ENSG00000252481
14	103804186	103804311	SNORA28	126	ENSG00000207315
15	66639545	66639680	SCARNA14	136	ENSG00000252712
15	66795581	66795652	SNORD18A	72	ENSG00000200623
16	2012335	2012467	SNORA10	133	ENSG00000206811
16	2012974	2013107	SNORA64	134	ENSG00000207405
16	2205024	2205106	SNORD60	83	ENSG00000206630
16	58582403	58582537	SNORA46	135	ENSG00000207493
17	7478031	7478165	SNORA48	135	ENSG00000209582
17	7481273	7481409	SNORA67	137	ENSG00000207152
17	16343350	16343420	SNORD49A	71	ENSG00000206956
17	16344540	16344612	SNORD65	73	ENSG00000212381

17	18965225	18965440	SNORD3B-1	216	ENSG00000200229
17	18967234	18967449	SNORD3B-2	216	ENSG00000201750
17	19015734	19015949	SNORD3D	216	ENSG00000199663
17	19091329	19091544	SNORD3A	216	ENSG00000202364
17	19093343	19093558	SNORD3C	216	ENSG00000199298
17	37009116	37009247	SNORA21	132	ENSG00000199293
17	62223443	62223512	SNORD104	70	ENSG00000199753
17	75085389	75085575	SCARNA16	210	ENSG00000251790
19	17973397	17973529	SNORA68	133	ENSG00000207166
19	49993222	49993305	SNORD32A	84	ENSG00000201675
19	49993872	49993956	SNORD33	85	ENSG00000199631
19	49994432	49994517	SNORD35A	85	ENSG00000200259
20	2635713	2635844	SNORA51	132	ENSG00000207427
20	37062508	37062641	SNORA71D	134	ENSG00000200354
20	37078013	37078146	SNORA60	134	ENSG00000199266
21	33749496	33749631	SNORA80	136	ENSG00000200792
22	39709824	39709916	SNORD83B	93	ENSG00000209480
22	39711218	39711312	SNORD83A	95	ENSG00000209482
x	153628622	153628756	SNORA70	135	ENSG00000207165

Datas are sorted by chromosome position and then for the enrichment calculated over the biological reference (always $p<0.0001$).

5.6.2 MITOTIC CARs BELONGING TO snoRNA

	Genomic Position				
Chr	**Start**	**End**	**Description**	**Lenght**	**ENSEMBL Transcript ID**
1	28906893	28907024	SNORA44	132	ENSG00000207314
1	93302846	93302940	SNORD21	95	ENSG00000206680
1	173833284	173833360	SNORD81	77	ENSG00000200710
1	173833507	173833583	SNORD47	77	ENSG00000202394
1	173833966	173834043	SNORD80	78	ENSG00000201692
1	173835772	173835853	SNORD76	82	ENSG00000200016
2	203141154	203141241	SNORD70	88	ENSG00000212534
6	133138358	133138487	SNORA33	130	ENSG00000200534
8	33370993	33371096	SNORD13	104	ENSG00000239039
8	56986394	56986460	SNORD54	67	ENSG00000238650
9	130210780	130210909	SNORA65	130	ENSG00000201302
9	136216251	136216325	SNORD24	75	ENSG00000206611
11	17096201	17096291	SNORD14A	91	ENSG00000201784
11	62621376	62621440	SNORD29	65	ENSG00000206653
11	62622093	62622167	SNORD28	75	ENSG00000207437
11	122929617	122929703	SNORD14D	87	ENSG00000207118
13	45911615	45911744	SNORA31	130	ENSG00000199477
16	2205024	2205106	SNORD60	83	ENSG00000206630
16	89627842	89627925	AC092123.1	84	ENSG00000200084
17	8076772	8076905	AC129492.1	134	ENSG00000200463
17	16344540	16344612	SNORD65	73	ENSG00000212381
19	10220433	10220511	SNORD105B	79	ENSG00000238531
19	12817263	12817332	SNORD41	70	ENSG00000209702
19	17973397	17973529	SNORA68	133	ENSG00000207166
19	49993222	49993305	SNORD32A	84	ENSG00000201675
19	49993872	49993956	SNORD33	85	ENSG00000199631
20	47897220	47897309	SNORD12	90	ENSG00000212304

Datas are sorted by chromosome position and then for the enrichment calculated over the biological reference (always $p<0.0001$).

BIBLIOGRAPHY

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 Biol *8*, R148.

Alonso, A., Hasson, D., Cheung, F., and Warburton, P.E. (2010). A paucity of heterochromatin at functional human neocentromeres. Epigenetics Chromatin *3*, 6.

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. Hum Mol Genet *12*, 2711-2721.

Amor, D.J., and Choo, K.H. (2002). Neocentromeres: role in human disease, evolution, and centromere study. Am J Hum Genet *71*, 695-714.

Amor, D.J., Kalitsis, P., Sumer, H., and Choo, K.H. (2004). Building the centromere: from foundation proteins to 3D organization. Trends Cell Biol *14*, 359-368.

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. Mol Cell Biol *22*, 2229-2241.

Bergmann, J.H., Rodriguez, M.G., Martins, N.M., Kimura, H., Kelly, D.A., Masumoto, H., Larionov, V., Jansen, L.E., and Earnshaw, W.C. (2010). Epigenetic engineering shows H3K4me2 is required for HJURP targeting and CENP-A assembly on a synthetic human kinetochore. EMBO J.

Bieda, M., Xu, X., Singer, M.A., Green, R., and Farnham, P.J. (2006). Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. Genome Res *16*, 595-605.

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

Black, B.E., Jansen, L.E., Maddox, P.S., Foltz, D.R., Desai, A.B., Shah, J.V., and Cleveland, D.W. (2007). Centromere identity maintained by nucleosomes assembled with histone H3 containing the CENP-A targeting domain. Mol Cell *25*, 309-322.

Blower, M.D., and Karpen, G.H. (2001). The role of Drosophila CID in kinetochore formation, cell-cycle progression and heterochromatin interactions. Nat Cell Biol *3*, 730-739.

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

Bonaccorsi, S., Gatti, M., Pisano, C., and Lohe, A. (1990). Transcription of a satellite DNA on two Y chromosome loops of Drosophila melanogaster. Chromosoma *99*, 260-266.

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

Brown, M.T. (1995). Sequence similarities between the yeast chromosome segregation protein Mif2 and the mammalian centromere protein CENP-C. Gene *160*, 111-116.

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 Res *19*, 778-784.

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.

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 Biol *7*, R91.

Carone, D.M., Longo, M.S., Ferreri, G.C., Hall, L., Harris, M., Shook, N., Bulazel, K.V., Carone, B.R., Obergfell, C., O'Neill, M.J.*, et al.* (2009). A new class of retroviral and satellite encoded small RNAs emanates from mammalian centromeres. Chromosoma *118*, 113-125.

Carroll, C.W., and Straight, A.F. (2006). Centromere formation: from epigenetics to self-assembly. Trends Cell Biol *16*, 70-78.

Castillo, A.G., Mellone, B.G., Partridge, J.F., Richardson, 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 Genet *3*, e121.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G.*, et al.* (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science *308*, 1149-1154.

Cheng, Z., Dong, F., Langdon, T., Ouyang, S., Buell, C.R., Gu, M., Blattner, F.R., and Jiang, J. (2002). Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. Plant Cell *14*, 1691-1704.

Choo, K.H. (1997). Centromere DNA dynamics: latent centromeres and neocentromere formation. Am J Hum Genet *61*, 1225-1233.

Chueh, A.C., Northrop, E.L., Brettingham-Moore, K.H., Choo, K.H., and Wong, L.H. (2009). LINE retrotransposon RNA is an essential structural and functional epigenetic component of a core neocentromeric chromatin. PLoS Genet *5*, e1000354.

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

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

Dej, K.J., and Orr-Weaver, T.L. (2000). Separation anxiety at the centromere. Trends Cell Biol *10*, 392-399.

Diaz, M.O., Barsacchi-Pilone, G., Mahon, K.A., and Gall, J.G. (1981). Transcripts from both strands of a satellite DNA occur on lampbrush chromosome loops of the newt Notophthalmus. Cell *24*, 649-659.

du Sart, D., Cancilla, M.R., Earle, E., Mao, J.I., Saffery, R., Tainton, K.M., Kalitsis, P., Martyn, J., Barry, A.E., and Choo, K.H. (1997). A functional neo-centromere formed through activation of a latent human centromere and consisting of non-alpha-satellite DNA. Nat Genet *16*, 144-153.

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.

Earnshaw, W., Bordwell, B., Marino, C., and Rothfield, N. (1986). Three human chromosomal autoantigens are recognized by sera from patients with anti-centromere antibodies. J Clin Invest *77*, 426-430.

Epstein, L.M., Mahon, K.A., and Gall, J.G. (1986). Transcription of a satellite DNA in the newt. J Cell Biol *103*, 1137-1144.

Erhardt, S., Mellone, B.G., Betts, C.M., Zhang, W., Karpen, G.H., and Straight, A.F. (2008). Genome-wide analysis reveals a cell cycle-dependent mechanism controlling centromere propagation. J Cell Biol *183*, 805-818.

Felsenstein, K.M., and Emmons, S.W. (1987). Structure and evolution of a family of interspersed repetitive DNA sequences in Caenorhabditis elegans. J Mol Evol *25*, 230-240.

Ferreri, G.C., Marzelli, M., Rens, W., and O'Neill, R.J. (2004). A centromere-specific retroviral element associated with breaks of synteny in macropodine marsupials. Cytogenet Genome Res *107*, 115-118.

Foltz, D.R., Jansen, L.E., Bailey, A.O., Yates, J.R., 3rd, 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.

Foltz, D.R., Jansen, L.E., Black, B.E., Bailey, A.O., Yates, J.R., 3rd, and Cleveland, D.W. (2006). The human CENP-A centromeric nucleosome-associated complex. Nat Cell Biol *8*, 458-469.

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 hMis18alpha, hMis18beta, and M18BP1. Dev Cell *12*, 17-30.

Fukagawa, T., Nogami, M., Yoshikawa, M., Ikeno, M., Okazaki, T., Takami, Y., Nakayama, T., and Oshimura, M. (2004). Dicer is essential for formation of the heterochromatin structure in vertebrate cells. Nat Cell Biol *6*, 784-791.

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 J *18*, 4196-4209.

Furuyama, T., and Henikoff, S. (2009). Centromeric nucleosomes induce positive DNA supercoils. Cell *138*, 104-113.

Gilbert, N., and Allan, J. (2001). Distinctive higher-order chromatin structure at mammalian centromeres. Proc Natl Acad Sci U S A *98*, 11949-11954.

Gilbert, N., Boyle, S., Fiegler, H., Woodfine, K., Carter, N.P., and Bickmore, W.A. (2004). Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. Cell *118*, 555-566.

Gopalakrishnan, S., Sullivan, B.A., Trazzi, S., Della Valle, G., and Robertson, K.D. (2009). DNMT3B interacts with constitutive centromere protein CENP-C to modulate DNA methylation and the histone code at centromeric regions. Hum Mol Genet *18*, 3178-3193.

Goshima, G., Wollman, R., Goodwin, S.S., Zhang, N., Scholey, J.M., Vale, R.D., and Stuurman, N. (2007). Genes required for mitotic spindle assembly in Drosophila S2 cells. Science *316*, 417-421.

Harrington, J.J., Van Bokkelen, G., Mays, R.W., Gustashaw, K., and Willard, H.F. (1997). Formation of de novo centromeres and construction of first-generation human artificial microchromosomes. Nat Genet *15*, 345-355.

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.

Heard, E. (2005). Delving into the diversity of facultative heterochromatin: the epigenetics of the inactive X chromosome. Curr Opin Genet Dev *15*, 482-489.
Henikoff, S., Ahmad, K., and Malik, H.S. (2001). The centromere paradox: stable inheritance with rapidly evolving DNA. Science *293*, 1098-1102.
Howman, E.V., Fowler, K.J., Newson, A.J., Redward, S., MacDonald, A.C., Kalitsis, P., and Choo, K.H. (2000). Early disruption of centromeric chromatin organization in centromere protein A (Cenpa) null mice. Proc Natl Acad Sci U S A *97*, 1148-1153.
Jaco, I., Canela, A., Vera, E., and Blasco, M.A. (2008). Centromere mitotic recombination in mammalian cells. J Cell Biol *181*, 885-892.
Jansen, L.E., Black, B.E., Foltz, D.R., and Cleveland, D.W. (2007). Propagation of centromeric chromatin requires exit from mitosis. J Cell Biol *176*, 795-805.
Jiang, J., Birchler, J.A., Parrott, W.A., and Dawe, R.K. (2003). A molecular view of plant centromeres. Trends Plant Sci *8*, 570-575.
Kanellopoulou, C., Muljo, S.A., Kung, A.L., Ganesan, S., Drapkin, R., Jenuwein, T., Livingston, D.M., and Rajewsky, K. (2005). Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. Genes Dev *19*, 489-501.
Kato, T., Sato, N., Hayama, S., Yamabuki, T., Ito, T., Miyamoto, M., Kondo, S., Nakamura, Y., and Daigo, Y. (2007). Activation of Holliday junction recognizing protein involved in the chromosomal stability and immortality of cancer cells. Cancer Res *67*, 8544-8553.
Kimura, T., Mills, F.C., Allan, J., and Gould, H. (1983). Selective unfolding of erythroid chromatin in the region of the active beta-globin gene. Nature *306*, 709-712.
Kino, T., Hurt, D.E., Ichijo, T., Nader, N., and Chrousos, G.P. (2010). Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. Sci Signal *3*, ra8.
Lachner, M., and Jenuwein, T. (2002). The many faces of histone lysine methylation. Curr Opin Cell Biol *14*, 286-298.
Lam, A.L., Boivin, C.D., Bonney, C.F., Rudd, M.K., and Sullivan, B.A. (2006). Human centromeric chromatin is a dynamic chromosomal domain that can spread over noncentromeric DNA. Proc Natl Acad Sci U S A *103*, 4186-4191.
Lee, H.R., Neumann, P., Macas, J., and Jiang, J. (2006). Transcription and evolutionary dynamics of the centromeric satellite repeat CentO in rice. Mol Biol Evol *23*, 2505-2520.
Lee, J.S., and Shilatifard, A. (2007). A site to remember: H3K36 methylation a mark for histone deacetylation. Mutat Res *618*, 130-134.
Lehnertz, B., Ueda, Y., Derijck, A.A., Braunschweig, U., Perez-Burgos, L., Kubicek, S., Chen, T., Li, E., Jenuwein, T., and Peters, A.H. (2003). Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. Curr Biol *13*, 1192-1200.
Li, Y.X., and Kirby, M.L. (2003). Coordinated and conserved expression of alphoid repeat and alphoid repeat-tagged coding sequences. Dev Dyn *228*, 72-81.
Liu, S.T., Rattner, J.B., Jablonski, S.A., and Yen, T.J. (2006). Mapping the assembly pathways that specify formation of the trilaminar kinetochore plates in human cells. J Cell Biol *175*, 41-53.
Lo, A.W., Craig, J.M., Saffery, R., Kalitsis, P., Irvine, D.V., Earle, E., Magliano, D.J., and Choo, K.H. (2001a). A 330 kb CENP-A binding domain and altered replication timing at a human neocentromere. EMBO J *20*, 2087-2096.
Lo, A.W., Magliano, D.J., Sibson, M.C., Kalitsis, P., Craig, J.M., and Choo, K.H. (2001b). A novel chromatin immunoprecipitation and array (CIA) analysis identifies a 460-kb CENP-A-binding neocentromere DNA. Genome Res *11*, 448-457.
Maison, C., Bailly, D., Peters, A.H., Quivy, J.P., Roche, D., Taddei, A., Lachner, M., Jenuwein, T., and Almouzni, G. (2002). Higher-order structure in pericentric heterochromatin involves a distinct pattern of histone modification and an RNA component. Nat Genet *30*, 329-334.
Marshall, O.J., Chueh, A.C., Wong, L.H., and Choo, K.H. (2008). Neocentromeres: new insights into centromere structure, disease development, and karyotype evolution. Am J Hum Genet *82*, 261-282.
Martinson, H.G., and True, R.J. (1979). Amino acid contacts between histones are the same for plants and mammals. Binding-site studies using ultraviolet light and tetranitromethane. Biochemistry *18*, 1947-1951.
May, B.P., Lippman, Z.B., Fang, Y., Spector, D.L., and Martienssen, R.A. (2005). Differential regulation of strand-specific transcripts from Arabidopsis centromeric satellite repeats. PLoS Genet *1*, e79.
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.
Montefalcone, G., Tempesta, S., Rocchi, M., and Archidiacono, N. (1999). Centromere repositioning. Genome Res *9*, 1184-1188.

Mourtada-Maarabouni, M., Hedge, V.L., Kirkham, L., Farzaneh, F., and Williams, G.T. (2008). Growth arrest in human T-cells is controlled by the non-coding RNA growth-arrest-specific transcript 5 (GAS5). J Cell Sci *121*, 939-946.

Mourtada-Maarabouni, M., Pickard, M.R., Hedge, V.L., Farzaneh, F., and Williams, G.T. (2009). GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer. Oncogene *28*, 195-208.

Murchison, E.P., Partridge, J.F., Tam, O.H., Cheloufi, S., and Hannon, G.J. (2005). Characterization of Dicer-deficient murine embryonic stem cells. Proc Natl Acad Sci U S A *102*, 12135-12140.

Nagaki, K., Neumann, P., Zhang, D., Ouyang, S., Buell, C.R., Cheng, Z., and Jiang, J. (2005). Structure, divergence, and distribution of the CRR centromeric retrotransposon family in rice. Mol Biol Evol *22*, 845-855.

Nagaki, K., Talbert, P.B., Zhong, C.X., Dawe, R.K., Henikoff, S., and Jiang, J. (2003). Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of Arabidopsis thaliana centromeres. Genetics *163*, 1221-1225.

Nakagawa, H., Lee, J.K., Hurwitz, J., Allshire, R.C., Nakayama, J., Grewal, S.I., Tanaka, K., and Murakami, Y. (2002). Fission yeast CENP-B homologs nucleate centromeric heterochromatin by promoting heterochromatin-specific histone tail modifications. Genes Dev *16*, 1766-1778.

Neumann, P., Yan, H., and Jiang, J. (2007). The centromeric retrotransposons of rice are transcribed and differentially processed by RNA interference. Genetics *176*, 749-761.

O'Neill, R.J., and Carone, D.M. (2009). The role of ncRNA in centromeres: a lesson from marsupials. Prog Mol Subcell Biol *48*, 77-101.

Okada, M., Cheeseman, I.M., Hori, T., Okawa, K., McLeod, I.X., Yates, J.R., 3rd, Desai, A., and Fukagawa, T. (2006). The CENP-H-I complex is required for the efficient incorporation of newly synthesized CENP-A into centromeres. Nat Cell Biol *8*, 446-457.

Ono, M., Yamada, K., Avolio, F., Scott, M.S., van Koningsbruggen, S., Barton, G.J., and Lamond, A.I. (2010). Analysis of human small nucleolar RNAs (snoRNA) and the development of snoRNA modulator of gene expression vectors. Mol Biol Cell *21*, 1569-1584.

Orr, B., and Sunkel, C.E. (2010). Drosophila CENP-C is essential for centromere identity. Chromosoma.

Page, S.L., Earnshaw, W.C., Choo, K.H., and Shaffer, L.G. (1995). Further evidence that CENP-C is a necessary component of active centromeres: studies of a dic(X; 15) with simultaneous immunofluorescence and FISH. Hum Mol Genet *4*, 289-294.

Palmer, D.K., O'Day, K., Trong, H.L., Charbonneau, H., and Margolis, R.L. (1991). Purification of the centromere-specific protein CENP-A and demonstration that it is a distinctive histone. Proc Natl Acad Sci U S A *88*, 3734-3738.

Panchenko, T., and Black, B.E. (2009). The epigenetic basis for centromere identity. Prog Mol Subcell Biol *48*, 1-32.

Pearson, C.G., Yeh, E., Gardner, M., Odde, D., Salmon, E.D., and Bloom, K. (2004). Stable kinetochore-microtubule attachment constrains centromere positioning in metaphase. Curr Biol *14*, 1962-1967.

Peters, A.H., O'Carroll, D., Scherthan, H., Mechtler, K., Sauer, S., Schofer, C., Weipoltshammer, K., Pagani, M., Lachner, M., Kohlmaier, A*., et al.* (2001). Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability. Cell *107*, 323-337.

Petit, P., and Fryns, J.P. (1997). Interstitial deletion 2p accompanied by marker chromosome formation of the deleted segment resulting in a stable acentric marker chromosome. Genet Couns *8*, 341-343.

Pluta, A.F., Cooke, C.A., and Earnshaw, W.C. (1990). Structure of the human centromere at metaphase. Trends Biochem Sci *15*, 181-185.

Pluta, A.F., Mackay, A.M., Ainsztein, A.M., Goldberg, I.G., and Earnshaw, W.C. (1995). The centromere: hub of chromosomal activities. Science *270*, 1591-1594.

Politi, V., Perini, G., Trazzi, S., Pliss, A., Raska, I., Earnshaw, W.C., and Della Valle, G. (2002). CENP-C binds the alpha-satellite DNA in vivo at specific centromere domains. J Cell Sci *115*, 2317-2327.

Richmond, T.J., Rechsteiner, T., and Luger, K. (1993). Studies of nucleosome structure. Cold Spring Harb Symp Quant Biol *58*, 265-272.

Rieder, C.L., and Salmon, E.D. (1994). Motile kinetochores and polar ejection forces dictate chromosome position on the vertebrate mitotic spindle. J Cell Biol *124*, 223-233.

Rieder, C.L., and Salmon, E.D. (1998). The vertebrate cell kinetochore and its roles during mitosis. Trends Cell Biol *8*, 310-318.

Rudd, M.K., and Willard, H.F. (2004). Analysis of the centromeric regions of the human genome assembly. Trends Genet *20*, 529-533.

Rudert, F., Bronner, S., Garnier, J.M., and Dolle, P. (1995). Transcripts from opposite strands of gamma satellite DNA are differentially expressed during mouse development. Mamm Genome *6*, 76-83.

Saffery, R., Sumer, H., Hassan, S., Wong, L.H., Craig, J.M., Todokoro, K., Anderson, M., Stafford, A., and Choo, K.H. (2003). Transcription within a functional human centromere. Mol Cell *12*, 509-516.
Saffery, R., Wong, L.H., Irvine, D.V., Bateman, M.A., Griffiths, B., Cutts, S.M., Cancilla, M.R., Cendron, A.C., Stafford, A.J., and Choo, K.H. (2001). Construction of neocentromere-based human minichromosomes by telomere-associated chromosomal truncation. Proc Natl Acad Sci U S A *98*, 5705-5710.
Saitoh, H., Tomkiel, J., Cooke, C.A., Ratrie, H., 3rd, 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.
Satinover, D.L., Vance, G.H., Van Dyke, D.L., and Schwartz, S. (2001). Cytogenetic analysis and construction of a BAC contig across a common neocentromeric region from 9p. Chromosoma *110*, 275-283.
Schueler, M.G., and Sullivan, B.A. (2006). Structural and functional dynamics of human centromeric chromatin. Annu Rev Genomics Hum Genet *7*, 301-313.
Schuh, M., Lehner, C.F., and Heidmann, S. (2007). Incorporation of Drosophila CID/CENP-A and CENP-C into centromeres during early embryonic anaphase. Curr Biol *17*, 237-243.
Shelby, R.D., Monier, K., and Sullivan, K.F. (2000). Chromatin assembly at kinetochores is uncoupled from DNA replication. J Cell Biol *151*, 1113-1118.
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. J Cell Biol *136*, 501-513.
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 Cancer *29*, 117-129.
Sugimoto, K., Fukuda, R., and Himeno, M. (2000). Centromere/kinetochore localization of human centromere protein A (CENP-A) exogenously expressed as a fusion to green fluorescent protein. Cell Struct Funct *25*, 253-261.
Sugimoto, K., Tsutsui, M., AuCoin, D., and Vig, B.K. (1999). Visualization of prekinetochore locus on the centromeric region of highly extended chromatin fibers: does kinetochore autoantigen CENP-C constitute a kinetochore organizing center? Chromosome Res *7*, 9-19.
Sullivan, B.A., Blower, M.D., and Karpen, G.H. (2001). Determining centromere identity: cyclical stories and forking paths. Nat Rev Genet *2*, 584-596.
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. J Cell Biol *127*, 581-592.
Talbert, P.B., Bryson, T.D., and Henikoff, S. (2004). Adaptive evolution of centromere proteins in plants and animals. J Biol *3*, 18.
Topp, C.N., Zhong, C.X., and Dawe, R.K. (2004). Centromere-encoded RNAs are integral components of the maize kinetochore. Proc Natl Acad Sci U S A *101*, 15986-15991.
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. J Struct Biol *140*, 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*, e5832.
Ugarkovic, D. (2005). Functional elements residing within satellite DNAs. EMBO Rep *6*, 1035-1039.
Vafa, O., and Sullivan, K.F. (1997). Chromatin containing CENP-A and alpha-satellite DNA is a major component of the inner kinetochore plate. Curr Biol *7*, 897-900.
Valgardsdottir, R., Chiodi, I., Giordano, M., Cobianchi, F., Riva, S., and Biamonti, G. (2005). Structural and functional characterization of noncoding repetitive RNAs transcribed in stressed human cells. Mol Biol Cell *16*, 2597-2604.
Van Hooser, A.A., Ouspenski, II, 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. J Cell Sci *114*, 3529-3542.
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 *316*, 243-246.
Ventura, M., Weigl, S., Carbone, L., Cardone, M.F., Misceo, D., Teti, M., D'Addabbo, P., Wandall, A., Bjorck, E., de Jong, P.J*., et al.* (2004). Recurrent sites for new centromere seeding. Genome Res *14*, 1696-1703.
Volpe, T., Schramke, V., Hamilton, G.L., White, S.A., Teng, G., Martienssen, R.A., and Allshire, R.C. (2003). RNA interference is required for normal centromere function in fission yeast. Chromosome Res *11*, 137-146.

Volpe, T.A., Kidner, C., Hall, I.M., Teng, G., Grewal, S.I., and Martienssen, R.A. (2002). Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. Science *297*, 1833-1837.

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? Am J Hum Genet *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 *326*, 865-867.

Wang, L.H., Schwarzbraun, T., Speicher, M.R., and Nigg, E.A. (2008). Persistence of DNA threads in human anaphase cells suggests late completion of sister chromatid decatenation. Chromosoma *117*, 123-135.

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

Warburton, P.E., Cooke, C.A., Bourassa, S., Vafa, O., Sullivan, B.A., Stetten, G., Gimelli, G., Warburton, D., Tyler-Smith, C., Sullivan, K.F.*, et al.* (1997). Immunolocalization of CENP-A suggests a distinct nucleosome structure at the inner kinetochore plate of active centromeres. Curr Biol *7*, 901-904.

White, S.A., and Allshire, R.C. (2004). Loss of Dicer fowls up centromeres. Nat Cell Biol *6*, 696-697.

Willard, H.F. (1989). The genomics of long tandem arrays of satellite DNA in the human genome. Genome *31*, 737-744.

Wong, L.H., Brettingham-Moore, K.H., Chan, L., Quach, J.M., Anderson, M.A., Northrop, E.L., Hannan, R., Saffery, R., Shaw, M.L., Williams, E.*, et al.* (2007). Centromere RNA is a key component for the assembly of nucleoproteins at the nucleolus and centromere. Genome Res *17*, 1146-1160.

Wong, L.H., and Choo, K.H. (2004). Evolutionary dynamics of transposable elements at the centromere. Trends Genet *20*, 611-616.

Xiao, F., Widlak, P., and Garrard, W.T. (2007). Engineered apoptotic nucleases for chromatin research. Nucleic Acids Res *35*, e93.

Yan, H., Ito, H., Nobuta, K., Ouyang, S., Jin, W., Tian, S., Lu, C., Venu, R.C., Wang, G.L., Green, P.J.*, et al.* (2006). Genomic and genetic characterization of rice Cen3 reveals extensive transcription and evolutionary implications of a complex centromere. Plant Cell *18*, 2123-2133.

Yan, H., and Jiang, J. (2007). Rice as a model for centromere and heterochromatin research. Chromosome Res *15*, 77-84.

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 U S A *97*, 7266-7271.

Zeitlin, S.G., Patel, S., Kavli, B., and Slupphaug, G. (2005). Xenopus CENP-A assembly into chromatin requires base excision repair proteins. DNA Repair (Amst) *4*, 760-772.

Zeitlin, S.G., Shelby, R.D., and Sullivan, K.F. (2001). CENP-A is phosphorylated by Aurora B kinase and plays an unexpected role in completion of cytokinesis. J Cell Biol *155*, 1147-1157.

Zhong, C.X., Marshall, J.B., Topp, C., Mroczek, R., Kato, A., Nagaki, K., Birchler, J.A., Jiang, J., and Dawe, R.K. (2002). Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. Plant Cell *14*, 2825-2836.