Chromatin landscape of the “dark matter of the genome”: centromeres of *S. cerevisiae* and repeat sequences of *D. melanogaster.*

Kristina Krassovsky

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Steven Henikoff, Chair

Harmit Malik

Charles Laird

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Chromatin landscape of the “dark matter of the genome”: centromeres of S. cerevisiae and repeat sequences of D. melanogaster.

Kristina Krassovsky

Chair of the Supervisory Committee:

Dr. Steven Henikoff

Basic Sciences, FHCRC

The chromatin landscape plays a major role in defining cell phenotypes through transcriptional regulation and specification of the main features of chromosome, such as centromeres and pericentric heterochromatin. Studies of the chromatin landscape so far have been mostly confined to the protein-coding part of the genome. In this work I present a study of the chromatin landscape of two non-coding regions: centromeres in budding yeast Saccharomyces *cerevisiae* and pericentric repeat sequences of the fruit fly Drosophila melanogaster. I have shown that the centromere of budding yeast contains a nucleosome with a special structure, called a hemisome. This finding eliminated other previously proposed models of the centromeric nucleosome and reconciled previous conflicting observations. I also developed a method to quantify enrichment of repeat sequences in Chip-Seq experiments and used it to construct an epigenetic map of heterochromatin in *D. melanogaster* using public datasets Drosophila Genetic Reference Panel (DGRP) and modENCODE. This analysis yielded several

unexpected biologically interesting findings such as preferential association of HP1a protein with transposable elements and depletion of nucleosomes from AT-rich short repeats sequences.

List of Figures

List of Tables

Table of Contents

1 Chapter 1 Introduction

Eukaryotic organisms package DNA into nuclei by wrapping it around histone proteins. Canonical histone proteins H3, H4, H2A and H2B form an octameric complex and 147 bp of DNA wraps this complex in a left handed orientation[1]. This protein/DNA complex is known as a nucleosome. This basic packaging mechanism is modified in various ways to permit or occlude access to DNA of DNA binding proteins involved in different cellular processes. Canonical histones can be replaced by their variants which are associated with specific functions[2]. All histone proteins have flexible ends, called tails that can be modified by covalent binding of different chemical groups, for example methyl or acetyl groups[3]. Modification of different residues on histone tails with chemical groups correlate with different regulatory functions. For example, methylation is often associated with transcriptional repression, while acetylation correlates with activation[4]. Nucleosome positioning, histone variants and histone tail modifications together make up the chromatin landscape of the cell. The chromatin landscape influences transcription and hence the phenotype of the cell. It also demarcates functionally distinct regions of chromosomes such as centromeres, telomeres and pericentric chromatin. The study and description of the chromatin landscape is crucial for our understanding of the inner workings of the cell.

Most studies of the chromatin landscape have focused on the part of the genome containing the protein-coding genes, even though only a small fraction of DNA in the cell encodes protein (1.1% for human [5]). Genomes also encode non-coding RNAs that have a variety of regulatory functions, as well as a role in immune response against foreign nucleic acids[6]. In addition to encoding for protein and regulatory molecules

genomes have elements that are necessary for the stable existence and propagation of the chromosome itself. Specifically, all chromosomes have telomeres that protect their ends, and centromeres that serve as points of attachment of microtubules during mitosis. In addition it has been suggested that pericentric chromatin plays a role in responding to the physical force produced by mitotic spindle[7]. Here I describe the chromatin landscape of the two noncoding regions: centromeres and pericentric heterochromatin.

1.1 Centromeres

Centromeres are specialized loci of the chromosome that serve as points of attachment for microtubules through a protein complex called the kinetochore. There is a considerable variation of centromere forms among eukaryotic organisms[8]. The smallest centromere is found in the budding yeast, where each of the chromosomes has exactly one nucleosome attaching to one microtubule[9]. Other organisms, including humans and flies, have so-called regional centromeres. In these organisms centromeres span kilobases of DNA and contain many nucleosomes and many microtubule attachments [10]. Another distinct form of centromere is the holocentromere, which spans the entire chromosome. While most studies of holocentric centromeres came from *C. elegans*, holocentricity has been found in many diverse organisms that include both plants and animals[11]. Although centromere forms vary considerably, the general mechanism of centromere specification is conserved. Despite the difference in size there is some evidence that the basic building block of the centromere – a centromeric nucleosome – is conserved in all these organisms [12],[13]. There is no particular DNA sequence that marks centromeres. Rather in all the

organisms studied centromeric nucleosomes are marked by incorporation of specialized histone variant cenH3 in place of canonical histone H3. The centromere is thus defined epigenetically by the presence of cenH3, rather than genetically by DNA sequence. Budding yeast and other members of Saccharomyces genus maintain some genetic definition component, and in that regard they differ from all other organisms. All 16 chromosomes of the budding yeast contain a DNA sequence, the Centromere DNA Element (CDE) that is further divided into parts CDEI, CDEII and CDEIII [14], [15], [16], [17]. The CDEIII sequence is 26 bp long and contains a 6 bp sequence that is identical on all the chromosomes. Trinucleotide CCG from this common sequence has been shown to be essential for centromere function[18], [19]. CDEIII is a binding site of CBF3 protein complex, which is necessary for centromere formation and function[20]. This special feature of budding yeast makes it a good model to study centromeres. Genetic definition of the centromere means that its position on a chromosome is precisely known. Knowledge of the exact position coupled with the ease of making genetic changes in yeast allows for manipulation of centromere and kinetochore components, DNA protection and supercoiling assays and the study of centromere function relative to transcription. Such studies are impossible in other organisms with the current level of technology. While there are differences in the kinetochore proteins and their organization between different organisms [21] there is considerable evidence that the molecular core that connects centromeres to the kinetochore is well-conserved [12]. Conclusions obtained in budding yeast will thus likely be informative for other organisms, including humans.

The presence of the histone variant cenH3 at the centromere is well-established and accepted, but the composition of centromeric nucleosomes has been debated [22]. Proposed models of the centromeric nucleosome include: an octameric particle similar to a canonical nucleosome, with the cenH3 histone variant replacing histone H3 [23], [24],[25]; a mixed octamer where one of H3 molecules is replaced with cenH3[26]; a hemisome – a particle that consists of a single copy of histones CenH3, H4, H2B and H2A [27], [28]; a H3/H4 tetramer[29]; and a hexasome – CenH3/H4 heterotetramer with two copies of non-histone protein HJURP (Scm3 in yeast) [30], [31] (Figure 1.1).

The centromeric nucleosome is the point of contact between chromosome and kinetochore and hence its exact structure is of great interest. The unusual structure of this nucleosome has been proposed to alter higher chromatin folding [32, 33], thus creating a platform for kinetochore assembly. Knowledge of the centromeric nucleosome structure will help to answer fundamental questions regarding the chromatin segregation machinery: how does the kinetochore recognize the centromere; and how is centromere position propagated from one generation to the next?

Two lines of evidence are usually cited in support of an octameric form of centromeric nucleosome: in vitro reconstitution studies of CenH3-containing nucleosomes [24] and sequential chromatin immunoprecipitation (CHIP) of differentially tagged CenH3 histones [23]. In reconstitution experiments, nucleosomes are assembled using salt dialysis: histones and DNA are mixed in the presence of high salt and then dialyzed into lower salt. Particles obtained in such a way were shown to be octamers by measuring the size of DNA protected by nucleosome from micrococcal nuclease digestion and by visualizing it with the help of crystal structure. However it has never been shown that

these particles occur at centromeres in vivo. Recently it has been shown that by reducing the size of DNA fragments used for in vitro assembly it is possible to obtain Cse4 hemisome particles on yeast centromeric DNA as well [34].

Another line of evidence for the octamer model comes from sequential chromatin IP experiments with differentially tagged CenH3 histones. In this experiment CenH3 containing nucleosomes are first pulled down with an antibody to one of the tags, eluted, and then pulled down with antibody to another tag[23]. The conclusion that CenH3 nucleosome contains two molecules of CenH3 histone came from observing enrichment of the second IP signal relative to the control genomic region. However enrichment of the second IP signal at the centromere is expected even in the absence of a second molecule because the background that is present in any IP experiment follows the distribution of input material. This means that if the first IP was successful, the input material for the second IP already follows a non-random distribution and has enrichment at the centromere relative to control region. In the absence of knowledge about the relative efficiency of antibodies and expression levels of tagged histones, it is impossible to establish minimum enrichment in the second IP that is consistent with the presence of a second molecule.

Hemisome model was first proposed for centromeric nucleosomes in Drosophila based on biochemical analysis and measurements of the size by Atomic Force Microscopy (AFM) [28]. Later by use of supercoiling assay budding yeast centromeric nucleosomes were shown to have right handed DNA wrap [27] in contrast to canonical nucleosomes that have left-handed wrap. Right handed DNA wrap is structurally impossible for octameric particle and is only consistent with hemisome or CenH3/H4 tetramer.

Proposed models can be most easily distinguished by measuring two properties of the centromeric nucleosome: length of the DNA wrap and presence or absence of histone H2A (Figure 1.2). In chapter 2 I present a study that measures these properties and establishes the composition of the centromeric nucleosome in budding yeast.

1.2 Heterochromatin in *D. melanogaster*

Another gene poor region of the chromosome region is heterochromatin. Heterochromatin is condensed and appears dark staining upon visualization under microscope[35]. It is located at the pericentric (around the centromere) regions, and while its full function is not established heterochromatin plays a role in formation and function of the centromere[7], [36]. While budding yeast does not have conventional heterochromatin, in higher eukaryotes a large portion of the genome is heterochromatic. Histone tail modifications are responsible for chromatin compaction. Specifically, H3 histone tails are modified by addition of a tri-methyl or di-methyl group to lysine 9 of H3 (H3K9me3). This modification is recognized by HP1 protein, which binds to nucleosomes that contain this modification and compacts them, creating a condensed structure [37]. Detailed studies of the chromatin landscape of heterochromatin have not been carried out so far. This is because a large fraction of heterochromatin is made up of highly repetitive sequences. The total proportion of repeat sequences in eukaryotic genomes is large, for example 44% in humans [38], [39]. Repeat sequences present problems for genome assembly and mapping of sequencing reads [40], which are the required steps for Chip-Seq, the most widely used method for the study of chromatin landscapes. In this work I have developed a method that allows quantification of Chip-Seq signal in repetitive sequences without reliance on the genome assembly. This

method can be applied to other problems involving repeat sequences, for example finding centromere sequences in different organisms.

D. melanogaster is a great model to study the problem of repeat sequences because it has been a model organism for more than 80 years and a large body of knowledge obtained by pre-sequencing methods is available. Such knowledge allows the unique comparison of quantification of sequencing datasets to the results obtained by other methods and allows validation of sequencing technologies.

Heterochromatin in *Drosophila* was first observed by Heitz in 1934 [41] when he applied his perfected technique of chromosome preparation to Drosophila nuclei. He also observed that heterochromatin is depleted in cells from larvae – a unique developmental stage of Drosophila and related insects, which follows the embryonic stage. Later, analysis of DNA renaturation kinetics was widely used to estimate the amount of repetitive sequences in the genome [42]. In these experiments DNA is denatured by heat and then cooled down to the temperature that allows renaturation. The amount of renatured DNA is measured as a function of time, producing what is known as C_0t curve. Because repetitive DNA renatures faster than single copy DNA, genomes with different percentages of repetitive sequences will have different renaturation kinetics. By fitting curves it is possible to precisely estimate the percentage of DNA with different renaturation propensities. In most genomes it was observed that there are two types of repetitive DNA: middle repetitive and highly repetitive [43]. In Drosophila the percentage of highly repetitive DNA sequence was estimated as 10-15% [44]. Later it was shown by the renaturation kinetic method that larval cells are depleted of repetitive sequences[45]. Another method used to study genome composition was

CsCl gradient centrifugation. During centrifugation in a CsCl gradient, DNA fragments separate according to their buoyant density, which depends on their GC content. Repetitive sequences have buoyant densities that are different from single copy DNA. When *Drosophila* DNA is centrifuged in a CsCl gradient it separates into four bands, one major band and three so-called satellite bands[46]. By extracting DNA from satellite bands it was possible not only to estimate the percentage of repeats but also to find specific repeat sequences[47]. It was found that satellite bands contain 11 short repeat sequences (repeat unit of 5-12 bp) and a larger 359 bp sequence.

With the widespread use of high-throughput sequencing, comprehensive *Drosophila* sequencing datasets became available. One is called the Drosophila Genetic Reference Panel (DGRP) [48]. This dataset contains sequencing data from 200 inbred fly lines that were created from wild type populations. Another dataset is modENCODE, a collection of experiments profiling histone tail modifications, histone variants and DNA-binding proteins. So far analysis of these datasets was confined to the assembled part of the genome. Euchromatic microsatellite variability in wildtype fly populations was investigated [49] and combinatorial patterns of histone tail modification in cell lines [50], [51] were studied. In this work I employ genome mapping independent methods to quantify repeated sequences in public datasets from DGRP and modENCODE. I compare the repeat content of Drosophila melanogaster recovered in high-throughput sequencing datasets to that were obtained by classical studies and profile histone tail modifications and HP1 proteins associated with repeat sequences.

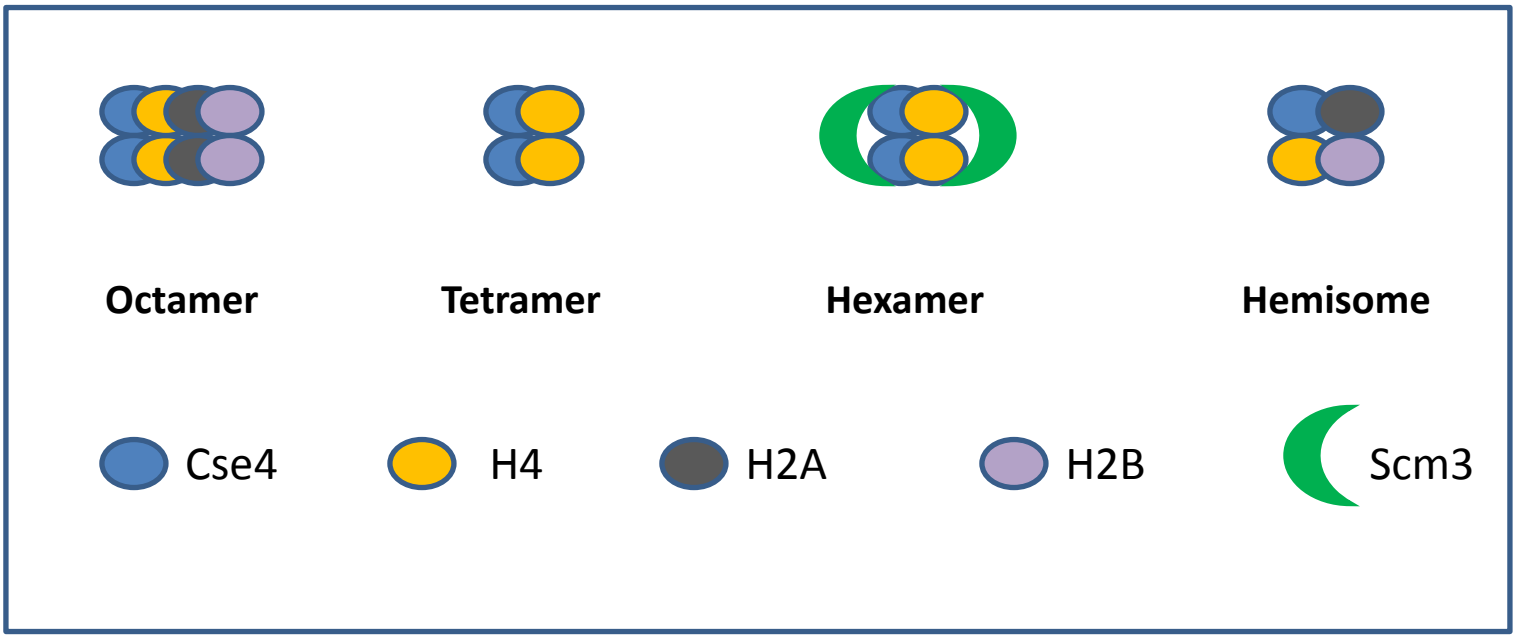


Figure 1.1 Proposed models of the centromeric nucleosome.

Histone composition of centromeric nucleosome is thought to be either an octamer, a particle which contains two copies of each of the histones Cse4, H4, H2A and H2B; a tetramer consisting of two copies of histones Cse4 and H4; a hexamer that includes two copies of histones Cse4 and H4 with the addition of two copies of nonhistone protein Scm3; or a hemisome containing one copy of each of the histones CenH3, H4, H2A and H2B.

	H2A +	H2A -
DNA wrap long	**Octamer**	**Hexasome**
DNA wrap short	**Hemisome**	**Tetramer**

Figure 1.2 Two properties that distinguish proposed models of the centromeric nucleosome.

2 Chapter 2 Tripartite organization of centromeric chromatin in budding yeast[1]

2.1 Introduction

The centromere is the genetic locus that specifies the location of the kinetochore, the complex proteinaceous structure that attaches to spindle microtubules for regular segregation to the poles at mitosis and meiosis [52]. Every chromosome must have one and only one centromere, because acentric and dicentric chromosomes are lost leading to aneuploidy and cell death. This stringent requirement for a single centromere has led to the expectation that centromeres would be defined by DNA sequence, and indeed this is the case in the budding yeast, Saccharomyces cerevisiae, where each of the 16 centromeres consists of a ~120-bp sequence that is entirely responsible for centromere specification [16]. However, a common feature of centromeres in multicellular eukaryotes is that they are embedded in highly repetitive satellite DNA, which has made their molecular study difficult [53]. Furthermore, the existence of neocentromeres that entirely lack centromeric satellites indicates that specific DNA sequences are not necessary for centromere function [54].

Despite the fundamental differences between budding yeast and multicellular eukaryotes with respect to sequence determinants of centromere identity, there are common protein determinants. Most important is the histone variant, CenH3 (CENP-A in mammals and Cse4 in yeast) which replaces histone H3 in centromeric nucleosomes

[1] This chapter appeared previously as Krassovsky, K., J. G. Henikoff and S. Henikoff (2012). "Tripartite organization of centromeric chromatin in budding yeast." Proc Natl Acad Sci U S A 109(1): 243-248.

and is essential for recruitment of the other structural components of the kinetochore [55]. It has been previously shown that the Cse4 nucleosome wraps DNA in a right-handed orientation [27, 56], consistent with in vivo observations of heterotypic tetrameric nucleosomes in Drosophila [57] and humans [58]. However, several studies have shown that CenH3 nucleosomes are left-handed octamers in vitro [23, 59-62].

To help reconcile these findings, we have mapped yeast centromeric particles at single base-pair resolution [63]. This revealed that the Centromere DNA Element (CDE) is well protected in intact particles that also contain H2A, but is preferentially cleaved internally at sites of binding for sequence-specific factors, termed CDEI and CDEIII, where we mapped distinct particles that respectively correspond to the known binding sites for the Cbf1 protein and the CBF3 complex [64]. Using a yeast strain containing multiple copies of Cse4, we found that Cse4-containing particles incorporate at canonical nucleosome positions throughout the genome, and are enriched at sites of rapidly turning over nucleosomes. The existence of two Cse4 nucleosomal species, a stable particle with a single DNA wrap at centromeres and an unstable octamer in chromosome arms supports a general model in which unstable CenH3 nucleosomes are rapidly turned over on chromosome arms to maintain one and only one centromere per chromosome.

2.2 Results

2.2.1 Cse4-containing chromatin particles map precisely to functional centromeres

We performed MNase digestion of crude yeast nuclei from log-phase cells grown in rich medium according to a standard protocol [65], except that we included a needle extraction step to gently but thoroughly solubilize MNase-digested chromatin [63]. This procedure resulted in essentially complete recovery of MNase nucleosome ladder DNA

(Figure 2.1A) and solubilization of most of Cen3 by qPCR (Figure 2.8). In the experiment shown in Figure 2.1, we used MNase digestion times of 2.5', 5', 10' and 20', and followed this by ChIP of FLAG-Cse4. We applied a modified protocol for Solexa library preparation that results in recovery and sequencing of particles down to ~25 bp [63]. After paired-end library preparation and sequencing on an Illumina Hi-Seq 2000 instrument, we obtained on average ~45 million mapped paired-end reads for solubilized chromatin (the input for ChIP). These showed two prominent size features: a broad distribution of fragments ranging in size from ~20-80 bp with a peak at ~30 bp, and a narrow distribution corresponding to nucleosome-sized DNA fragments (Figure 2.1B). The nucleosomal peaks showed stepwise reductions in average size as expected for an MNase series, ranging from 163 bp (2.5') to 151 bp (20'). For the ChIP material, we obtained on average ~23 million mapped paired-end reads. These showed a broader size distribution with indistinct nucleosome-sized peaks, and a broad peak at ~50 bp for 10' and 20' digestions (Figure 2.1C). A sharp peak at ~90-bp seen in the 10’ and 20’ samples for both the soluble chromatin input and the ChIP is attributable to internal cleavage of canonical nucleosomes [63]. In the analyses described below, we used all mapped fragments regardless of size.

We first mapped the ratio of Cse4/Input using as metric the paired-end read count density [63]. For all 16 chromosomes, the maximum Cse4/Input occupancy was over the centromere, consistent with a previous genome-wide mapping study [66]. However, that study used cross-linking and sonication prior to immunoprecipitation (X-ChIP), resulting in a low resolution map, with enrichment extending >300 bp to either side of the center of functional centromeres (Figure 2.9). In striking contrast, our native

chromatin mapping of Cse4/Input at a centromere (Cen4) shows it to be confined primarily to the Cen4 CDE (Figure 2.2A and Figure 2.10). We attribute the broad distribution of Cse4 obtained by X-ChIP in part to the large size of sonicated fragments and in part to the cross-linking of centromeric nucleosomes to flanking nucleosomes and other proteins. The nearly precise mapping of Cse4 ChIP material to Cen4 that we observed using native chromatin delimits the span of the Cse4 nucleosome to the functional centromere sequence. Interestingly, the ChIP/Input signal is non-uniform, showing about twice the protection from MNase over CDEII than over CDEI and CDEIII for all samples.

To ascertain the generality of this result, we aligned ChIP/Input profiles for all 16 yeast centromeres around the midpoint of each CDE and averaged the signal over each base pair. This confirmed that the clear pattern seen for Cen4 for all four samples in the MNase series is general, with nearly precise protection of the functional centromere. It also confirms a previous report of precise positioning of MNase-protected particles over all 16 CDEs [50]. The greatest ChIP signal was centered over CDEII, with distinct shoulders on either side corresponding to partial MNase protection of CDEI and CDEIII (Figure 2.2B). CDEI corresponds to the 8-bp consensus sequence for Cbf1, a conserved general transcription factor that is found at a large number of sites throughout the yeast genome, including centromeres [67, 68]. CDEIII corresponds to a 26-bp consensus sequence that is the binding site for CBF3, a multisubunit complex that is specific for budding yeast centromeres [69]. CDEII is conserved for AT-richness and length (from 78-86 bp), but otherwise has no distinguishing features [16]. Owing to the documented presence of Cbf1 and the CBF3 complex at the CDE, we interpret the

higher ChIP signal for Cse4 over CDEII relative to CDEI and CDEIII as strong evidence for a well-positioned Cse4-containing particle precisely over CDEII.

2.2.2 Distinct particles over CDEI and CDEIII flank the Cse4 nucleosome

To obtain independent confirmation our ChIP/Input occupancy mapping of CDEs, we analyzed the size distribution of MNase-protected DNA around centromeres by plotting the length of each mapped fragment on the Y-axis versus the distance of its midpoint location from the midpoint of the CDE on the X-axis (a 'V-plot') [63]. Strikingly, we observed a clear V-shaped pattern centered precisely over the midpoint of the aligned CDEs from all centromeres (Figure 2.3 A). The sharp edges of the V map midpoints of fragments that are cleaved precisely at one edge of the CDE and extend beyond the opposite edge, and the vertex maps the midpoint of fragments that are cleaved precisely on both sides of the CDE (diagramed in Figure 2.3 B). The vertex corresponds to a fragment size of ~120 bp, which is the average size of annotated CDEs (http://www.yeastgenome.org), indicating that the CDE is almost perfectly protected from MNase digestion.

Each diagonal of a V-plot represents a single sharply defined cleavage on one side of a particle and random cleavage on the other side [63]. In the case of CDEs, we observed an additional pair of V-shaped patterns, one over CDEI and the other over CDEIII (Figure 2.3 A). These patterns were generated by cleavages between CDEI and CDEII and between CDEII and CDEIII, respectively. These internal cleavages were too rare to generate detectable numbers of CDEII-only fragments, which implies that each element of the CDE is protected by one of three closely packed particles that block

MNase from cleaving between them. Interestingly, these Vs in the Cse4 ChIPs fade below ~50 bp, suggesting that MNase digestion released these two protected particles from association with Cse4. Consistent with this interpretation, protected fragments of the expected size and in the expected position can be observed in total chromatin. Extrapolation of diagonals to the vertex of each V results in a minimum protected size of ~10 bp directly over CDEI and of ~20 bp directly over CDEIII for both the Cse4 ChIP and the total chromatin (red dotted diagonal lines in Figure 2.3 A). The identification of distinct protected particles over both CDEI and CDEIII in soluble chromatin indicates that the distinct shoulders observed in the density plots (Figure 2.2) represent partial protection by Cbf1 and the CBF3 complex that is independent of association with Cse4.

Centromeric sequences were recovered in the solubilized chromatin fraction used for ChIP at ~2/3 the level of that from total nuclei (Figure 2.4 A-B), as if kinetochore attachment rendered a subset of these sequences relatively insoluble. To investigate this possibility, we sequenced libraries made from pellet-extracted DNA [63] which we found to be enriched ~100-fold for the CDE in the 20' digestion sample relative to the total nuclear DNA (Figure 2.4 A-B). To determine the minimally protected region of the insoluble kinetochore, we displayed the data from the pellet fraction as V-plots (Figure 2.4 C and Figure 2.11 A). The patterns were nearly identical to those for Cse4 ChIP (Figure 2.11 B-C), which demonstrates that the tripartite structure of centromeric chromatin can be observed without ChIP, based exclusively on reduced kinetochore solubility.

2.2.3 Loss of Cbf1 reduces the size and shifts the location of the centromere-protected region

It remained formally possible that a single Cse4-containing particle spans the entire CDE, while Cbf1 protein and the CBF3 complex bind DNA that is exposed on the nucleosomal surface. Surface binding is unlikely considering that both Cbf1 and CBF3 sharply bend DNA [70, 71], and also that Cbf1 binding excludes canonical nucleosomes [63, 68]. To directly test the possibility that the Cse4 nucleosome occupies the entire CDE, we asked whether loss of one of the flanking particles causes the expected loss of protection of the centromeric element that the particle occupies. The CBF3 complex is essential for viability and for Cse4 nucleosome localization, however, Cbf1 null mutations are viable. Previously, Kent and co-workers [68] had performed paired-end DNA sequencing on MNase-protected fragments in both wildtype and cbf1Δ strains. We remapped their raw data to the yeast genome such that all mappable fragment sizes were included, and constructed average centromere density plots. For analysis, we separated the paired-end reads into four size classes: ≤80 bp, 81-110 bp, 111-140 bp and >140 bp. We observed that loss of Cbf1 caused a striking reduction of 22 bp in the median size distribution of fragments that map to the CDE, with increases in ≤80 bp, 81-110 bp and 111-140 bp size classes relative to the >140 bp size class (Figure 2.5 A-B), with no noticeable change in overall occupancy (Figure 2.5 C). We also observed a conspicuous shift of the median fragment center, 10 bp closer to the CDEIII side of CDEII, with encroachment of ≤80 bp particles into CDEI. It is possible that the continued occupancy of CDEI is caused by the presence of other DNA-binding proteins in yeast that bind to the same CACGTG consensus sequence as Cbf1 [72]. The reduction and shift in protection seen in most cells are as expected if Cbf1 protects CDEI in wildtype, but that loss of Cbf1 allows MNase better accessibility despite occupation of CDEI by

other small particles. Alternatively, loss of Cbf1, which helps to exclude H3 nucleosomes in its vicinity [63, 68] might have reduced protection of the CDE, allowing transient occupancy by an H3 nucleosome in a small subset of cells.

2.2.4 Small particles and well-positioned nucleosomes flank the CDE

V-plots revealed moderate enrichment of subnucleosome-sized particles in total chromatin immediately flanking CDEs (Figure 2.3 A). In the Cse4 ChIP material, we also observed strong enrichment of ≤80 bp fragments centered ~50 bp on either side of the CDEs, which were rapidly depleted with increasing MNase digestion (Figure 2.12). Most subnucleosomal particles mapped elsewhere in the yeast genome are relatively stable to MNase digestion [63], which suggests that whatever is protecting CDE-flanking DNA on both sides might span the CDE. The subnucleosomal particles on either side of the CDE are themselves flanked by well-positioned H3 nucleosomes (Figure 2.3 and Figure 2.11). The fact that all centromeres have these positioned nucleosomes at approximately the same distance from the CDE on both sides confirms and extends previous studies showing that centromeres are flanked by phased nucleosomes [65, 73].

2.2.5 H2A is as abundant at centromeres as it is genome-wide

Previous studies have reported depletion of H2A and H2B nucleosomes at yeast centromeres, suggesting that the only histones in the Cse4 nucleosome are Cse4 and H4 [30, 74]. To determine the composition of Cse4 nucleosomes in our chromatin preparations, we performed ChIP of FLAG-tagged H2A followed by paired-end DNA sequencing. If Cse4 nucleosomes were deficient in H2A, then we would expect that normalized counts for the H2A ChIP would be fewer than for the corresponding input DNA over centromeres. Rather, we found that at all 16 centromeres, the H2A ChIP

signal was equal to that for input DNA (Figure 2.6). The precise positioning of centromeric nucleosomes is evident from these profiles, insofar as flanking regions showed wide variations in H2A occupancy, with almost no variation within the CDE region and between centromeres for H2A occupancy over centromeres.

To confirm that our measurements were sufficiently sensitive to detect differences in H2A occupancy, we examined the single nucleosome occupying the Gal4 UAS in cells grown in glucose, which had been shown to be depleted of H2A owing to the high relative abundance of the H2A.Z variant [75]. Our H2A ChIP data showed clear depletion of H2A from this nucleosome relative to neighboring nucleosomes, confirming the sensitivity of our profiling assay (Figure 2.13 A). For further confirmation of the abundance of H2A at centromeres, we similarly analyzed published X-ChIP-chip H2A data from the 6 centromeres represented on the microarray, and obtained a very similar profile to that for our native ChIP-seq data from all 16 centromeres (Figure 2.13 B).

2.2.6 Misincorporated Cse4 particles are enriched at sites of rapid turnover

To directly compare the incorporation of Cse4 to that of a canonical histone, we performed ChIP using a strain expressing both FLAG-tagged H2A and 5-6 copies of Myc-tagged Cse4 (Figure 2.14). Although FLAG-H2A yielded a nucleosomal size distribution expected for the degree of MNase digestion used (~155 bp), DNA from Cse4-Myc nucleosomes immunoprecipitated from the same solubilized chromatin displayed a broader size distribution, with a peak at ~135 bp that is not evident in control ChIPs from single-copy Cse4 strains (Figure 2.7 A). These ~135-bp protected Cse4 particles are phased at canonical nucleosomal positions in highly expressed genes (Figure 2.7 B), indicating that they correspond to mislocalized Cse4

nucleosomes. Notably, their size distribution matches that of partially unwrapped left-handed CenH3 octamers produced in vitro from purified components (Figure 2.15 A) [59, 62]. Therefore, it is likely that excessive amounts of Cse4 led to the formation of conventional left-handed octameric particles in vivo that were deposited as nucleosomes genome-wide, in contrast to the much smaller Cse4 particle confined to the ~80-bp CDEII.

We asked where the larger Cse4 particles in the overproduction strain were assembled by mapping them genome-wide. We found that Cse4-Myc and FLAG-H2A ChIP peaks precisely coincided, which implies that overproduced Cse4-Myc nucleosomes are incorporated in place of H3 nucleosomes genome-wide (Figure 2.7 C). However, the profiles were quantitatively different. For example, by taking the ratio of Cse4-Myc to FLAG-H2A over the SNT1-FEN1 region, which has been previously characterized with respect to its rate of turnover [76], we found that peaks of Cse4 incorporation corresponded closely to sites of rapidly turning-over ("hot") nucleosomes (Figure 2.7 C) and around 5’ ORFs genome-wide (Figure 2.15 B). We also compared Cse4 and H2A abundances to nucleosome turnover rates globally, and found that Cse4 was in general enriched at sites of high turnover, whereas H2A was depleted from these same sites (Pearson's r = 0.52, Figure 2.7 D). Similar findings of Cse4 misincorporation at sites of hot nucleosomes have been reported for strains that overproduce Cse4 as a result of mutations in components of the nucleosome assembly apparatus [77]. Despite the fact that Cse4-Myc is the only Cse4 copy present in this overproducing strain, the misincorporated particles have evidently not formed functional centromeres, as the strain grew normally.

2.3 Discussion

We have used MNase mapping and paired-end sequencing to map Cse4 nucleosomes to centromeres at high resolution. We found that Cse4 nucleosomes are confined to the central ~80-bp CDEII region of functional centromeres, tightly flanked by distinct small particles over CDEI and CDEIII. This confirms that DNA wraps only once around a Cse4-containing core [78], as implied by our previous study showing that Cse4 nucleosomes induce positive supercoiling at functional centromeres in vivo. Our findings are consistent with the observation that singly wrapped CenH3 particles also occupy Drosophila and human centromeres [57, 58], suggesting that this organization is a universal feature of centromeric nucleosomes. Singly wrapped tetrameric nucleosomes are universal for archaea [79], as if centromeres have retained the ancestral nucleosomal organization.

Our native ChIP-seq analysis also showed that H2A is present at all 16 yeast centromeres at the same level as over the rest of the genome. We confirmed this result by analyzing X-ChIP data from a published study [80]. Because that study was performed on formaldehyde crosslinked chromatin by the same laboratory that previously reported deficiency of H2A over centromeres [30], we might attribute the different outcomes to the use of sonication in the first study to fragment and solubilize DNA, versus the use of MNase in the second study [80]. It is possible that sonication caused loss of poorly crosslinked nucleosomes, and differences in the degree of crosslinking of different histones [57] might have resulted in discrepant ChIP efficiencies between them. The excellent concordance between our study using native chromatin and that of Luk and co-workers using crosslinked chromatin confirms that H2A is as

abundant over centromeres as it is over the entire genome, as expected for a centromeric particle containing all four histones. Our findings are consistent with a report that the Mif2 kinetochore-specific protein co-immunopurifies with Cse4, H4, H2A, and H2B, but not detectably to H3, when purified without cross-linking or enzymatic DNA fragmentation [12].

The low recovery of Cse4 nucleosomes isolated using standard MNase digestion protocols that is evident from our work and that of others [50] (Figure 2.16) confirms previous findings of centromere hypersensitivity to MNase digestion in S. pombe [81] and Drosophila [57]. It also raises questions about reports of what appear to be conventional CenH3 octameric nucleosomes isolated from diverse eukaryotes [22, 82], because these particles were extracted under conditions that led to depletion of yeast centromeric chromatin, which we might attribute to catastrophic loss by cleavage within the singly wrapped CenH3 particle. Our finding that non-functional octamer-like Cse4-containing particles are present at non-centromeric sites of high turnover provides a possible alternative explanation for the immunoprecipitation of octameric nucleosomes in some studies [22, 82], but tetramers in others [57, 58]. The ~90% AT-richness of CDEII might be an adaptation to prevent formation of these aberrant particles at functional centromeres, which would explain why octameric Cse4 nucleosomes fail to assemble on CDEs in vitro [23, 61, 83].

Misincorporation of Cse4 at sites of hot nucleosomes have been reported for strains that overproduce Cse4 as a result of mutations in components of the nucleosome assembly apparatus [77]. Cse4 nucleosomes are also enriched at promoters of highly transcribed genes in strains that do not overproduce Cse4 [23, 66]

(Figure 2.15 C), suggesting that high turnover is a normal mechanism for evicting misincorporated CenH3 from chromosome arms [60] (Figure 2.15 E). At centromeres multiple factors would favor incorporation of a single stable tetramer, including exclusion of H3 octamers by Cbf1 [63, 68], the 90% AT-richness of CDEII, which resists assembly of Cse4 octamers [23, 83], and the recruitment of Cse4 by the adjacent CBF3 complex [84]. In multicellular eukaryotes, heterochromatin condensation would help to stabilize CenH3 tetramers by preventing nucleosome turnover [85]. Our detection of two distinct forms of Cse4 particles, one at centromeres and the other on chromosome arms, thus reconciles seemingly conflicting reports of left-handed CenH3 octamers produced in vitro [23, 59-62] and of right-handed wrapping [27] and heterotypic CenH3 tetramers [57, 58] observed in vivo.

2.4 Materials and Methods

Strains used in this study are listed in Table 2-1. Preparation of yeast nuclei, MNase digestion and DNA extraction steps were performed as described [63]. ChIP was performed as described [65].

Paired-end libraries were prepared using either our modified protocol [63] or the standard Illumina protocol, followed by at least 20 rounds of paired-end sequencing in an Illumina Hi-Seq 2000 by the FHCRC Genomics Shared Resource. Data were processed and mapped using Novoalign as described [63]. SRA SRR058444 and SRR058445 data were mapped similarly with Bowtie using default parameters. Solexa data analysis was performed as previously described [63], except that the fraction of mapped reads spanning each base pair was multiplied by the total number of base pairs in the reference sample to give a normalized count for that base pair. To construct V-

plots, a table of fragment midpoint and length pairs was displayed using the scatterplot function of Kaleidograph (version 4.1, Synergy Software, Inc.).

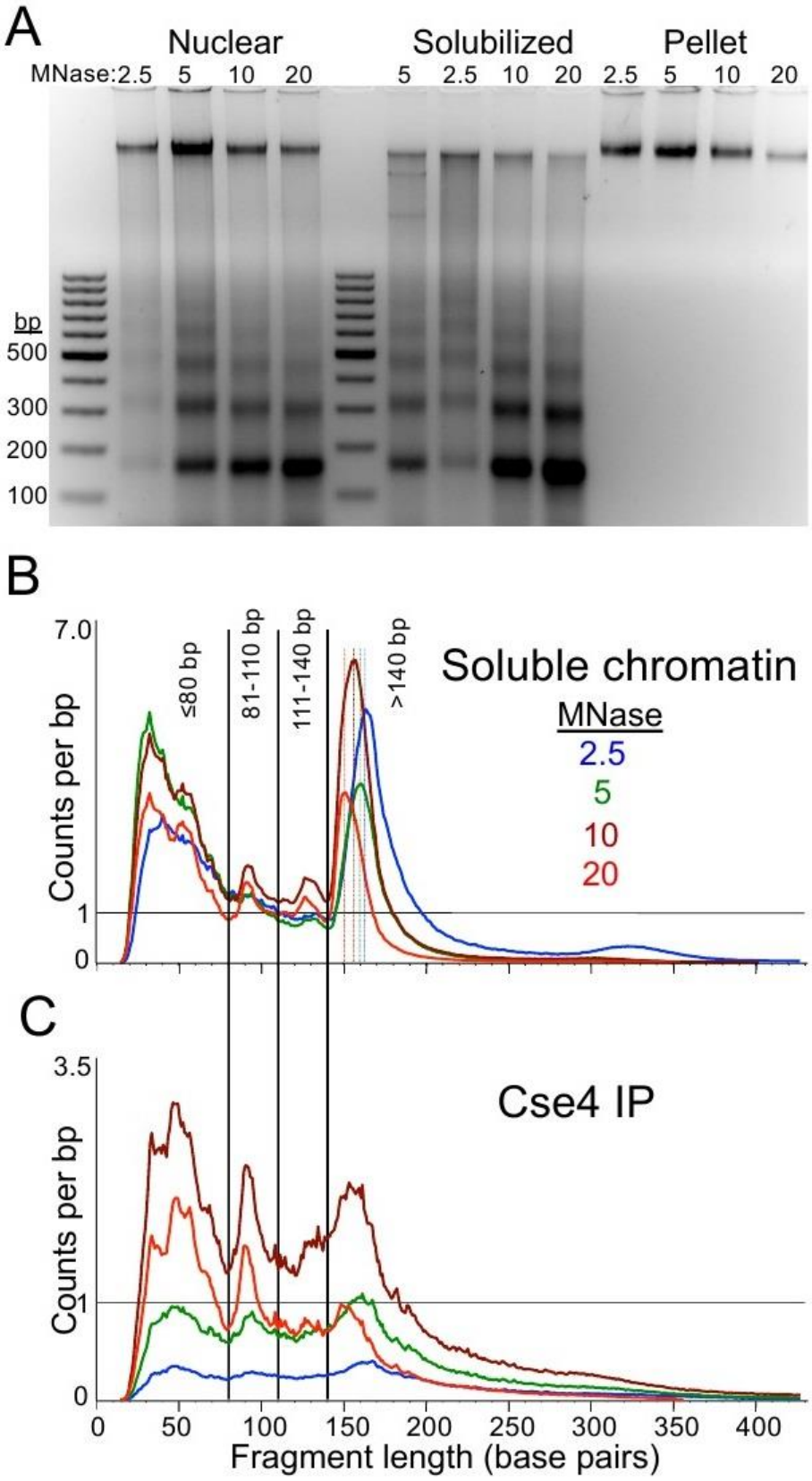

A
Nuclear
Solubilized
Pellet
MNase: 2.5 5 10 20 5 2.5 10 20 2.5 5 10 20
bp
500
300
200
100
B
Soluble chromatin
MNase
2.5
5
10
20
≤80 bp
81-110 bp
111-140 bp
>140 bp
Counts per bp
7.0
1
0
C
Cse4 IP
Counts per bp
3.5
1
0
0 50 100 150 200 250 300 350 400
Fragment length (base pairs)

Figure 2.1. Paired-end sequencing of soluble chromatin and Cse4 ChIP yields distinct size classes of MNase-protected particles.

(A) Agarose gel analysis of MNase time point samples showing DNA from whole nuclei extracted from strain SBY5146 after MNase digestion (Nuclear), DNA from soluble chromatin after needle extraction and pooling of extracts (Solubilized) and DNA from the insoluble residue after solubilization (Pellet). The gradual reduction in size of protected DNA fragments can be seen as jogs in the solubilized chromatin samples loaded out-of-order (5', 2.5', 10' and 20') in the middle set of lanes. Size distributions of mapped paired-end reads for Solubilized chromatin (B) and FLAG-Cse4 ChIP (C), showing the size classes chosen for further analysis.

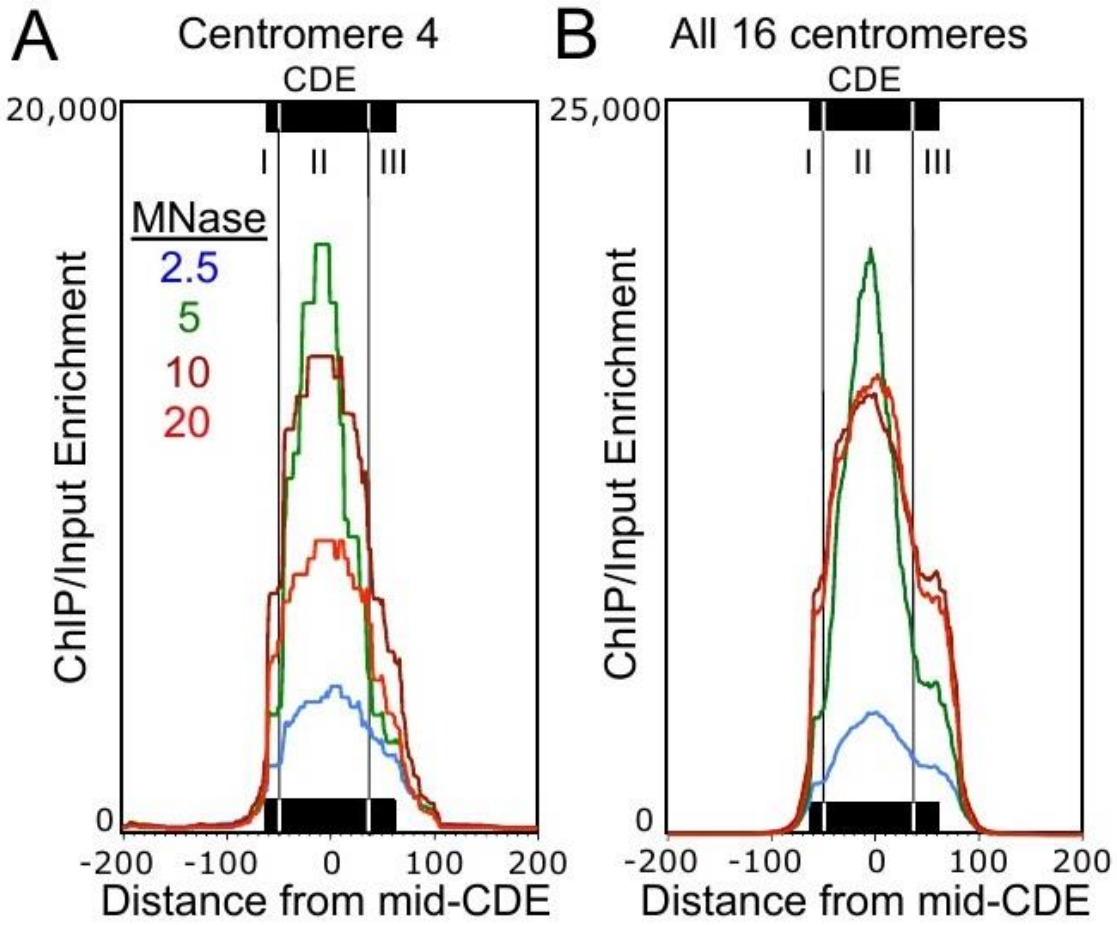


Figure 2.2. ChIP maps the Cse4 nucleosome to CDEII.

Mapped paired-end reads for Cse4 ChIP displayed as ChIP/Input ratios at 1 bp resolution. (A) MNase series of FLAG-Cse4 ChIP for Cen4, (B) Same as A but for all 16 centromeres aligned around the mid-CDE. See also Figure 2.8 for a comparison to Cse4 X-ChIP data.

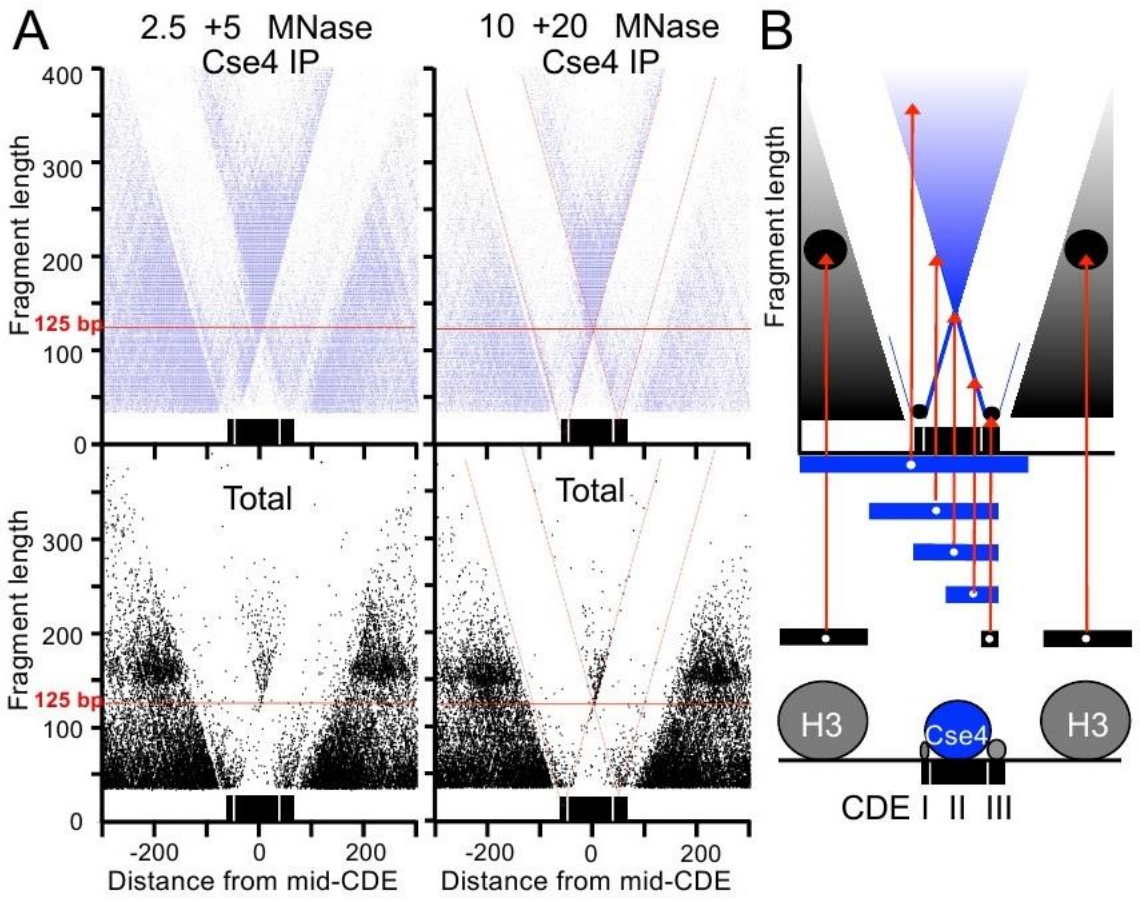


Figure 2.3. V-plot mapping reveals subnucleosome particles at CDEI and CDEIII.

(A) The distance from the midpoint of each fragment to the midpoint of the CDE that maps +/-300 bp of a mid-CDE is represented by a dot, where its position on the X-axis is the distance between its midpoint and the midpoint of the CDE, and its position on the Y-axis represents its fragment length. Cse4 IP (blue) and total soluble chromatin maps are shown for data from 2.5' and 5' MNase digestion time points (left) and from 10’ and 20’ MNase digestion time points (right). The diagonals are marked with red lines in the right panel to show that they intersect at the same positions within CDEI and CDEIII for both Cse4 ChIP and Total soluble chromatin. (B) Diagram showing the position of dots placed on the V-plot (red arrows) for fragments of various sizes and map positions. Below is a model of particles that protect DNA from MNase digestion deduced from the midpoint maps.

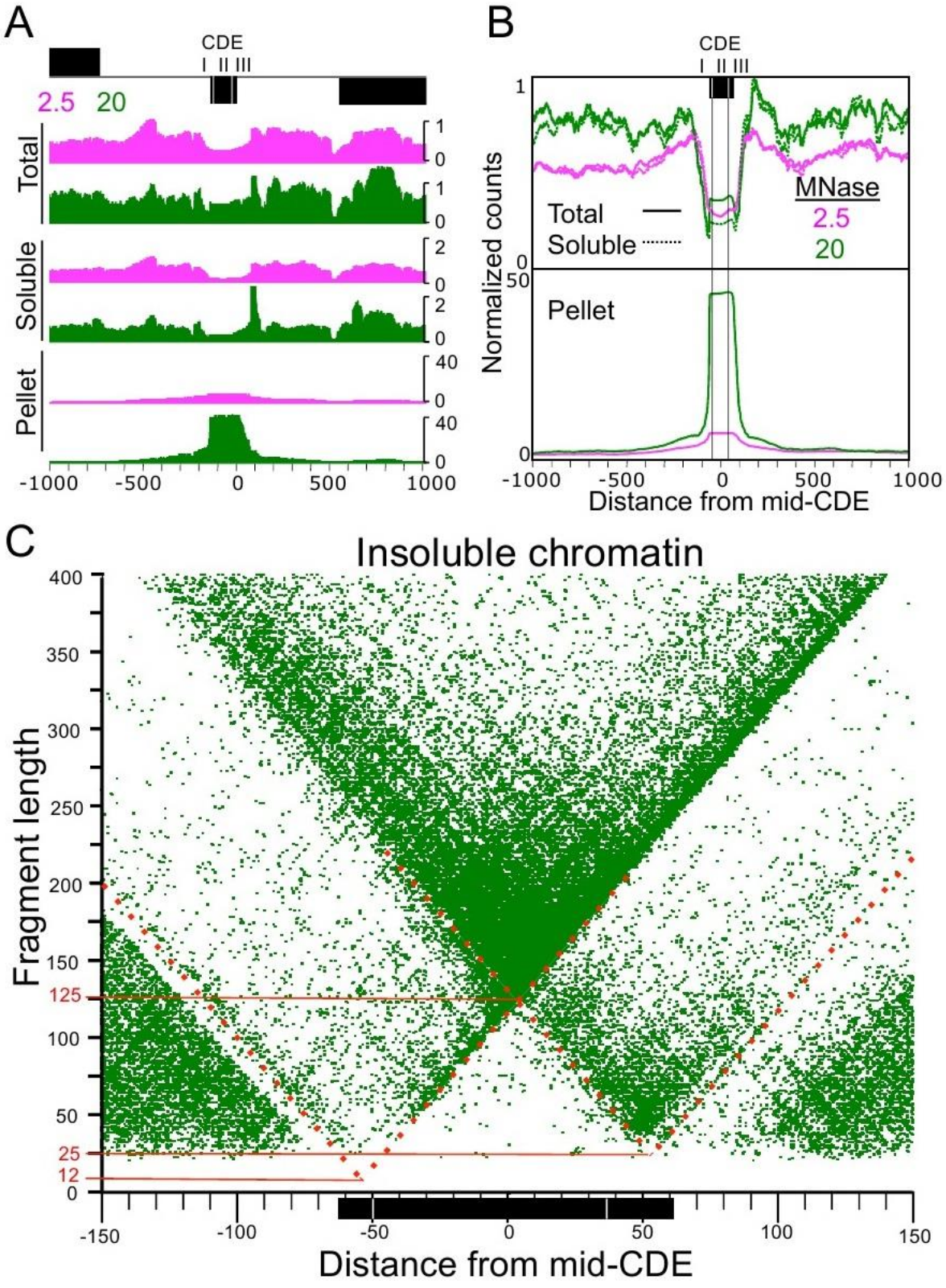


Figure 2.4. Kinetochore enrichment in insoluble chromatin.

(A) Occupancy maps of total nuclei and soluble and insoluble (pellet) fractions for the 2-kb region spanning Centromere 4 showing changes that occur between 2.5’ (magenta) to 20’ (green) MNase digestion. (B) An occupancy map for the same region showing the average of all 16 centromeres aligned at the midpoints of their CDEs. Based on the partitioning of total 20’ MNase-treated nuclear DNA into soluble and insoluble chromatin fractions, we estimate that ~1/3 of centromeric chromatin is insoluble and that insoluble centromeric chromatin is enriched ~100-fold relative to other insoluble chromatin in the pellet. (C) The central 300 bp of the V-plot for insoluble chromatin for all 16 aligned centromeres shows a V-plot that is nearly identical to those seen for ChIP from soluble chromatin (Figure 2.3). Dotted red lines indicate approximate extensions of diagonals, where each intersection maps the midpoint location of the protected region on the X axis and the minimal protected region on the Y-axis (solid red lines).

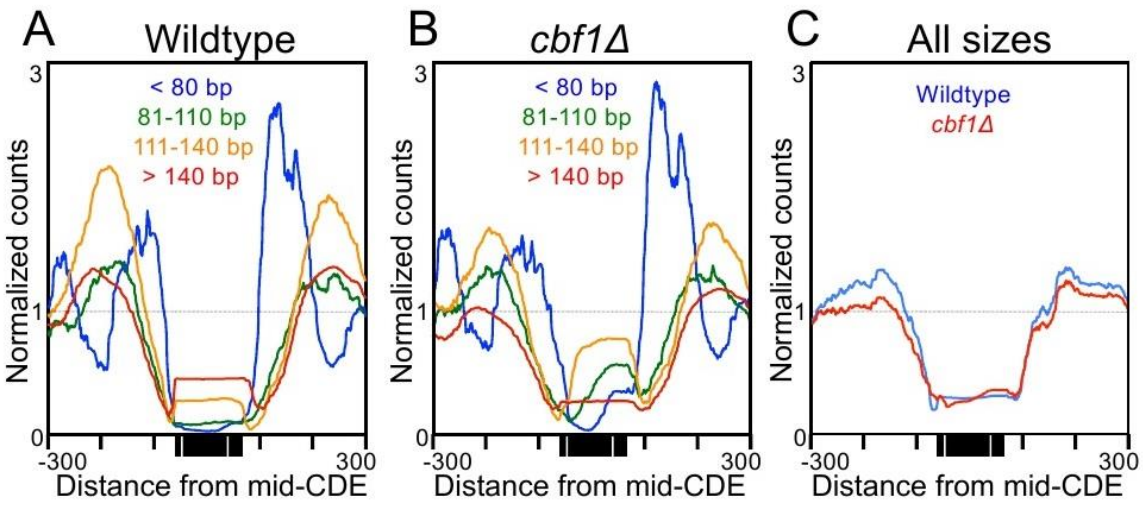


Figure 2.5. Loss of Cbf1 reduces the size and shifts the location of the centromere-protected region

(A-C) Occupancy maps of DNA size classes from MNase-protected particles using data obtained from SRA SRR058444 and SRR058445 (21), comparing wildtype to cbf1Δ strains to illustrate the reduction in size distribution with loss of Cbf1.

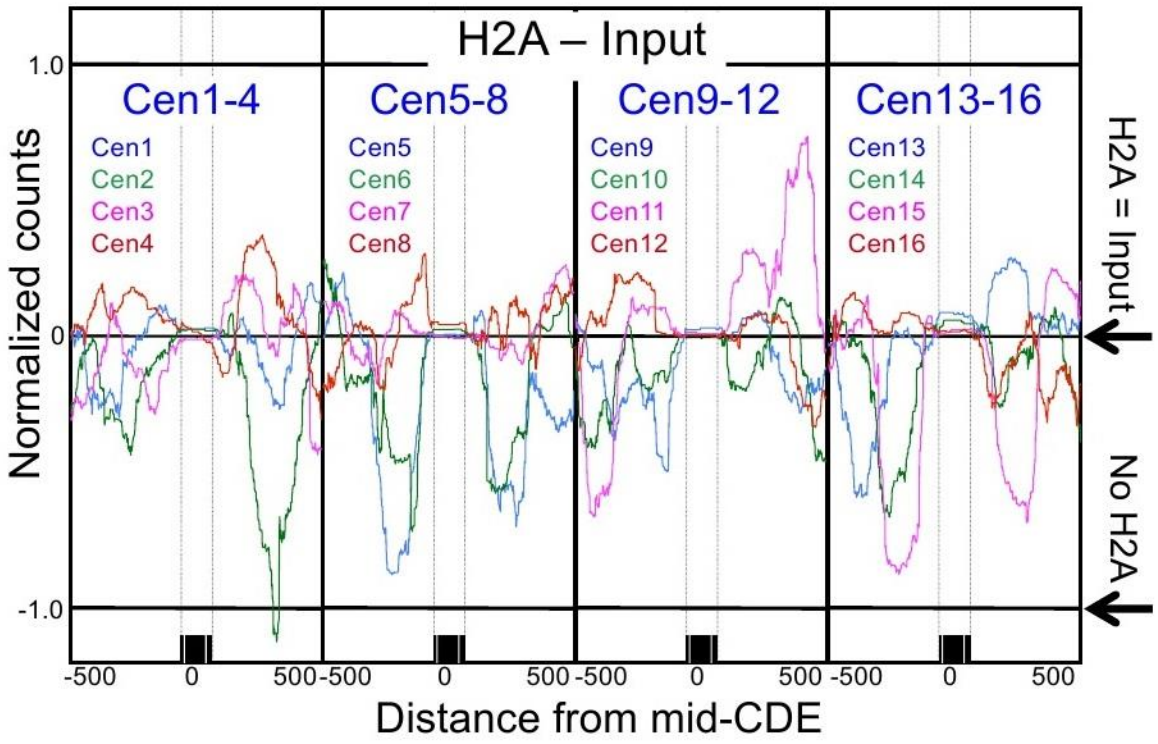


Figure 2.6. The Cse4 nucleosome contains H2A.

ChIP-seq profile of the difference in normalized counts between H2A and Input over all 16 centromeres after ChIP of H2A-FLAG from strain SBY2688, showing that H2A = Input (i.e. H2A-Input = 0). If H2A were absent from centromeres, H2A-Input would equal -1 (No H2A). Fragments larger than 200 bp were excluded to avoid any contribution from dinucleosomes. Similar centromeric results were obtained by plotting published data obtained using crosslinked chromatin followed by ChIP-chip (Figure 2.13B). The expected depletion of H2A at the H2A.Z-enriched Gal4 UAS is confirmed in Figure 2.13A. The variability in flanking nucleosomes is likely due to the variable abundance of H2A.Z in canonical nucleosomes.

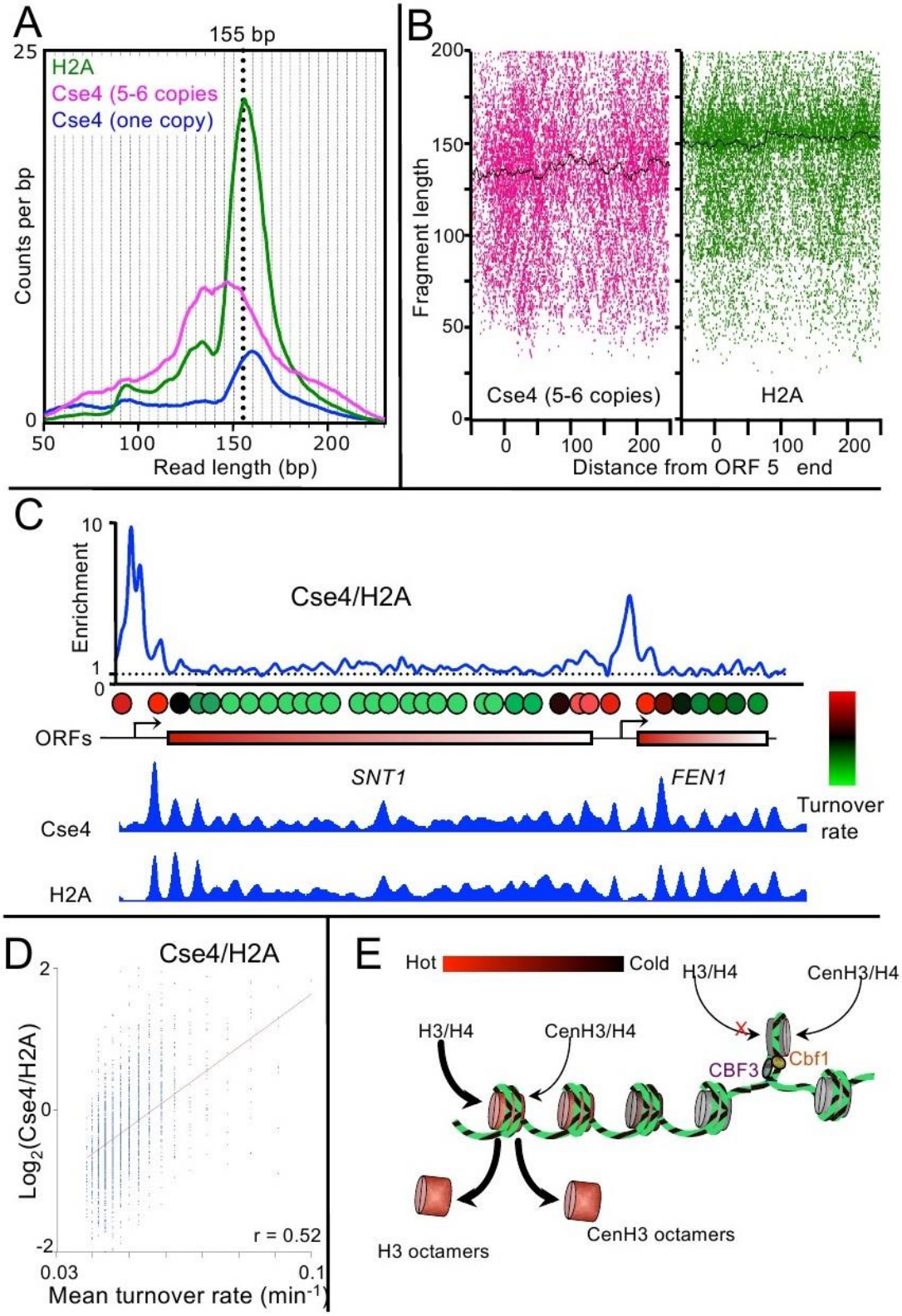

A
155 bp
H2A
Cse4 (5-6 copies
Cse4 (one copy)
Counts per bp
Read length (bp)
B
Fragment length
Cse4 (5-6 copies)
H2A
Distance from ORF 5 end
C
Enrichment
Cse4/H2A
ORFs
SNT1
FEN1
Cse4
H2A
Turnover rate
D
Cse4/H2A
Log2(Cse4/H2A)
r = 0.52
Mean turnover rate (min-1)
E
Hot
Cold
H3/H4
CenH3/H4
CBF3
Cbf1
H3 octamers
CenH3 octamers

Figure 2.7. Overproduced CenH3 particles occupy canonical nucleosome positions and protect ~135 bp.

Size distributions (A) of mapped fragments for FLAG-H2A (magenta) and Cse4-Myc (green) ChIPs, including a FLAG-Cse4 size distribution from a control (blue dotted line). See also Figure 2.15A. (B) V-plots show nucleosomal fragments from the 20 most highly expressed genes aligned at their 5’ ORF ends. Black lines indicate median fragment sizes. (C) Overproduced Cse4-Myc and H2A over the SNT1-FEN1 region showing enrichment of Cse4/H2A (top), an aligned hot nucleosome map reproduced from Ref. [76] (middle) and enrichment of Cse4 and H2A (below). (D) Scatterplot of ChIP signals versus turnover rates for Cse4/H2A. (E) Model in which CenH3 hemisomes are stably held in place by Cbf1 and CBF3, whereas unstable CenH3 octamers are rapidly evicted from chromosome arms and targeted for degradation.

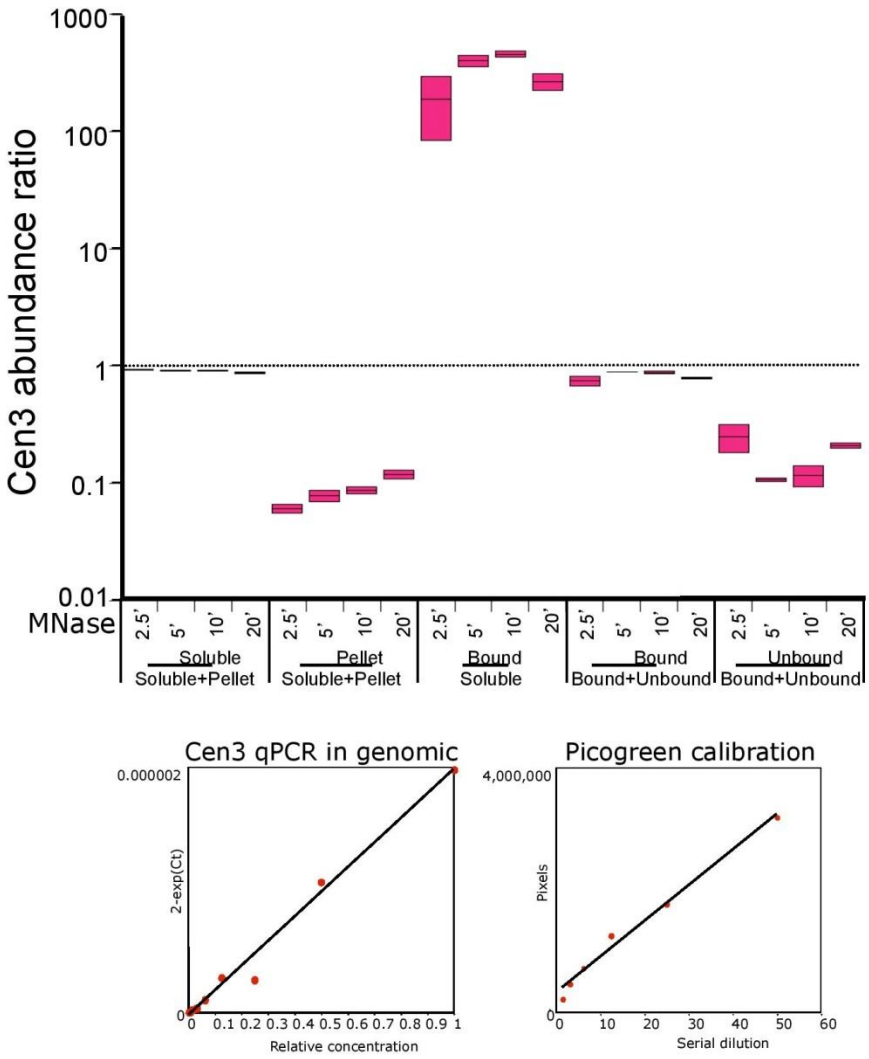


Figure 2.8. qPCR measurement ratios showing relative abundance of Cen3 in chromatin fractions.

PCR primers were designed to amplify the 125-bp Cen3 sequence (ChrIII:114383-114404 and 114487-114508) using a 46oC annealing/extension temperature. Each qPCR measurement was divided by the total DNA level as determined by Picogreen fluorescence in the presence of RNAse A. Ratios are calculated as (qPCR2/Picogreen2)/ (qPCR1/Picogreen1)

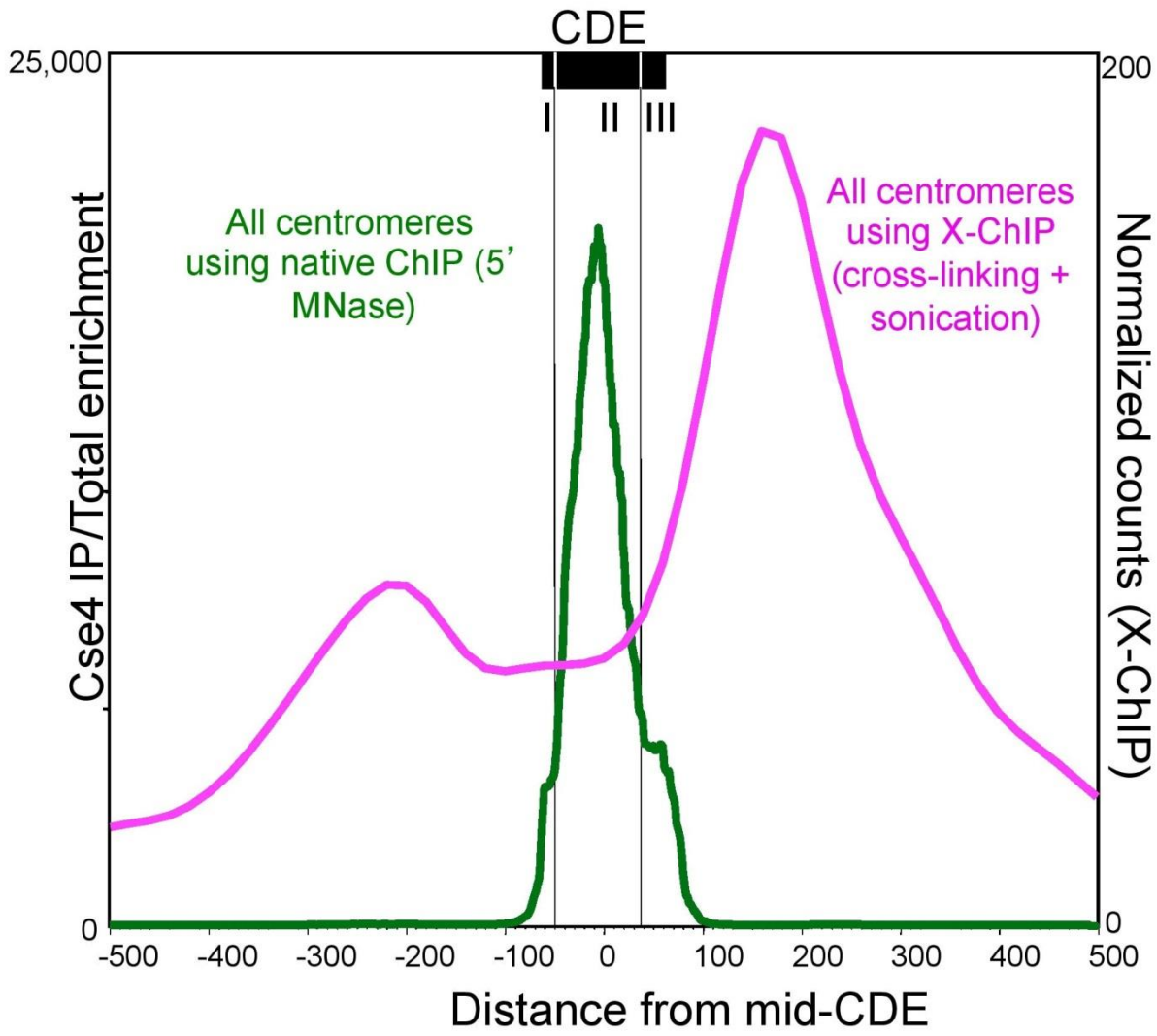


Figure 2.9. Comparison between Native-ChIP and X-ChIP mapping profiles for all 16 aligned yeast centromeres.

A normalized enrichment profile for all 16 yeast centromeres was computed from triplicate Cse4 X-ChIP data obtained from GEO (Accession # GSE13322) [66]. This is shown superimposed over the N-ChIP profiles from Figure 2B. For X-ChIP data, plus and minus reads were offset by the maximum cross-correlation value for normalized counts around the mid-CDE (130 bp), and the average of offset plus and minus reads within each 20 kb window was plotted.

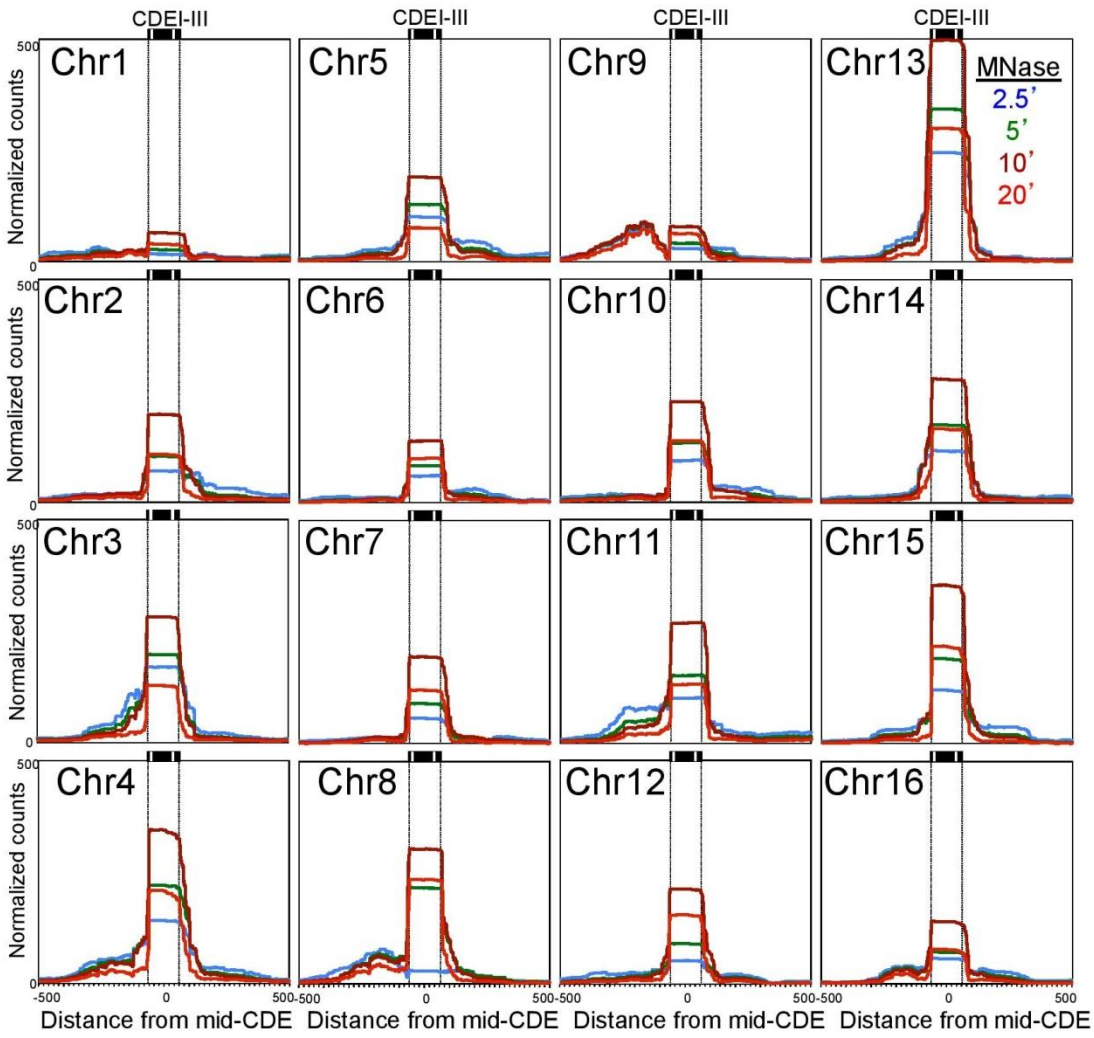
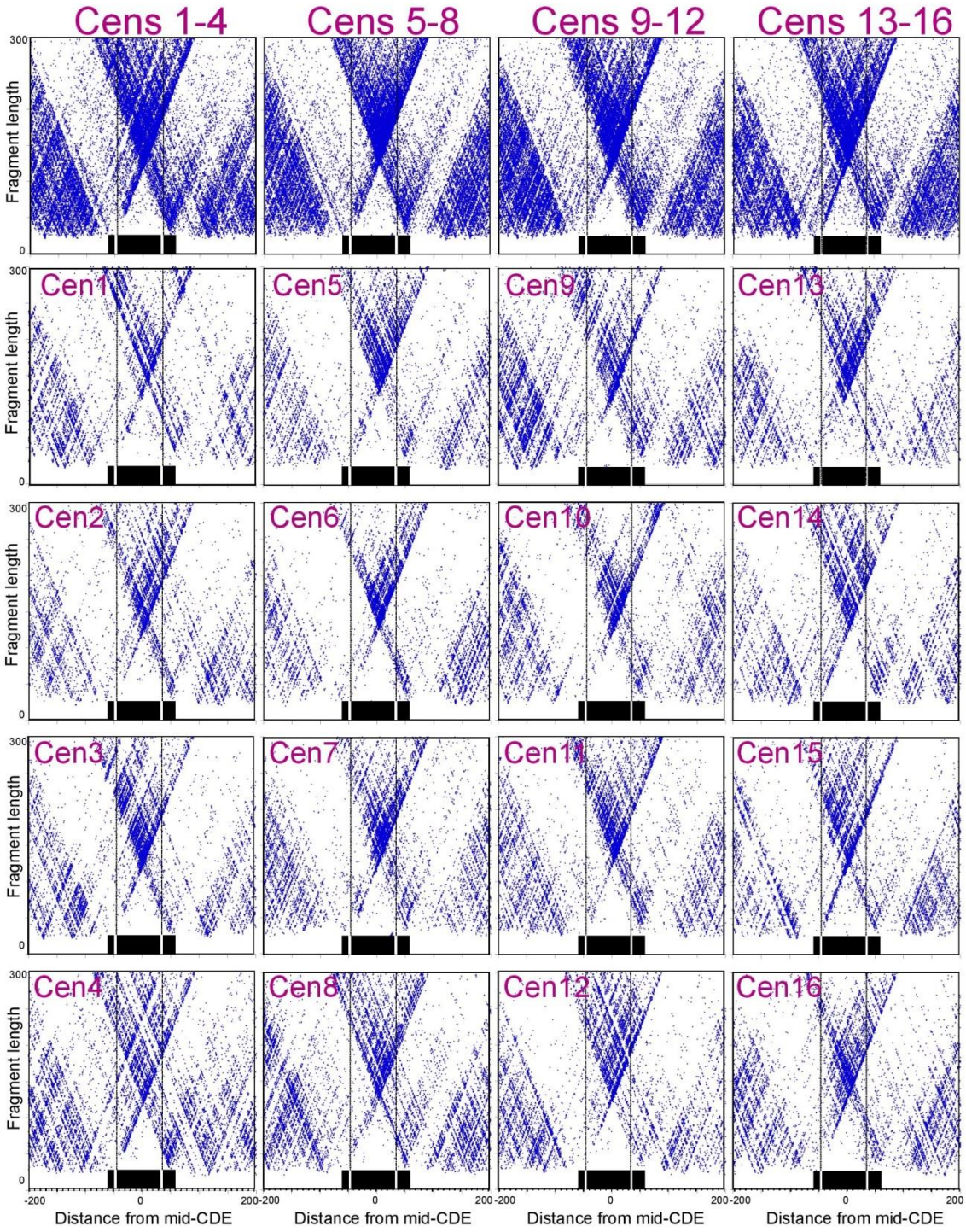


Figure 2.10. Cse4-containing particles are precisely positioned over the CDE

Average Cse4 ChIP profiles for all centromeres and all mapped fragment sizes. The square signal over each CDE increases during digestion up to 10' then generally decreases, whereas the rounded signal to the left of Chr9 shows little if any change in signal, indicating that robustness of the profile regardless of the level of MNase digestion.

A


Cens 1-4
Cens 5-8
Cens 9-12
Cens 13-16
Fragment length
300
0
Cen1
Cen5
Cen9
Cen13
Cen2
Cen6
Cen10
Cen14
Cen3
Cen7
Cen11
Cen15
Cen4
Cen8
Cen12
Cen16
-200
0
200
Distance from mid-CDE

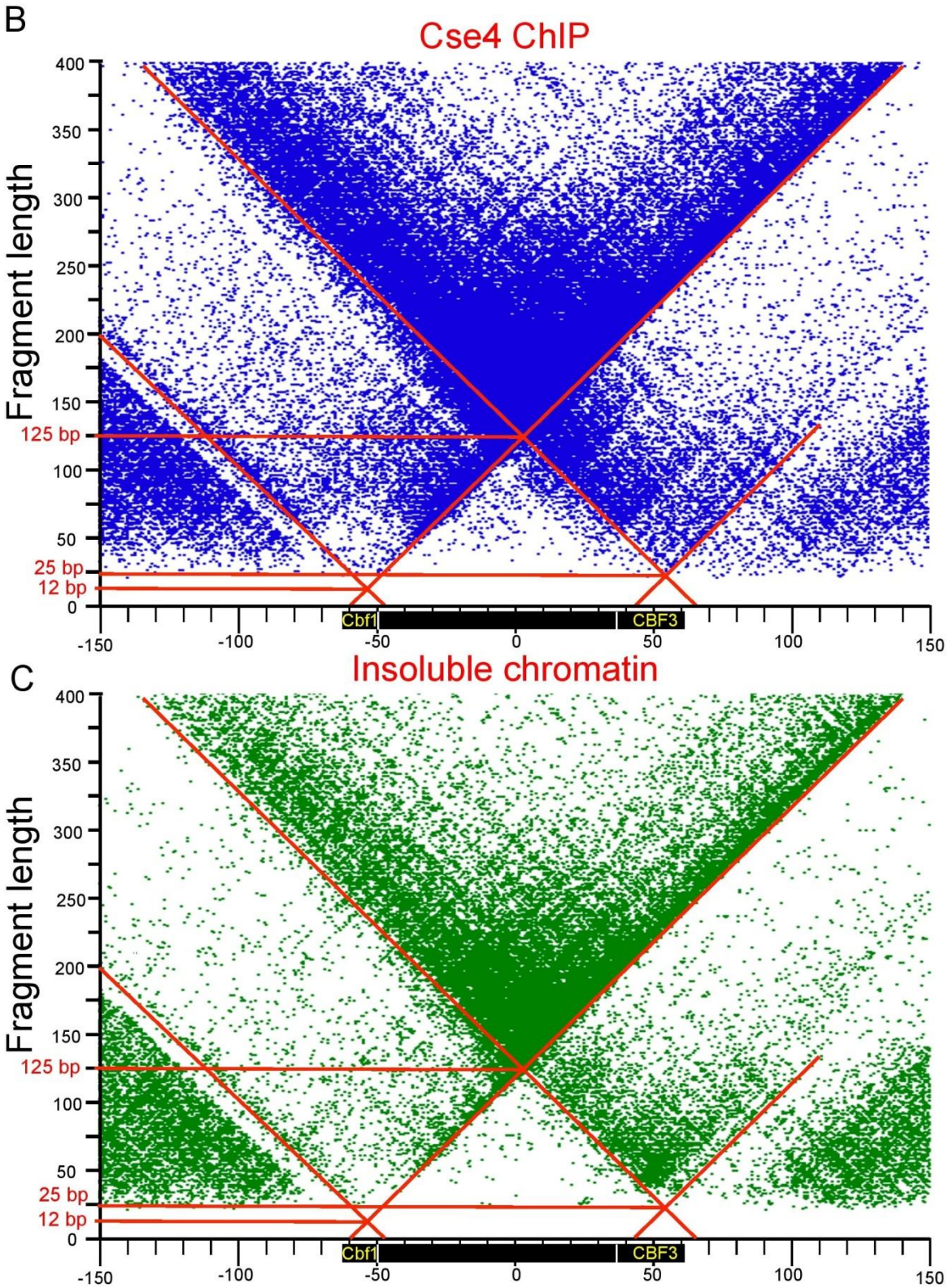

B
Cse4 ChIP
Fragment length
400
350
300
250
200
150
125 bp
100
50
25 bp
12 bp
0
Cbf1
CBF3
-150
-100
-50
0
50
100
150
C
Insoluble chromatin
Fragment length
400
350
300
250
200
150
125 bp
100
50
25 bp
12 bp
0
Cbf1
CBF3
-150
-100
-50
0
50
100
150

Figure 2.11. V-plots of DNA from insoluble chromatin for individual centromeres

See the legend to Figure 2.3 for details on interpretation of V-plots. (A) All 16 centromeres are shown either in groups of 4 or individually. (B-C) Direct comparison of Cse4 ChIP-seq (A) and pellet (B) from the same 20’ MNase sample, showing vertexes corresponding to ~10 bp for intersection of diagonals over CDEI, ~20 bp for intersection of the diagonals over CDEIII, and ~125 bp for intersection of the diagonals over the full CDE.

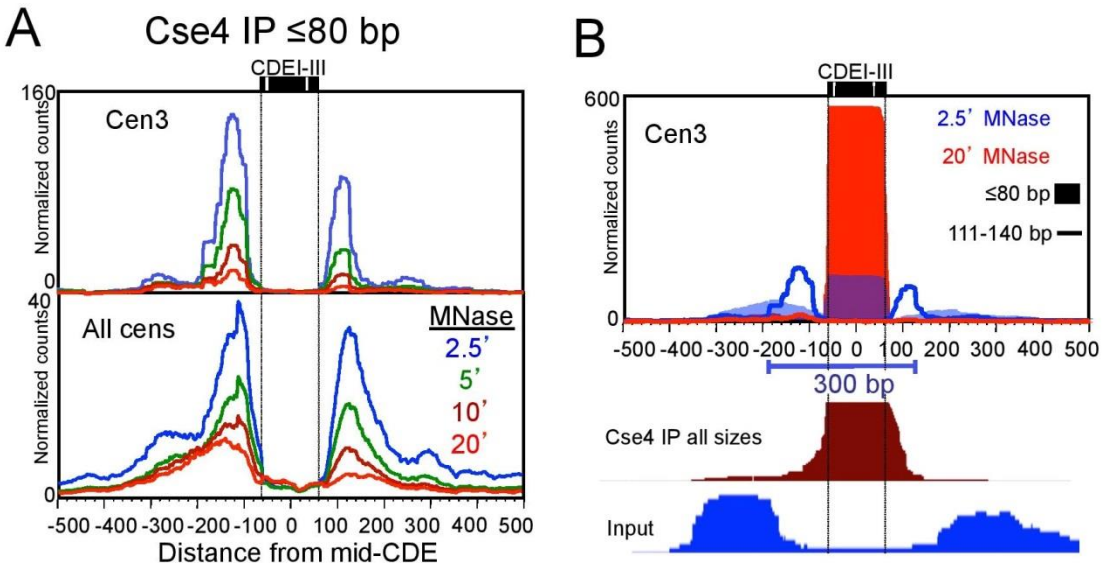


Figure 2.12. The CDE is closely flanked by subnucleosomal particles and phased nucleosomes

(A) Mapped paired-end read counts of ≤80 bp fragments show enrichment after Cse4 ChIP and high sensitivity to MNase digestion on both sides of the centromere after ChIP. (B) Mapped paired-end reads for Cse4 ChIP at two MNase concentrations show subnucleosomal particles flanking the Cen3 CDE, where a square peak indicates nearly perfect protection of the 111-140 bp centromeric DNA segments. The ≤80 bp fragments are lost from the immunoprecipitated material with concomitant enrichment of 110-140 bp fragments during MNase digestion. Below are densities for Cse4 IP (brown) and >140 bp soluble chromatin fragments lined up to illustrate the relative locations of the different MNase-protected particles after 20' digestion.

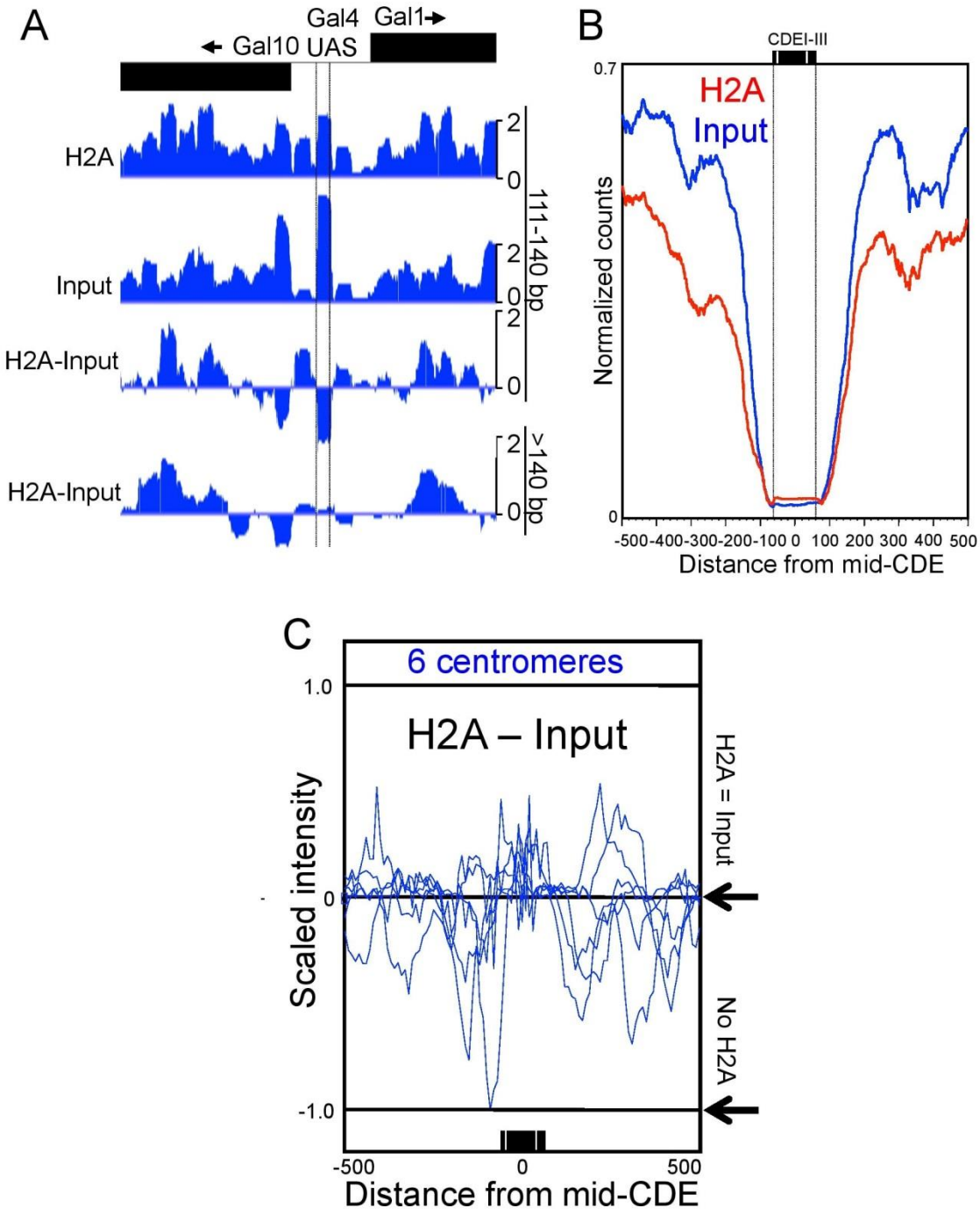


Figure 2.13. Depletion of H2A at the H2A.Z-enriched nucleosome over the Gal4 UAS

(A) Positive control for the experiment shown in Figure 2.5A, showing the 2.5' MNase-digestion profile for total chromatin at the Gal4 region, where the well-positioned nucleosome over the Gal4 UAS was reported to wrap only ~135-bp, and to be depleted

for H2A and enriched for H2A.Z, in contrast to flanking nucleosomes which showed the opposite enrichment [75]. Note that there is high occupancy of input for this nucleosome but lower occupancy of H2A relative to the genome as a whole (H2A-Input < 0) and to flanking nucleosomes. (B) X-ChIP data showing that the Cse4 nucleosome contains H2A. Data are converted from the sum of scaled intensities from ‘AA’ and ‘AZ’ (total H2A-containing) nucleosomes [80]. See the legend to Figure 2.6 for details.

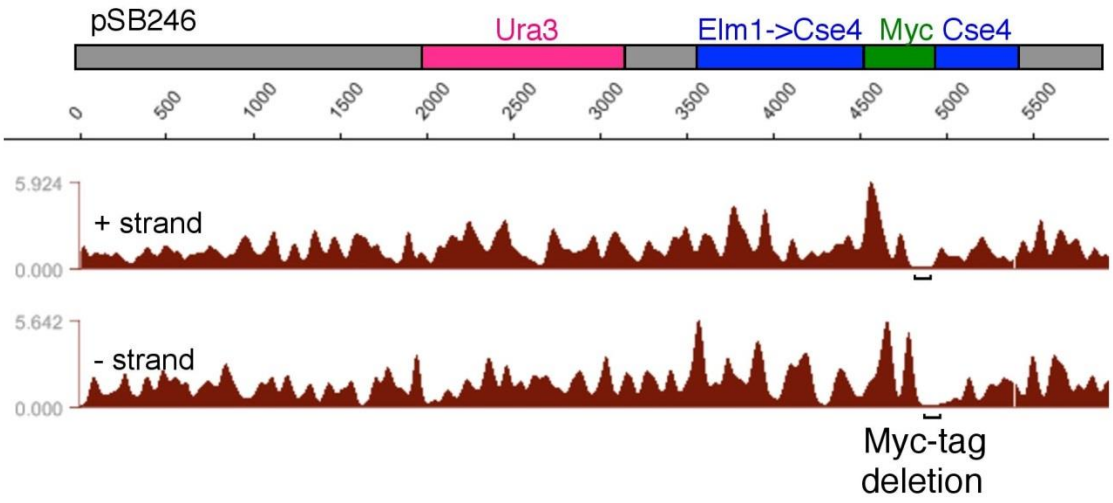


Figure 2.14. Multiple copies of CSE4-Myc in strain SBY8796

The plasmid pSB246 was searched using the first one million single-end 36-bp reads either not mapped to the yeast genome (364160) or mapped uniquely (635840). The 1736 hits with an exact match for 36 bases (834+ and 902-) are displayed, using the left end of the mapped segment in all cases. Copy number estimate based on read density: (1,736 plasmid reads/6 kb per plasmid)/(635,840 genomic reads/12,000 kb per genome) = 5.5 copies per plasmid. This value was confirmed by average URA3 and CSE4 normalized peak heights for the whole dataset (~120/20 = 6). The map of pSB246 shows Escherichia coli sequence in grey, Ura3 sequence in red, Cse4 and upstream Elm1 sequence in blue, and 11 tandem Myc sites in green. Sequence assembly revealed that three of the 11 Myc sites have been deleted from all copies (bracketed region).

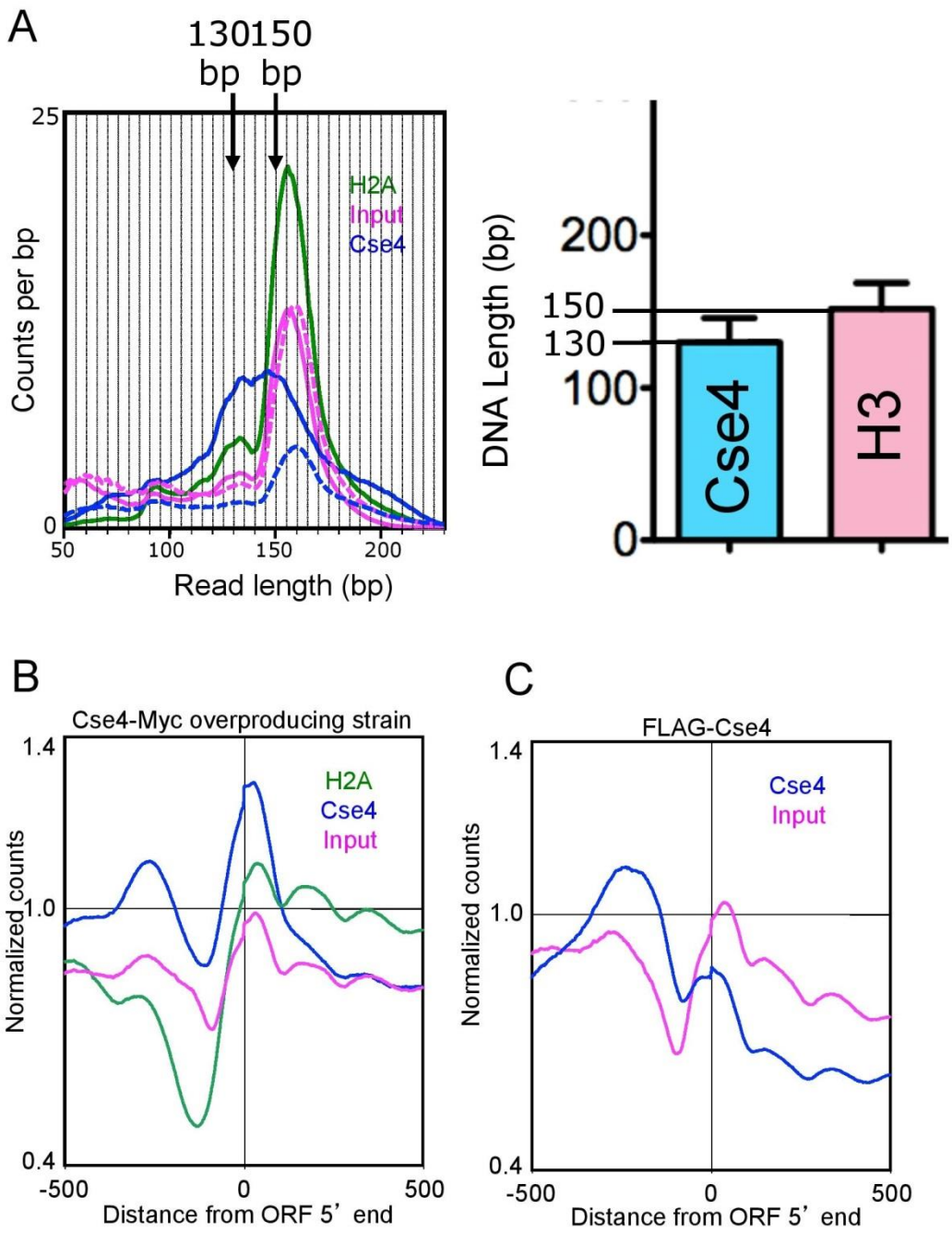


Figure 2.15. Overproduced Cse4 nucleosomes deposit at 'hot' nucleosome positions

(A) Length distributions of mapped fragments genome-wide, showing a comparison to lengths of MNase-protected fragments extracted from purified nucleosomes assembled in vitro and measured on agarose gels [62]. The results from two separate experiments are shown, where little difference is seen in the degree of MNase digestion based on the slight offset in input peak position. (B) To confirm that 'hot' nucleosomes are generally enriched for Cse4 nucleosomes relative to H2A nucleosomes around

promoters (Figure 2.7B), we aligned all yeast genes at their 5’ ORF ends and plotted the total number of normalized counts at each base pair. Strong enrichment is seen for overproduced Cse4-Myc and depletion of FLAG-H2A relative to Input from the -1 to +1 nucleosome positions. (C) In a single-copy control strain, Cse4 enrichment is seen over highly active promoter regions, as previously reported [23, 66].

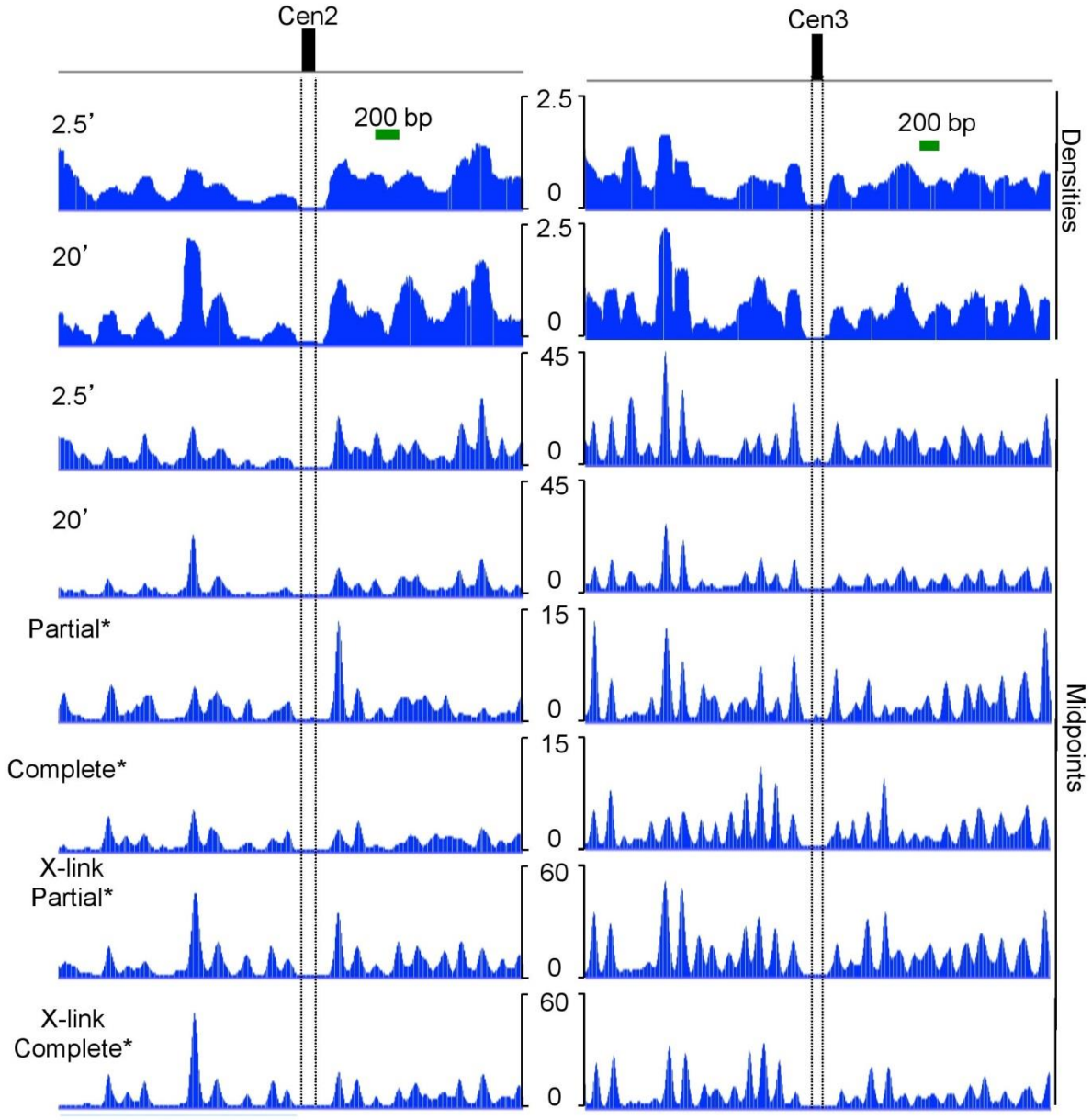


Figure 2.16. Centromeric DNA is depleted from both native and cross-linked chromatin

Regions around Cen2 and Cen3 are shown. For midpoint landscapes from single-end reads, an offset of 75 bp was added to each plus end and subtracted from each minus end. A Gaussian kernel density function with bandwidth = 10 and window = 5 was applied to both single-end and paired-end reads.

Table 2-1 Yeast strains used in this study

Strain*	**Genotype**
SBY8796	MATa leu2-3,112 his3-11,15 trp1-1 lys- RAD5 hta2-htb2::NAT HTA1-L3FLAG-kanMX cse4ΔKAN ura3-1:CSE4-myc12:URA
SBY5146	MATa leu2-3,112 his3-11:pCUP1-GFP12-LacI12:HIS3 trp1-1:256lacO:TRP1 can1-100 ade2-1 Δlys2 Δbar1 cse4ΔKan ura3-1:pCse4-3xFLAG-CSE4 + 500bp
SBY2688	MATa leu2-3,112 his3-11,15 trp1-1 LYS2 RAD5 hta2-htb2::NAT HTA1-L3FLAG-kanMX ura3-1:CSE4-myc12:URA

*All strains were derived from W303.

3 Chapter 3 Distinct chromatin features characterize different classes of repeat sequences in Drosophila melanogaster[2]

3.1 Background

A large fraction of almost all eukaryotic genomes consists of tandemly repeated sequences, often called satellite DNA [86]. Because satellite DNA repeat units are short and vary little if at all in sequence, they are mostly excluded from genome assemblies [87]. This is unfortunate, because some satellite sequences are known to have important functions. For example centromeres – chromosome loci that form microtubule attachment sites during mitosis - are known to be positioned on repeat sequences in many organisms [86]. Another example is telomeres – sequences that cap chromosome ends. Also, changes in satellite sequences can play roles in evolution and disease [88]. Satellite sequences might have other functions: For example, they have been shown to be important for meiotic recombination [89].

Whole genome sequencing has become a widely used tool for the discovery of genetic variation, nucleosome position, chromatin modifications and DNA binding proteins. Analysis of such experiments relies on the alignment of individual sequence segments to the reference genome. Because it is impossible to uniquely align satellite sequences they are usually excluded from the genome assembly. Thus alternative methods for analysis of repeats in sequencing data are required.

[2] This chapter appeared previously as Krassovsky, K. and S. Henikoff (2014). "Distinct chromatin features characterize different classes of repeat sequences in Drosophila melanogaster." BMC Genomics 15(1): 105.

Several groups have used methods independent of the alignment to the reference genome to analyze repeat content. Parker et al. [88] used direct counting of telomere repeat sequences to estimate changes of telomere repeats in tumor cells. Hayden and Willard [90] used k-mer analysis and repeat library alignment to describe canine centromere sequences. In this study we adapt these approaches with some modifications to study the repeat content of *Drosophila melanogaster*. Drosophila is a particularly attractive model because of previous extensive characterization of its satellite repeat content by methods other than sequencing. This provides a unique opportunity to verify recovery of satellites in sequencing data.

Drosophila repeat families were initially discovered by detection of satellite bands that form during CsCl equilibrium gradient centrifugation [91, 92]. When centrifuged at high force CsCl creates a gradient of Cs^+ ions. While moving through this gradient long DNA fragments separate into distinct bands based on the buoyant density, which depends primarily on GC content. Tandem arrays of short repeat units typically have biased base composition [*e.g.* $(AAGAG)_n$ is 60% A+T and 40% G+C)], so that they effectively separate from long DNA fragments of average base composition comprising single-copy DNA. The bands can then be extracted, cloned and sequenced. Three of four of such bands were shown to consist of short (5 to 10 bp) repeats, while the fourth one consisted of longer (359 bp) repeat sequences [47]. Both classes of these tandem repeats are highly abundant in the genome and map primarily to centromeric and pericentric regions of chromosomes. Another class of repeats is derived from transposable elements, found in all eukaryotic genomes. These are DNA sequences that have inserted copies of themselves into new positions in the genome, and are

interspersed with single-copy or satellite sequences. Transposons have been shown to comprise ~15% of the *Drosophila melanogaster* genome [93].

Most of the repeated sequences are packaged into heterochromatin – condensed and mostly transcriptionally silent chromatin identified cytologically as being more refractile and more densely staining [35]. Heterochromatin can be divided into constitutive, chromatin that is permanently condensed and is found in pericentric and telomeric regions, and facultative, gene-containing chromatin where condensation is associated with repression of gene expression [94]. It is thought that this condensation and gene repression is achieved partly by posttranslational histone modifications, which are known to be enriched at different functional elements. For example, H3K4me3 is found at promoters of active genes [95] in a variety of organisms. In flies it has been shown that constitutive heterochromatin is associated with H3K9me2 while repressed genes in facultative heterochromatin are enriched in H3K27me3 [96].

Associations of specific DNA binding proteins with histone modifications are currently studied by chromatin immunoprecipitation followed by sequencing (Chip-Seq). Analysis of such experiments has thus far been limited to single-copy sequences and interspersed repeats. Studies of tandemly repeated sequences in heterochromatin by Chip-Seq are impeded by the inability to uniquely align repeat-containing reads to the reference genome.

Recently two large-scale initiatives generated comprehensive *D. melanogaster* sequencing datasets. One is the Drosophila Genetic Reference Panel (DGRP) which included sequencing of 200 inbred fly lines generated from wild caught flies [48]. Data

generated by DGRP were used to study phenotype-genotype associations and evolution of the subset of repeat sequences that could be mapped uniquely. The other large-scale initiative is modENCODE, which included Chip-Seq experiments for a number of DNA binding proteins and histone tail modifications from different developmental stages of Drosophila. In this study we used these publicly available resources to analyze the repeat content of the *D. melanogaster* genome and to identify histone tail modifications and DNA binding proteins associated with satellites.

3.2 Results and Discussion

3.2.1 Strategy for quantifying repeats

We used three independent metrics to describe repeat content: (1) alignment to the libraries of known repeats; (2) estimation of the proportion of low complexity sequences; (3) classification of the most frequent k-mers (Figure 3.1).

Repeat libraries were constructed for short repeats (FlyBase), 359 bp repeats [97] and transposons (FlyBase) by extraction from existing genome assemblies including unassembled contigs. A complexity score similar to the DUST score used by the BLAST program to exclude low-complexity sequences was calculated for each sequenced fragment. Short repeat units have low complexity scores. This means that finding the number of sequences with a low complexity score allows us to estimate the percentage of short repeats independent of alignment programs.

Another alternative to alignment to reference libraries is k-mer analysis. A k-mer is a sequence of length k found in the sequencing dataset. For example, the 5-mer AAGAG is one of the 5-mers found in the sequence AAGAGAAGAG. By counting all k-mers we can find sequences that occur very frequently in the genome. Satellites result in k-mers

that have a much higher count than the rest of the genome. K-mer and low complexity analyses provide an estimation of the completeness of repeat libraries and find abundant sequences not included in the libraries.

To find the overall fold enrichment of satellites in the ChIP-seq experiments we calculated fold enrichment of each k-mer present at least twice in both ChIP and input samples. We then grouped k-mers by their enrichment value and classified each k-mer as being in one of the repeat families, in the euchromatin, or not previously mapped, by aligning to each of the repeat libraries and the whole annotated genome. Classification and grouping of k-mers in this manner allows visualization of enrichment or depletion of a particular repeat family.

DGRP datasets include multiple sequencing runs for the same fly line, which allowed us to distinguish variability introduced by experimental variation from biologically relevant variation that occurred due to differences between individual flies in the wild population. We calculated the percentage of reads that mapped to each of the repeat families as well as the percentage of low complexity sequences (Figure 3.2) and averaged the percentages between experiments of the same fly line. To determine whether variation is statistically significant we performed ANOVA tests (Table 3-1) and found that short repeats and 359 bp repeats did not differ significantly between DGRP fly lines. Low complexity sequences and transposons changed slightly but significantly overall. These findings imply that sequence variation is confined to interspersed, but not tandem repeat sequence families.

3.2.2 All repeat classes are depleted in larval relative to adult and embryonic developmental stages

An unusual feature of the Drosophila genome is changes in the repeat content for some cells of the organism. The best-known example is the polytene chromosomes of larval salivary glands [45, 98]. During larval stages of development rapid growth of the organism requires high levels of gene expression. To efficiently accommodate this need cells undergo multiple rounds of replication without mitoses or cell divisions. This process results in banded polytene chromosomes, which are composed of multiple precisely aligned copies of sister chromatids and homologs. Chromosomes in most larval tissues are polytene, with salivary gland chromosomes being the most extremely polytenized. Polytene chromosomes are depleted of heterochromatin, especially satellite sequences. Another cell type that is depleted of satellite DNA is nurse cells, which produce yolk that is stored in the egg and consumed during embryonic development [99]. Unlike salivary glands and other larval tissues, nurse cell nuclei lack polytene structure.

We took advantage of the sequencing datasets from different development stages available from modENCODE. In order to determine the abundance of different repeat families in the genomes of embryos, larvae and adults, we mapped sequences from input datasets to the repeat libraries and calculated the percentage of mapped reads. We also calculated the percentages of low complexity sequences in each dataset (Figure 3.3). To determine whether repeat sequence abundance changes between developmental stages, we compared the median percentage abundance between sequences from the same developmental stage to that across all sequences (Table 3-2). As expected, we found that all repeat families are significantly depleted in larvae relative to embryos and adults (Table 3-3).

The percentage of short repeats has been previously estimated at 5-10% [44] using cot curves and at 18-22% [47, 92] using CsCl gradients for embryos of the Oregon R wild-type lab strain. In the datasets examined the short repeat content is 3% on average for DGRP flies, 12% for embryos and adult heads and 3% for larvae of modENCODE flies. The lower repeat content in DGRP samples might be explained by the use of whole flies in these experiments, where nurse and follicle cells that make up most of the mass of healthy adult female flies in uncrowded cultures [100] will lower the satellite repeat content.

Multiple experimental replicates available in both modENCODE and DGRP datasets present an opportunity to examine the reliability of modern sequencing methods for recovery of repeated sequences. On the one hand, the abundance of short repeats varies only slightly between distinct DGRP fly lines and replicate datasets derived from the same fly line. On the other hand, there is considerable variation between modENCODE replicate datasets. This variability is unlikely to be due to alignment bias because we obtained similar estimates using an alternative method of finding short repeat sequences based on sequence complexity. High variation might be due to random loss or amplification of repeats by PCR during Illumina library preparation and flow cell cluster generation. PCR is known to have biases in amplification due to composition [101]. Alternatively, variation might be due to real sequence heterogeneity among individuals of the same laboratory strain, as has been previously suggested for satellite sequences [89]. This possibility is consistent with our observation that short repeat recovery is less variable for embryos than adult flies. It is possible that fewer adult flies are needed for the recovery of material necessary for constructing Illumina

sequencing libraries compared to embryos, where there are fewer cells per individual, making inter-individual heterogeneity more evident in adult than embryo samples. Another possibility is that the differences arose from the DNA preparation method used. Unlike DGRP samples where DNA was extracted from flies directly, modENCODE samples were prepared for ChIP by extraction of cross-linked chromatin. Sonication of chromatin as opposed to sonication of naked DNA might produce additional variability between the experiments.

3.2.3 The most frequent k-mers in the fly genome are known short repeats and transposons

We wanted to find the most abundant repeat sequences of the fly genome independent of the mapping to the known repeat libraries. Such an estimation is important to ascertain the completeness of existing annotations and libraries. We found the occurrence of all k-mers of length 31 in all samples of DGRP. This k-mer length was chosen to be large enough to allow distinguishing unique from repeated sequences by BLAST searching (22 bp, Ref. [102]) but smaller than the minimum read length (45 bp). We then divided all k-mers of length 31 into equal quintiles based on their count and classified them by repeat family, if known. The classification of repeats averaged across all datasets of 10 DGRP fly lines is shown in Figure 3.4. The most frequent k-mers belong to the short repeat class. Almost all of the frequent k-mers are classified as belonging to one of the known repeat families.

We noticed that ~2% of the fly genome in DGRP datasets can be classified as low-complexity sequences that are not present in the short repeat library. Two-thirds of the low-complexity sequences represent imperfect short repeats, i.e. runs of short repeats interspersed with a small number of changes. By manual scrutiny we observed that

many of the remaining sequences contained long runs of single nucleotides of which the large majority consisted of stretches of As or Ts, with a small percentage with stretches of Gs or Cs. Examples of such sequences together with their proportion of the total genome are presented in Figure 3.5. Occurrences of tracks of 'T' and 'A' were previously examined in several genomes [103] and found to be more abundant than expected by chance. Such sequences have been previously described near promoters of some genes and have been shown to effectively exclude nucleosomes [104].

Previous studies of fly repeats using cloning and sequencing of satellite bands isolated from CsCl gradients showed that satellite DNA in Drosophila includes 11 individual short sequences [47]. A limitation of this approach recognized by the authors is the instability of repeats when cloned into *Escherichia coli* and hence possible biases against them, as well as co-sedimentation of repeat-containing fragments with single-copy DNA. We wanted to compare this previous result with sequences obtained by modern high thoughput sequencing. To do so we counted the number of reads mapped to each sequence entry in the short repeat library in DGRP and modENCODE input datasets. We then grouped individual repeats by their frequency (Figure 3.6a and Figure 3.6b). We found that although our short repeat library contains more than 200 individual repeats, only 13-14 of them make up ~90% of the short repeat satellite reads. We also identified the most abundant short repeats and compared them to the sequences identified previously by cloning and sequencing CsCl gradient bands (Figure 3.6c and Figure 3.6d) [47]. Out of 11 sequences identified in previous work we found that 4 ("AATAACATAG", "AAGAGAG", "AAGAG", "AAGAC") are among the most abundant repeats in DGRP flies and 6 ("AATAACATAG", "AAGAGAG", "AAGAG", "AATAT",

“AAGAC”, “AATAGAC”) in modENCODE flies. Interestingly, some of the short repeats differ only in the interchange of two nucleotides, such as “AATAACATAG” and “AATAAGATAC”. Such sequences would have the nearly same buoyant density and will band together in a CsCl gradient.

In both DGRP and modENCODE samples we found repeat sequences that were not previously identified as abundant repeats. We might attribute this discrepancy to differences in the method of DNA extraction or to evolutionary changes that occurred in the Oregon R strain, which has been maintained in various laboratories for several decades. Abundances of specific repeat sequences among the DGRP fly lines are very similar, indicating strong homogeneity of the satellites among individual flies in the wild outbred population. We did not find four of the previously reported repeats (“AACAA”, “AATAAC”, “AATAC” and “AATAG”) among the top repeats in modENCODE samples, although they are present at very low abundance in both modENCODE and DGRP samples.

3.2.4 Histone H3 modifications are differentially associated with short repeat sequences

Histone tails can have various post-translational modifications that are associated with different states of the chromatin. For example, di- and tri-methylated H3K9 are known markers of heterochromatin, trimethylated H3K27 is found at facultatively silenced genes, and trimethylated H3K4me3 is associated with gene activity. All of these marks have been studied in the mappable single-copy segments of the genome. We wanted to investigate associations of these marks with different classes of repeated sequences. To do so we identified k-mers that are present in both input and IP samples obtained from embryos in the modENCODE datasets and for each k-mer we calculated its

enrichment relative to the input. We then separated these k-mers into groups based on their fold enrichment (Figure 3.7 left). As expected, the distribution of k-mer enrichment resembles a normal distribution, with a majority of the sequences neither enriched nor depleted. We then mapped k-mer groups to each of our repeat libraries or to the genome assembly. In this way each k-mer was classified as either one of the repeat types, part of the genome assembly, or unmapped. We then plotted the percentage of each repeat type in each group as well as the percentage of unmapped k-mers in each group (Figure 3.7 middle). Such k-mer classification allows a visualization of enrichment of particular histone modifications in each repeat class. As expected, H3K4me3 is virtually absent from all the repeat types and H3K9me3 and H3K9me2 are enriched for some short repeats and transposons. Surprisingly, we found H3K9me1 to be depleted from short repeats but enriched in the 359 bp repeats. H3K9me1 has been shown to be a substrate for a histone methyltransferase that catalyzes di- and tri-methylation in mouse and Arabidopsis [105], [106], but the specificity of chromatin association of this modification in Drosophila has not been reported previously. H3K27me3 is depleted from short and 359 bp repeats but enriched in transposons, which is consistent with it being a mark of facultative heterochromatin. As described below, some short repeats are also depleted for all histone modifications examined.

Posttranslational histone tail modifications are known to be involved in transposon silencing. Transposons are classified into groups based on their structure and mechanism of transposition. Retrotransposons, which mobilize via an RNA intermediate, are further divided into LTR (long terminal repeats) and non-LTR classes. Previous studies investigated whether some transposon families are preferentially

associated with specific histone modifications. For example, a screen of 100 transposon sequences by microarray analysis found that retrotransposons have higher enrichment in H3K9me2 than other elements [94]. In contrast, *roo* retrotransposons, which are abundant in euchromatin, were found to have lower H3K9me2 association. Four families of LTR retrotransposons (*roo*, *tirant*, *412* and *F*) were also screened for preferential association with H3K9me2 and H3K27me3 in different strains of *D. melanogaster* and were found to have large variations in enrichment between the strains. However, our systematic investigation based on classification of Illumina sequencing reads both by k-mer analysis and direct counting of reads mapped to different transposon groups detected no preferential association of LTR, non-LTR or IR transposon classes with histone modifications (Table 3-6).

3.2.5 All three HP1 proteins localize to transposons

We also examined ChIP datasets of Heterochromatin-associated Protein 1 (HP1) for association with different repeat classes. HP1 has been implicated in the formation of heterochromatin by the binding of its “chromodomain” to di- and tri-methylated H3K9 [37, 107] and by dimerization of its “chromo-shadow” domain, bringing neighboring nucleosomes together to condense chromatin [62]. Drosophila has three closely related HP1 proteins, HP1a, HP1b and HP1c, each of which has been shown to have a different localization pattern by cytology [108]. HP1a has been shown to be required for silencing of transposons and is exclusively localized to heterochromatin [109, 110]. HP1c localizes to euchromatin and HP1b localizes to both euchromatin and heterochromatin. However, k-mer analysis shows that all three of the HP1 proteins are enriched in transposons and depleted in other types of repeats (Figure 3.8 middle). This

is unexpected because HP1a has not been shown to have preferential localization with different classes of heterochromatin, such as transposons versus satellites.

3.2.6 AT-rich repeats are depleted of nucleosomes

We noticed that even for the H3K9me3 and H3K9me2 heterochromatic marks that are enriched in short repeats a few specific repeat sequences are depleted of these marks. This prompted us to look for a common property of short repeats that are depleted of heterochromatic marks and HP1 proteins. We first classified each repeat by the length of the repeat unit but detected no consistent trends. However, when we classified repeats by AT content, we observed that the short repeats that are depleted of HP1 family proteins and histone modifications are also very AT rich (Figure 3.7 and Figure 3.8 right panels; Table 3-7).

We hypothesized that the consistent depletion of short AT-rich repeats from ChIP datasets of histone modifications and chromatin proteins that bind them is due to the depletion of nucleosomes themselves. To test this possibility we performed k-mer analysis on sequences enriched by ChIP-seq of H3 and H4 histones. We found that these histones are also depleted of short AT-rich repeats (Figure 3.9; Table 3-7). Hence depletion of histone modifications and HP1 proteins from AT-rich short repeat sequences is not due to selectivity against these chromatin features but rather is explained by the overall depletion of nucleosomes from AT-rich short repeats.

Highly AT-rich DNA has a narrow minor groove and reduced flexibility, which disfavors the tight wrapping of the double helix around the nucleosome core and results in preferential exclusion of nucleosomes [104, 111]. As the (AATAT)n, (AATATAT)n and other long arrays of pure AT sequences are predicted to be especially stiff [112], they

would be expected to prevent nucleosome formation. Alternatively, nucleosomes might be actively excluded by competing DNA-binding proteins. For example, D1 protein is a highly abundant nuclear protein that is preferentially bound to the narrow minor groove of AATAT *Drosophila* satellite arrays [113, 114]. With ~1 D1 protein per 10 nucleosomes, and ~0.7% of the genome consisting of AATAT-containing satellites, there is enough chromatin-bound D1 to occupy ~1/2 of all the AATAT sites [(1 D1/10 nucleosomes)/(30 AATAT sites in a 150 bp span) = 0.0033% of the genome]. These alternative possibilities are not mutually exclusive, as expansion of an AATAT array would both exclude nucleosomes and promote D1 binding, consistent with the possibility that D1 protein has evolved to package stiff AT-rich satellites.

3.3 Conclusions

We have shown that enrichment of repeated sequences can be quantified in Chip-Seq experiments despite being largely excluded from genome assemblies. The strategy of calculating k-mer enrichment relative to the input allows direct comparison of repeat sequences to single-copy regions of the genome. The strategy presented here can be applied to study other chromatin features known to be located in heterochromatin, for example centromeres.

We also have presented the first analysis of the chromatin landscape of repeat sequences in a genome-wide context. Different heterochromatic regions of *D. melanogaster* have distinct chromatin features. Satellite sequences associate with specific histone modifications such as H3K9me2 and H3K9me3. All three HP1 homologues are enriched at transposons and do not show preferences for particular types of transposons. AT-rich short repeats are depleted of nucleosomes and hence all

histone modifications. We conclude that ChIP-seq datasets can be mined to provide unexpected insights into chromatin landscapes of repetitive sequences.

3.4 Methods

3.4.1 Datasets

modEncode datasets listed in Table 3-4 were downloaded from http://data.modencode.org. DGRP datasets listed in Table 3-5 were downloaded from http://www.ncbi.nlm.nih.gov/sra?term=DGRP. For each fly line only sequences generated by the Illumina platform were used.

3.4.2 Repeat Libraries

The short repeats library was downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/dm3/bigZips/chromTrf.tar.gz. It was converted to a fasta file format and purged of duplicate entries. The 359 bp library was the one produced in [97] and obtained directly from Dr. Gustavo Kuhn.

The transposon library was downloaded from FlyBase r5.48 ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_5.48_FB2012_06/fasta/dmel-all-transposon-r5.48.fasta**.**

3.4.3 Determining repeat abundances

Sequences were mapped to a short repeat and 359 bp repeat library using BWA [101] and to transposons using Novoalign (www.novocraft.com). The number of sequences mapped to the library was divided by the total number of sequences to find the percentage abundance.

3.4.4 K-mer analysis

K-mers were obtained using Jellyfish [115] with the command “jellyfish count -m 31 -o output -c 3 -s 10000000 -t 12 -L 2”. K-mers were split into quintiles using a custom script and aligned to repeat libraries using BWA (short repeats and 359 bp repeats) and Novoalign (transposons).

3.4.5 Finding low complexity sequences

The percentage of low complexity sequences was found by running Prinseq [116] with the command “perl prinseq-lite.pl -fastq FileName.fastq -verbose -graph_data -out_good null -lc_method dust -lc_threshold 7”. This command separates sequences with a complexity score above 7 and records that number in the log file.

3.4.6 K-mer analysis of the ChIP-seq datasets

A k-mer count table was constructed for both Input and ChIP samples using Jellyfish and then merged using a custom R script. For each k-mer, enrichment was calculated by dividing the number of counts in the ChIP dataset by the number of counts in the corresponding Input dataset and normalized by multiplying by the ratio of the total number of sequences in input and ChIP samples. K-mers then were split into 16 groups based on enrichment. K-mer sequences from each group were aligned to repeat libraries and the genome assembly using BWA and Novoalign. The number of k-mers in each group mapped to a particular library was noted and then plotted using an R script. For experiments with two replicates the median number of k-mers in each bin is shown.

Table 3-1. Abundance of different repeat families in DGRP flies

Repeat Family	% of total genome (mean between lines)	Standard deviation between lines	p-value (significance of the difference between lines)
Total low Complexity	5.2	0.44	6.968e-06***
Low complexity, not short repeats	2.21	0.32	0.19
Short Repeats	2.98	0.33	0.03128*
359 bp Repeats	1.31	0.24	0.01315*
Transposons	11.3	0.88	1.613e-05***

Table 3-2 Abundance of different repeat families in fly genome by developmental stage (modENCODE dataset)

Developmental stage	Repeat Family	% of total genome (median)
Embryo (12-14 hr)	Low Complexity	10.22
	Short Repeats	7.75
	359 bp Repeats	3.00
	Transposons	16.0
Larvae	Low Complexity	6.13
	Short Repeats	3.90
	359 bp Repeats	1.43
	Transposons	11.6
Adult Head	Low Complexity	12.45
	Short Repeats	9.8
	359 bp Repeats	2.20
	Transposons	15.8

Table 3-3. p-values that indicate significance of the difference between the developmental stages

Developmental stage	Repeat Family	p-value	change
Embryo/Adult	Low Complexity	0.07273	2.23
	Short Repeats	0.2209	2.05
	359 bp Repeats	6.087e-06***	0.8
	Transposons	0.627	0.2
Adult/Larvae	Low Complexity	0.0003176***	6.32
	Short Repeats	0.001056**	5.9
	359 bp Repeats	0.002137**	0.77
	Transposons	0.0002104***	4.2
Embryo/Larvae	Low Complexity	1.281e-05***	4.09
	Short Repeats	1.922e-06***	3.85
	359 bp Repeats	2.677e-09***	1.57
	Transposons	7.374e-07***	4.4

Table 3-4. DGRP datasets used in the study

Fly Line	Experiment ID
DGRP-313	SRR018517, SRR018518, SRR018519
DGRP-357	SRR018285, SRR018286
DGRP-358	SRR018574, SRR018575, SRR029943, SRR034277, SRR034278
DGRP-362	SRR029164, SRR029166
DGRP-365	SRR018579, SRR034281, SRR034282, SRR034283
DGRP-375	SRR018287, SRR018288, SRR018289, SRR018290, SRR018291
DGRP-379	SRR018582, SRR018583, SRR018584
DGRP-380	SRR018591, SRR018592, SRR018593
DGRP-391	SRR018292, SRR018293, SRR018294, SRR060098
DGRP-399	SRR018295, SRR018296, SRR018297

Table 3-5. modENCODE datasets (http://data.modencode.org) used in this study.

Assay Factor	Development Stage	Experiment ID*
H3K9me1	Embryo	4123
H3K9me2	Embryo	4940
H3K9me3	Embryo	4939
H3K4me3	Embryo	5096
H3K27me3	Embryo	3955
H3	Embryo	5079
H4	Embryo	5107
HP1a	Embryo	3956
HP1b	Embryo	5111
HP1c	Embryo	5587
H3K9me2	Larvae	4958
H3K9me3	Larvae	4952
H3K4me3	Larvae	5097
H3K27me3	Larvae	5089
HP1a	Larvae	4936
HP1b	Larvae	5110
HP1c	Larvae	5112
H3K9me2	Adult	5259
H3K9me3	Adult	4933
H3K4me3	Adult	5098
H3K27me3	Adult	5583
HP1a	Adult	5592
HP1b	Adult	5590
HP1c	Adult	5591

*Each experiment contains both input and IP sequences and one or more replicates.

Table 3-6. Total enrichment of histone tail modifications by transposon family group

Percentage of reads mapped to each of the groups was calculated in both input and IP datasets. Enrichment was calculated as the ratio of percentage of reads in the Input dataset to IP dataset. Values were averaged between experiment replicates and mean values together with standard deviations are presented in the table.

Modification	LTR	Non LTR	SINE	IR
H3K9me1	0.91 ± 0.02	1.04 ± 0.01	0.77 ± 0.08	1.06 ± 0.03
H3K9me2	2 ± 0.07	2.23 ± 0.1	2.5 ± 0.13	1.79 ± 0.06
H3K9me3	1.82 ± 0.01	2.18 ± 0.01	2.79 ± 0.1	1.76 ± 0.01
H3K4me3	0.31 ± 0.19	0.41 ± 0.23	0.36 ± 0.19	0.35 ± 0.24
H3K27me3	0.72 ± 0.04	0.79 ± 0.02	0.72 ± 0.04	0.75 ± 0.02

Table 3-7. Sequences of 5-12 mer repeats identified as enriched (top 4 bins) or depleted (bottom 4 bins) in histone modifications IP samples

Only the most abundant sequences (top 90% of k-mer from each group) are shown for brevity. Percentage indicates portion of the k-mers from selected bins that map to particular sequence.

Chromatin feature	Enriched sequences	%	Depleted sequences	%
H3K9me1	GGTCCCGTACTC	26.96	AATAAGATAC	30.34
	CAGTACGGGAC	16.94	AATAT *	19.45
	CCGTACTGGTC	16.28	AATAACATAG *	11.63
	AAGAG *	5.95	AATAC *	5.79
	AGTACGGAACCG	4.78	AAGAT	4.08
	ACAAC	4.54	TGTAT	3.75
	CCTCT	4.28	ATATAAT	2.96
	AAGAGAG *	1.51	AATAG *	2.69
	AAGAC *	1.21	ATATAATA	2.22
	CAAACACAAACA	1.17	AATAGAC *	1.62
	GTACGGGACCGA	1.03	TAATAAA	1.48
	CGTACTCGGTTC	0.87	ATATATAA	1.44
	TGCTGCTGC	0.78	ATATTTT	1.34
	GGAACCAGTAC	0.76		
	AAGAGAGAAGAG	0.65		
	GACAC	0.64		
	AAGACATGAC	0.62		
	CAAGG	0.62		
H3K9me2	AAGAG *	33.59	AATAAGATAC	48.88
	AAGAGAG *	12.14	AATAACATAG *	17.92
	CCTCT	9.32	AATAT *	10.64
	ACAAC	7.35	AATAC *	8.16
	AAGACA	3.85		
	AAGAA	3.76		
	AAGACATGAC	3.42		
	AAGAC *	3.42		
	CTTCTC	3.08		
	TCTTCTCTTT	2.39		
	AAGAGAGAAGAG	2.14		
	TCTTTG	1.88		
	CAGTACGGGAC	1.45		
	AGAAAG	1.37		
H3K9me3	AAGAG *	34.53	AATAT *	78.43137

	AAGAGAG *	12.48	TAATAAA	15.68627
	ACAAC	7.56	TTTATTTA	5.882353
	CCTCT	6.85		
	AAGACA	3.95		
	AAGAA	3.87		
	AAGACATGAC	3.51		
	AAGAC *	3.51		
	CTTCTC	3.16		
	TCTTCTCTTT	2.46		
	AAGAGAGAAGAG	2.2		
	TCTTTG	1.93		
	CAGTACGGGAC	1.49		
	AGAAAG	1.41		
H3K27me3	AAGAG *	34.53	AATAT *	34.18
	AAGAGAG *	12.48	AATAAGATAC	29.11
	ACAAC	7.56	AAGAT	13.08
	CCTCT	6.85	AATAACATAG *	8.44
	AAGACA	3.95		
	AAGAA	3.87		
	AAGACATGAC	3.51		
	AAGAC *	3.51		
	CTTCTC	3.16		
	TCTTCTCTTT	2.46		
	AAGAGAGAAGAG	2.2		
	TCTTTG	1.93		
	CAGTACGGGAC	1.49		
	AGAAAG	1.41		
H3K4me3	AAGAG *	34.53	AATAAGATAC	50.37
	AAGAGAG *	12.48	AATAACATAG *	18.47
	ACAAC	7.56	AATAC *	8.41
	CCTCT	6.85	AATAT *	8
	AAGACA	3.95		
	AAGAA	3.87		
	AAGACATGAC	3.51		
	AAGAC *	3.51		
	CTTCTC	3.16		
	AGAAGAGAAA	2.46		
	AAGAGAGAAGAG	2.2		
	TCTTTG	1.93		
	CAGTACGGGAC	1.49		
	AGAAAG	1.41		

HP1a	AAGAG *	32	AATAAGATAC	38.39
	AAGAGAG *	11.56	AATAT *	19.83
	CCTCT	11.4	AATAACATAG *	14.31
	ACAAC	7	AATAC *	6.19
	AAGACATGAC	4.4	TGTAT	4.37
	AAGAA	3.75	ATATAAT	3.03
	AAGACA	3.66	ATATAATA	2.91
	AAGAC *	3.58		
	CTTCTC	2.93		
	AGAAGAGAAA	2.28		
	AAGAGAGAAGAG	2.04		
	TCTTTG	1.79		
	CTTGTCATGT	1.55		
HP1b	AAGAG *	32	AATAAGATAC	38.32
	AAGAGAG *	11.56	AATAT *	19.98
	CCTCT	11.4	AATAACATAG *	14.29
	ACAAC	7	AATAC *	6.17
	AAGACATGAC	4.4	TGTAT	4.36
	AAGAA	3.75	ATATAAT	3.03
	AAGACA	3.66	ATATAATA	2.91
	AAGAC *	3.58		
	CTTCTC	2.93		
	AGAAGAGAAA	2.28		
	AAGAGAGAAGAG	2.04		
	TCTTTG	1.79		
	CTTGTCATGT	1.55		
	CAGTACGGGAC	1.38		
HP1c	AAGAG *	32	AATAAGATAC	38.29
	AAGAGAG *	11.56	AATAT *	20.02
	CCTCT	11.4	AATAACATAG *	14.28
	ACAAC	7	AATAC *	6.17
	AAGACATGAC	4.4	TGTAT	4.36
	AAGAA	3.75	ATATAAT	3.02
	AAGACA	3.66	ATATAATA	2.9
	AAGAC *	3.58		
	CTTCTC	2.93		
	AGAAGAGAAA	2.28		
	AAGAGAGAAGAG	2.04		
	TCTTTG	1.79		
	CTTGTCATGT	1.55		

	CAGTACGGGAC	1.38		
H3	AAGAG *	32	AATAAGATAC	37.66
	AAGAGAG *	11.56	AATAT *	20.35
	CCTCT	11.4	AATAACATAG *	14.04
	ACAAC	7	AATAC *	6.07
	AAGACATGAC	4.4	TGTAT	4.28
	AAGAA	3.75	ATATAAT	2.97
	AAGACA	3.66	ATATAATA	2.86
	AAGAC *	3.58		
	CTTCTC	2.93		
	AGAAGAGAAA	2.28		
	AAGAGAGAAGAG	2.04		
	TCTTTG	1.79		
	CTTGTCATGT	1.55		
	CAGTACGGGAC	1.38		
H4	GGTCCCGTACTC	26.96	AATAAGATAC	30.44
	CAGTACGGGAC	16.94	AATAT *	19.38
	CCGTACTGGTC	16.28	AATAACATAG *	11.66
	AAGAG *	5.95	AATAC *	5.81
	AGTACGGAACCG	4.78	AAGAT	4.09
	ACAAC	4.54	TGTAT	3.76
	CCTCT	4.28	ATATAAT	2.97
	AAGAGAG *	1.51	AATAG *	2.7
	AAGAC *	1.21	ATATAATA	2.23
	CAAACACAAACA	1.17	AATAGAC *	1.63
	GTACGGGACCGA	1.03	TAATAAA	1.49
	CGTACTCGGTTC	0.87	ATATATAA	1.44
	TGCTGCTGC	0.78	ATATTTT	1.35
	GGAACCAGTAC	0.76		
	AAGAGAGAAGAG	0.65		
	GACAC	0.64		
	AAGACATGAC	0.62		
	CAAGG	0.62		

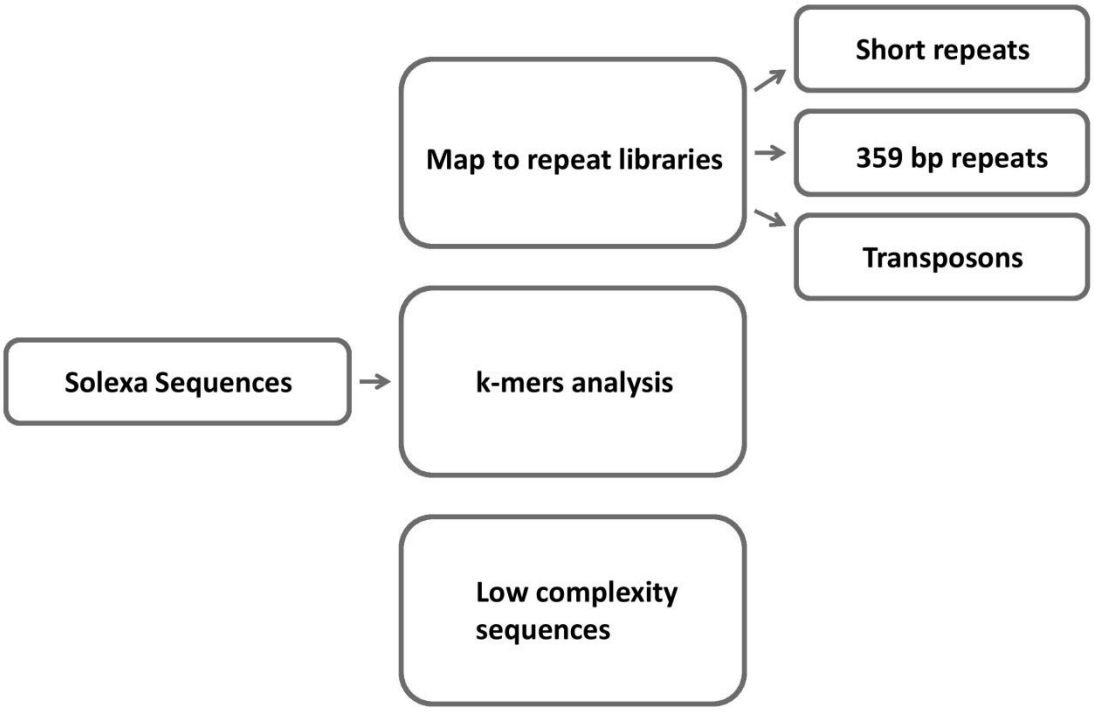


Figure 3.1**.** Strategy for quantifying repeats in sequencing datasets

Three independent approaches were used to quantify repeats: 1) map to repeat libraries; 2) count frequent k-mer; 3) extract and analyze low complexity sequences.

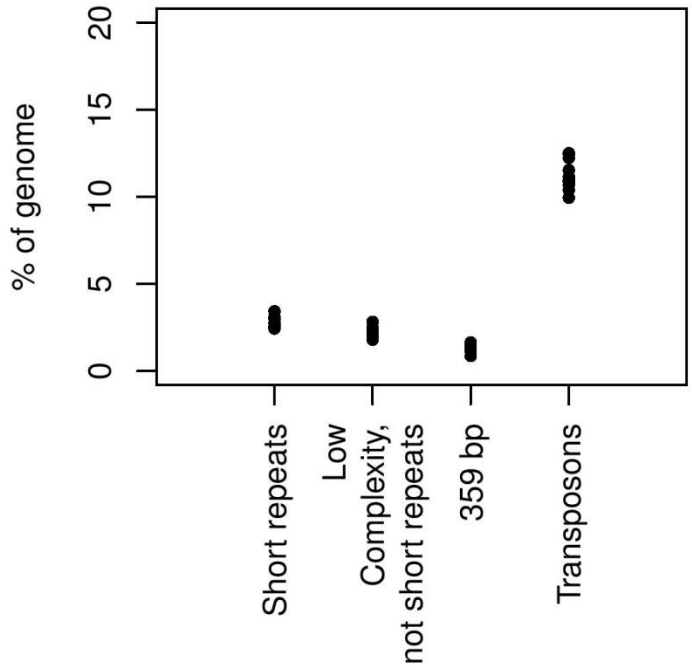


Figure 3.2. Percentage of four repeat classes in the genomes of 10 DGRP wild-caught fly lines

Paired-end reads from individual sequencing experiments were mapped to short repeat, 359 bp and transposon libraries and the percentage of total reads was calculated. The percentage of “low complexity, not short repeats” sequences was found by subtracting the percentage of short repeat sequences from the percentage of sequences with a low complexity score. Each value represents a median for a single fly line.

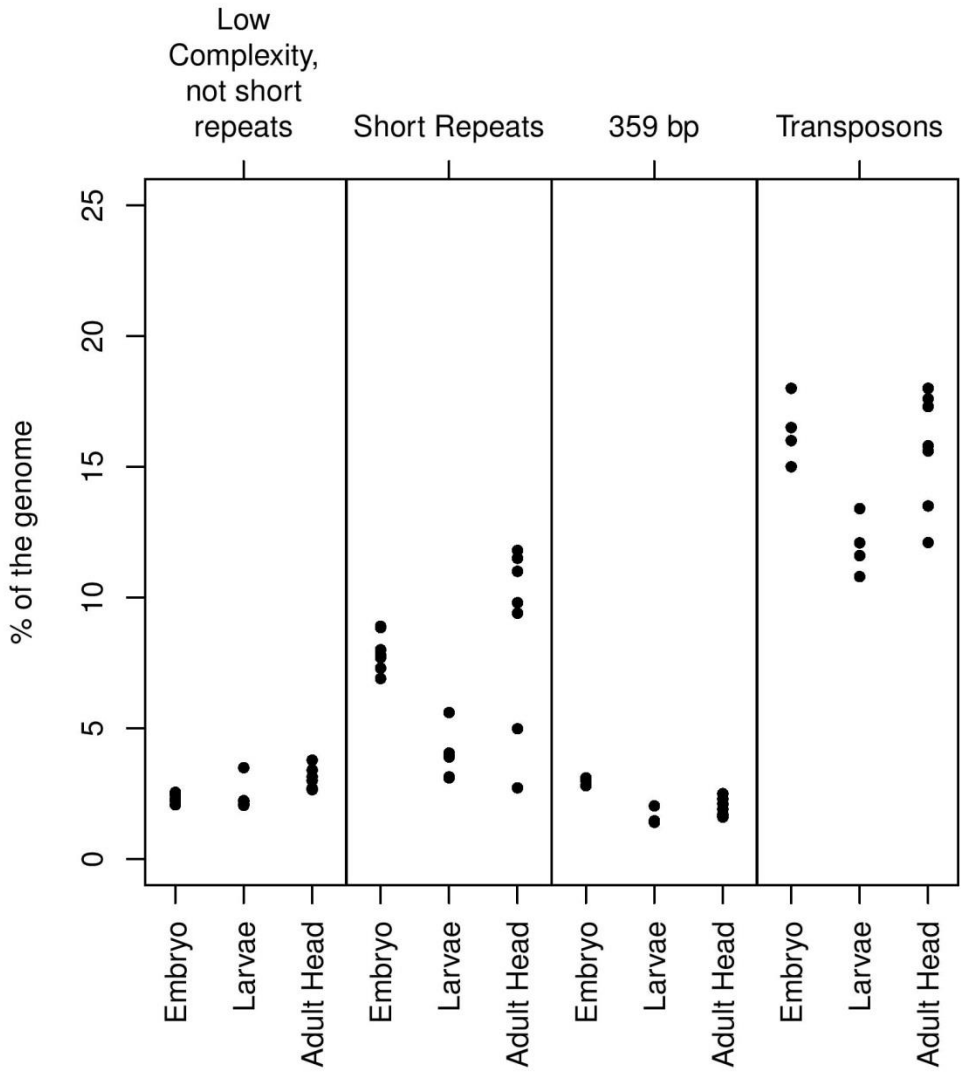


Figure 3.3. Percentage of four repeat classes in the genomes of embryo, larvae and adult modENCODE Oregon R flies

See the legend to Figure 3.2. Values for each sequencing experiment are shown, grouped by developmental stage.

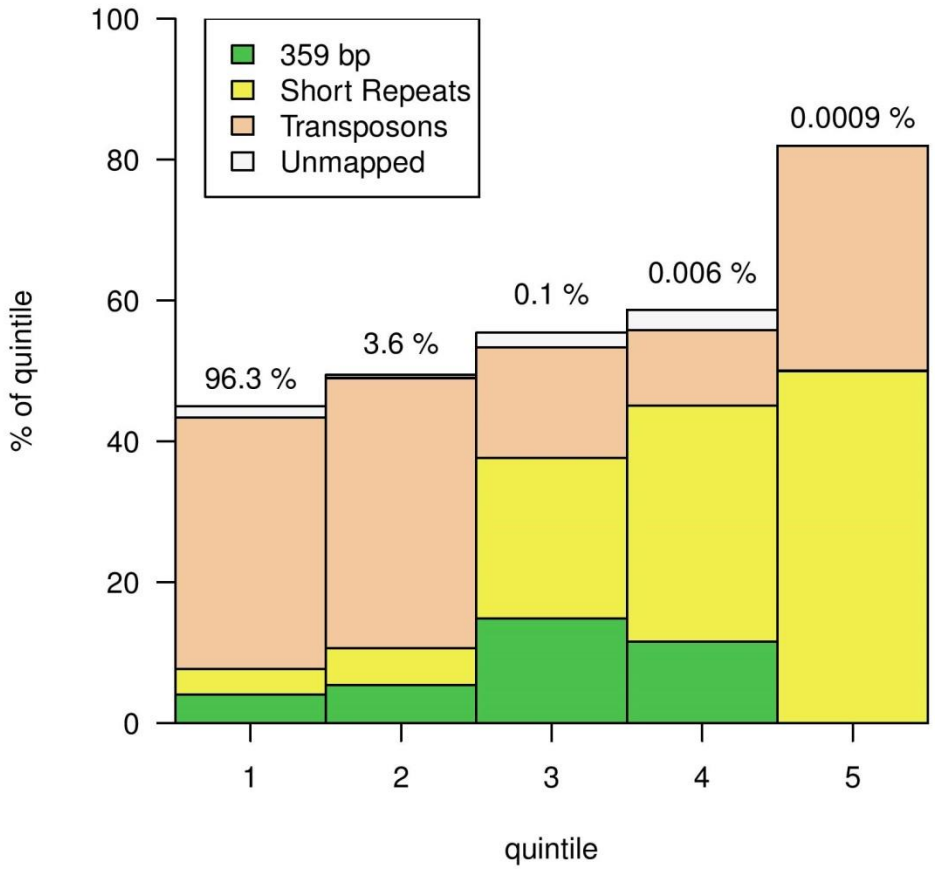


Figure 3.4. Averaged k-mer distribution for DGRP flies grouped by repeat class

K-mer frequencies were computed for each sequencing experiment. K-mers were split into quintiles by frequency, with quintile 5 being the most frequent. Each k-mer in the quintile was classified as mapping to short repeat, 359 bp repeat, transposon, and genome assembly or unmapped classes. The median number of k-mers that belonged to each of the repeat classes or that were unmapped for the 10 fly lines is plotted in each quintile. Numbers above indicate the percentage of the total number of k-mers falling into each quintile.

Figure 3.5. Examples of low complexity sequences that are not classified as short repeats

Sequences with a low complexity score were separated using Prinseq and mapped to the short repeat library. Low complexity sequences that did not map to the short repeat library were separated and some representative examples are shown.

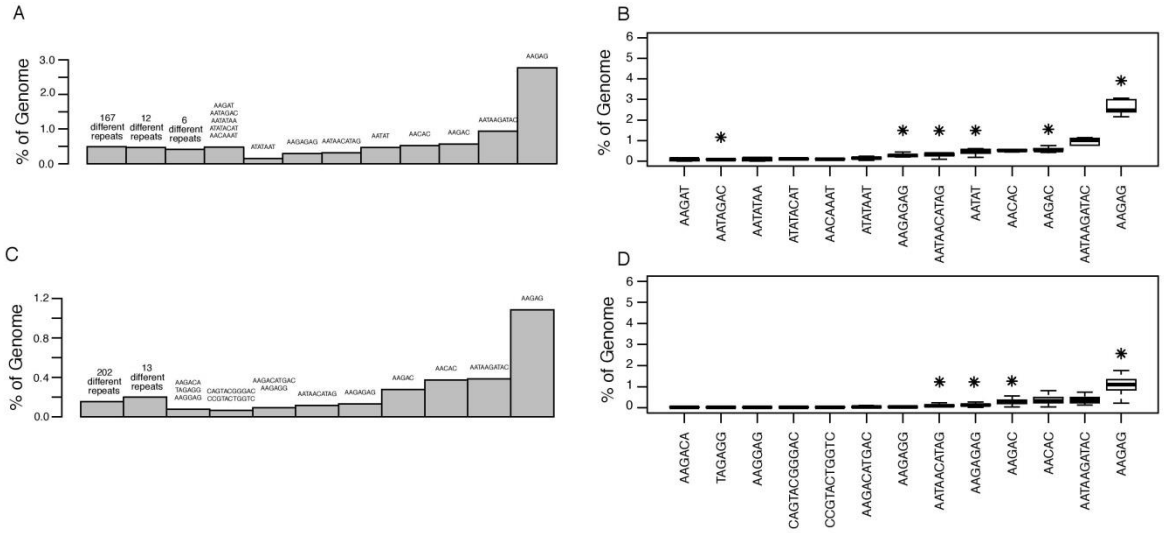


Figure 3.6. The most frequent short repeats in the fly genome

modENCODE A) and DGRP C) reads that mapped to a particular short repeat sequence were counted. Repeats were split into groups based on the number of reads that mapped to them. Repeat sequences in each group and the percentage of each group relative to the whole genome are indicated. B) and D) Repeats from the top groups shown in A) and C), respectively. The percentage of total reads mapped to each repeat over multiple experiments is shown in form of boxplots. An asterisk (*) marks each repeat group identified in previous studies based on cloning and sequencing of CsCl gradient bands.

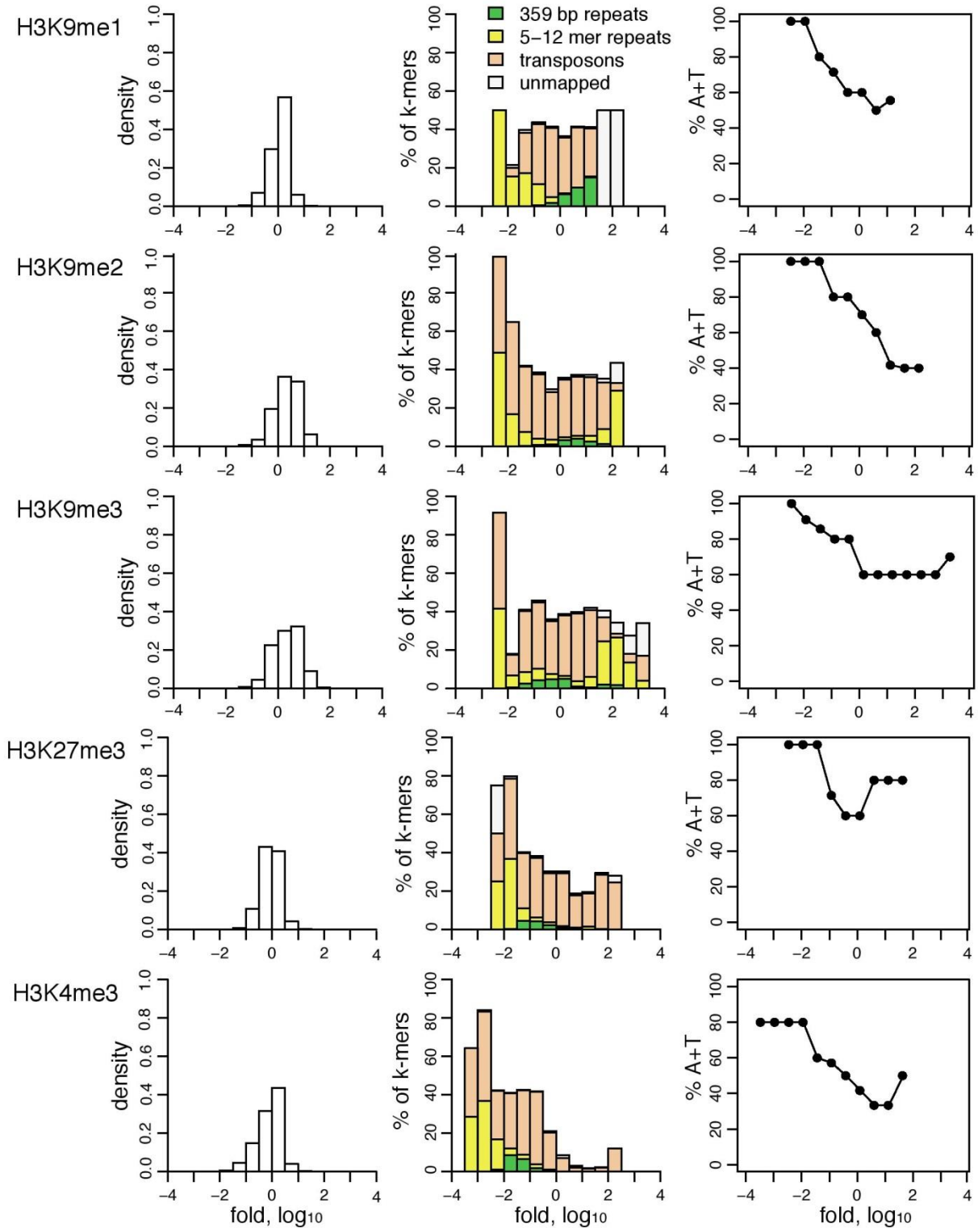


Figure 3.7**.** Association of epigenetic marks with repeat classes

For each Chip-Seq replicate k-mer frequencies were determined for both input and IP sequences. For each k-mer present in both input and IP least twice, the enrichment of IP over input was calculated. K-mers were grouped by enrichment. Left: Distribution of counts in each group. Middle: K-mers in each group were classified as short repeats, 359 bp, transposons, assembled genome or unmapped. The percentage of k-mers classified in each repeat class is shown. Some k-mers were classified as both short repeats and transposons, and they are included in both groups. Right: Percentage of A+T in short repeat k-mers. For all graphs the median between experimental replicates is shown. The number of replicates was two for all modifications.

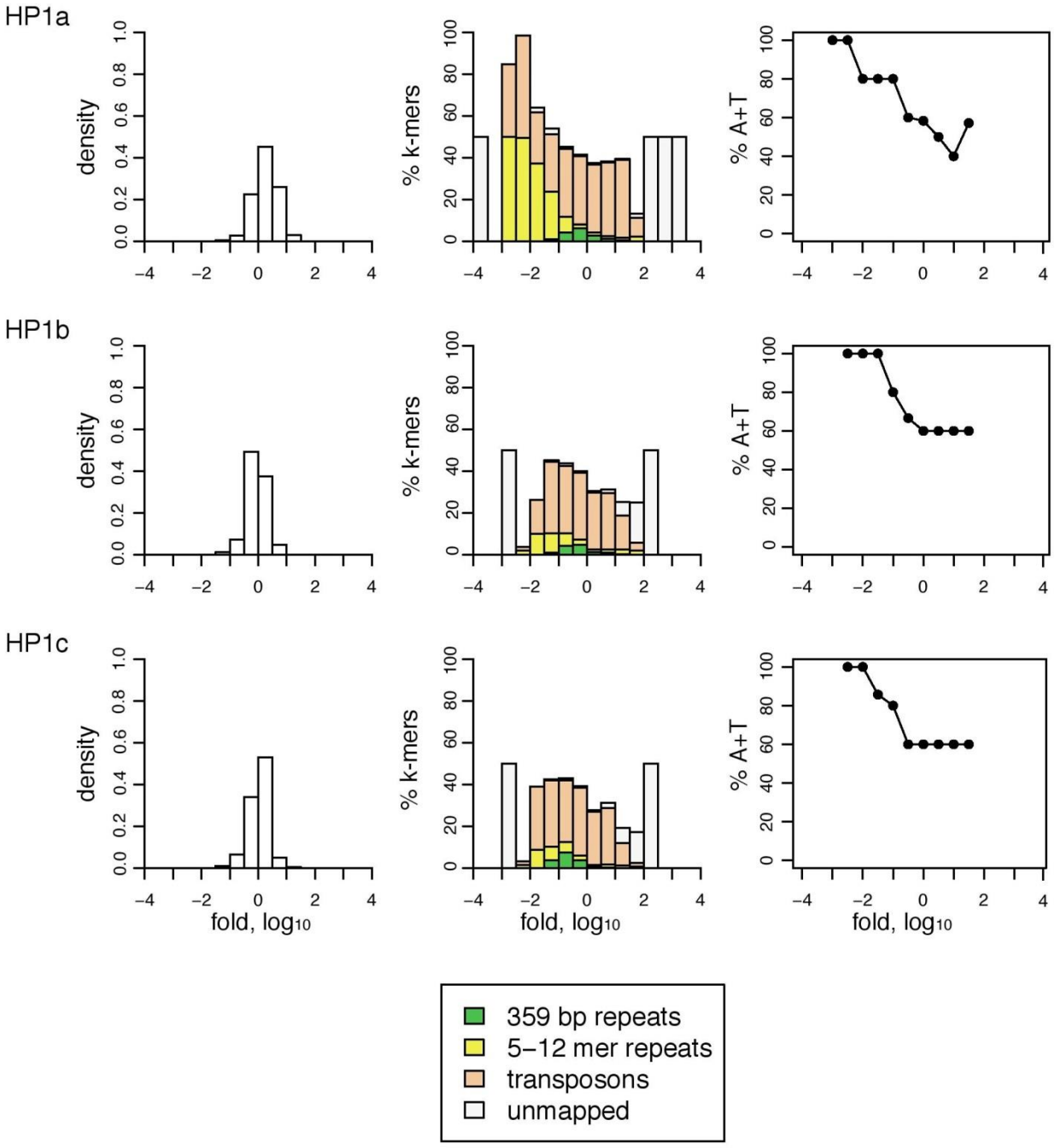


Figure 3.8. Association of HP1 proteins with repeat classes

Same analysis as in Figure 3.7 but for HP1 proteins. The number of replicates was 4 for HP1a, 1 for HP1c and 2 for HP1b.

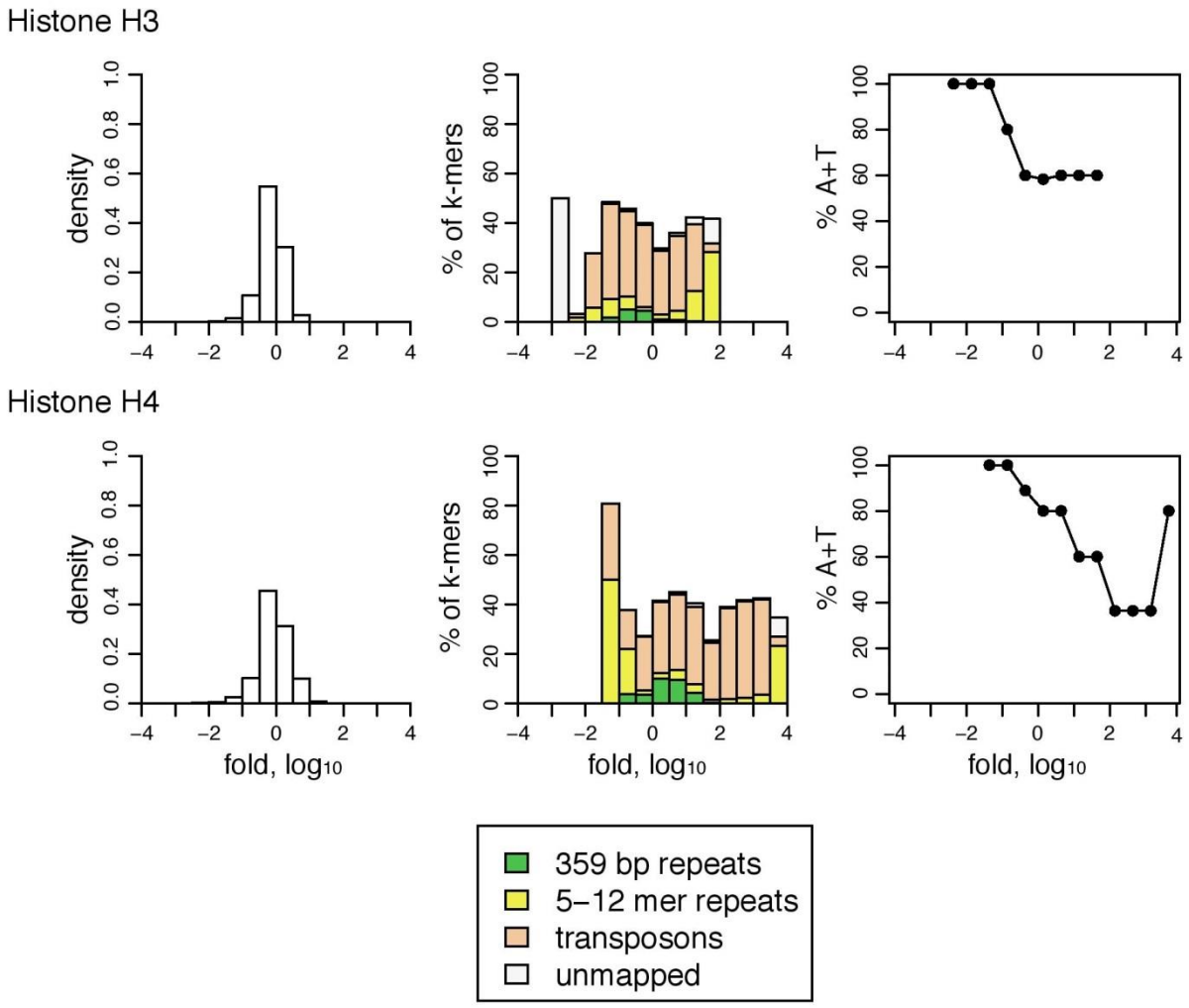


Figure 3.9**.** Association of histones H3 and H4 with repeat classes

Same analysis as in Figure 3.7 but for H3 and H4 histones. The number of replicates was 2 for both H3 and H4.

4 Future directions

In the work presented here I have analyzed chromatin landscape of the centromeres of S. cerevisiae and repeat sequences of D. melanogaster. Results obtained in chapter 2 allowed distinguishing between the models of centromeric nucleosomes that were recently proposed. Specifically, each centromere of S. cerevisiae consists of a single nucleosome which has a distinct form called hemisome. It incorporates one copy of the histone H4, H2A, H2B and H3 histone variant Cse4. Interestingly, histone Cse4 is also found in chromosome arms at low levels but doesn't form the hemisome structure there.

Instead it forms an octameric particle which is likely unstable and is quickly removed. In order to quantify association of repeat sequences of D. melanogaster with specific chromatin features I have developed a new analysis scheme. Repeat sequences were found to associate with specific histone modifications such as H3K9me3 and H3K9me2. Surprisingly, AT-rich short repeat sequences were found to be depleted of nucleosomes. It is expected that this new analysis scheme will be useful in other studies of unassembled parts of the genomes such as identification of centromeric sequences and changes of repeats in diseases.

Centromeres and pericentric chromatin are examples of non-coding regions of the genome that nevertheless are crucial for cell proliferation. Knowledge of the structure of centromeric nucleosomes is necessary for our understanding of the mechanisms of centromere formation, propagation and recognition by kinetochore complex. Analysis of consequences of the hemisome structure of centromeric nucleosome will be the subject of future research. Several ways that unique structure of this nucleosome specifies

centromere might be envisioned. For example, the right-handed wrap and unusual structure might alter the three-dimensional folding of chromatin at centromere creating a platform that is required for kinetochore assembly. In addition to that, hemisomes likely expose unique binding sites that are recognized by inner kinetochore proteins. It is also possible that the high proportion of A and T nucleotides that is common to many centromeres might favor formation of hemisomes instead of octamers, as such sequences are known to be stiff.

Since publication of the chapter 2 several studies concerning the structure of centromeric nucleosomes appeared. They employed various methods to investigate both yeast and human cells but arrived at different models. Measurement of DNA wrap in Cenp-A containing nucleosomes using chromatin IP followed by long read paired end sequencing showed trimodal distribution of fragments sizes[25]. These results were interpreted as evidence for partially unwrapped octameric particles at human centromeres. High resolution microscopy combined with fluorescence counting methods such as FRET and photobleaching were used to determine the number of cenH3 molecules at centromeric nucleosomes in yeast [29],[117] and human [118],[57] cells. Perhaps reflecting technical difficulties of these methods conflicting results were obtained. Two studies reported a single cenH3 molecule at the centromere through most of the duration of the cell cycle, with two molecules appearing at anaphase [117] or directly at S phase [57]. Other studies [29],[118] showed two molecules present throughout the cell cycle. Atomic force microscopy (AFM) was also employed to measure the height of cenH3 containing nucleosomes to infer stoichiometry. Previously cenH3 containing nucleosomes isolated from Drosophila cells were shown to have

smaller heights than H3 containing nucleosomes as measured by AFM [28] suggesting that they contain half the number of histones of the octamer. Recently, the height of in vitro reconstituted cenH3 octamers was measured by AFM and it was reported that it was about 30% smaller than H3 containing octamers[119]. This report challenged previous results, because it implied that reduced height is consistent with octameric form. However, the same measurements performed in two other independent laboratories did not show this size reduction[120],[121] of CenH3 containing octamers. Later examination of AFM results from many laboratories showed a wide variation in these types of measurements implying that AFM results alone cannot be used to infer stoichiometry[122].

To get a definitive answer on the structure of centromeric nucleosome a new method based on H4 anchored cleavage with localized hydroxyl radicals was used[123]. This in vivo method yields highly reproducible results and has the advantage over Chip-Seq method in that it doesn't require solubilization of chromatin and doesn't depend on the specificity of antibodies. Tightly localized hydroxyl radicals are produced around the unique residue in histone H4 at position 47, which was mutated to cysteine, and cleave DNA in the immediate proximity. CenH3 octamers and tetramers contain two H4 molecules, while the hemisome has only one. This means that these particles produce different cleavage patterns. Analysis of cleavage patterns at centromeres showed that they contain one molecule of H4. These patterns also showed that hemisome is present in two distinct rotational phases relative to CDE in equal proportions. Thus this study confirms results presented in chapter 2, and provides new details on rotational position of centromeric nucleosome.

Chip-seq remains the method of choice for the study of chromatin landscape. Analysis of repeat sequences in such datasets will be useful in the study of several problems. For example, centromeric sequences in various organisms can be identified using Chip-seq with antibodies against cenH3 histones and quantification of the enrichment of repeated sequences. Recently, a systematic study of epigenetic marks in healthy and cancerous human cells was carried out by the ENCODE project [124].This analysis was, however, performed only for the non-repetitive part of the genome. Quantification of repeat sequences in these experiments could produce unexpected findings and identify new roles of repeated sequences in disease.

Repeat sequences of different classes are abundant in eukaryotic organisms but specific knowledge of their role in the life of an organism remains absent. Roles of transposons in evolution [125] and gene regulation have been described [126], as well as expansion and contraction of trinucleotide repeat sequences in disease[127]. However, the role of the large stretches of satellite DNA remains relatively unexplored. The prevailing view is that simple repeat sequences result from slippage during DNA replication resulting in spontaneous expansion of such repeats [128], and that they don't perform any specific functions. While so far no function of satellite DNA has been shown experimentally, it is hard to accept its uselessness. After all, maintenance of such large quantities of DNA sequences involves replication and then compaction with epigenetic marks which requires substantial energetic resources. At the same time, at least some organisms are adapted to reducing amount of satellite DNA at some developmental stages, such as Drosophila larvae. I know of only one study undertaken in the last 40 years that systematically explored this problem[89]. In this study flies with deletions of

90% of heterochromatin on chromosome X were constructed. Flies exhibited no obvious phenotype, and they also produced viable progeny. However the authors observed lower genetic diversity that resulted from fewer recombination events during meiosis. They proposed that the function of noncoding heterochromatin was to space centromeres and gene regions, so that suppression of recombination near centromere (the “centromere effect”) does not affect protein-coding regions. Using modern genome editing technologies such as CRISPR [129] it will soon be possible to systematically delete large stretches of satellite DNA. It is likely that the consequences of such deletions will be slight, with no immediately observable cell phenotype.

Satellite sequences might play a role in the spatial organization of the genome. Repeat sequences are occupied by very well-positioned nucleosomes (Figure 4.1), and this can perhaps influence motions and mechanical properties of mitotic chromosomes, or, as proposed, affect meiotic recombination. Modern techniques such as 3D chromosome capture (Hi-C) and high resolution microscopy such as STORM can be used to study these effects.

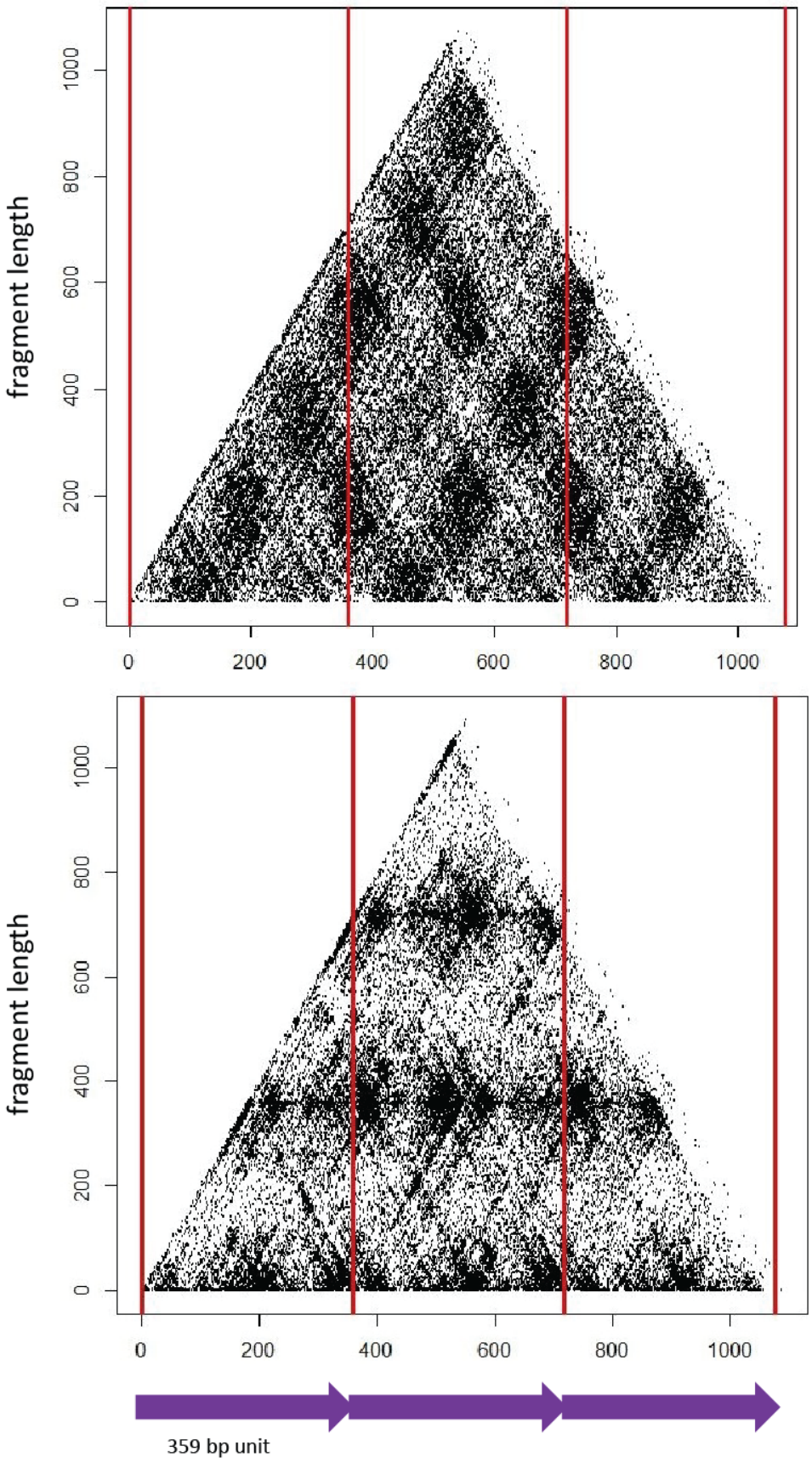

fragment length
0
200
400
600
800
1000
fragment length
0
200
400
600
800
1000
359 bp unit

Figure 4.1 Nucleosomes are well phased on 359 bp repeat sequences

DNA from nuclei digested with MNase of S2 Drosophila cells was sequenced and aligned to tandem of three 359 bp repeat units and plotted as midpoint map (Top). Nucleosomes are well-positioned on 359 bp repeats as evident from clusters of dots. This phasing is not due to an MNase sequence bias because the phased pattern is not present in the naked DNA digestion control (Bottom).

References

1. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ: **Crystal structure of the nucleosome core particle at 2.8 A resolution**. *Nature* 1997, **389**(6648):251-260.
2. Talbert PB, Henikoff S: **Histone variants--ancient wrap artists of the epigenome**. *Nat Rev Mol Cell Biol* 2010, **11**(4):264-275.
3. Zentner GE, Henikoff S: **Regulation of nucleosome dynamics by histone modifications**. *Nat Struct Mol Biol* 2013, **20**(3):259-266.
4. Allfrey VG, Faulkner R, Mirsky AE: **Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis**. *Proc Natl Acad Sci U S A* 1964, **51**:786-794.
5. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The sequence of the human genome**. *Science* 2001, **291**(5507):1304-1351.
6. Cech TR, Steitz JA: **The Noncoding RNA Revolution-Trashing Old Rules to Forge New Ones**. *Cell* 2014, **157**(1):77-94.
7. Stephens AD, Haggerty RA, Vasquez PA, Vicci L, Snider CE, Shi F, Quammen C, Mullins C, Haase J, Taylor RM, 2nd *et al*: **Pericentric chromatin loops function as a nonlinear spring in mitotic force balance**. *J Cell Biol* 2013, **200**(6):757-772.
8. Henikoff S, Ahmad K, Malik HS: **The centromere paradox: stable inheritance with rapidly evolving DNA**. *Science* 2001, **293**(5532):1098-1102.
9. Winey M, Mamay CL, O'Toole ET, Mastronarde DN, Giddings TH, Jr., McDonald KL, McIntosh JR: **Three-dimensional ultrastructural analysis of the Saccharomyces cerevisiae mitotic spindle**. *J Cell Biol* 1995, **129**(6):1601-1615.
10. Blower MD, Sullivan BA, Karpen GH: **Conserved organization of centromeric chromatin in flies and humans**. *Dev Cell* 2002, **2**(3):319-330.
11. Melters DP, Paliulis LV, Korf IF, Chan SW: **Holocentric chromosomes: convergent evolution, meiotic adaptations, and genomic analysis**. *Chromosome Res* 2012, **20**(5):579-593.
12. Westermann S, Cheeseman IM, Anderson S, Yates JR, 3rd, Drubin DG, Barnes G: **Architecture of the budding yeast kinetochore reveals a conserved molecular core**. *J Cell Biol* 2003, **163**(2):215-222.
13. Steiner FA, Henikoff S: **Holocentromeres are dispersed point centromeres localized at transcription factor hotspots**. *eLife* 2014, **3**:e02025.
14. Carbon J, Clarke L: **Structural and functional analysis of a yeast centromere (CEN3)**. *J Cell Sci Suppl* 1984, **1**:43-58.
15. Neitz M, Carbon J: **Identification and characterization of the centromere from chromosome XIV in Saccharomyces cerevisiae**. *Mol Cell Biol* 1985, **5**(11):2887-2893.
16. Fitzgerald-Hayes M, Buhler JM, Cooper TG, Carbon J: **Isolation and subcloning analysis of functional centromere DNA (CEN11) from Saccharomyces cerevisiae chromosome XI**. *Mol Cell Biol* 1982, **2**(1):82-87.
17. Jiang W, Carbon J: **Molecular analysis of the budding yeast centromere/kinetochore**. *Cold Spring Harb Symp Quant Biol* 1993, **58**:669-676.
18. Niedenthal R, Stoll R, Hegemann JH: **In vivo characterization of the Saccharomyces cerevisiae centromere DNA element I, a binding site for the helix-loop-helix protein CPF1**. *Mol Cell Biol* 1991, **11**(7):3545-3553.

19. Jehn B, Niedenthal R, Hegemann JH: **In vivo analysis of the Saccharomyces cerevisiae centromere CDEIII sequence: requirements for mitotic chromosome segregation**. *Mol Cell Biol* 1991, **11**(10):5212-5221.
20. McGrew J, Diehl B, Fitzgerald-Hayes M: **Single base-pair mutations in centromere element III cause aberrant chromosome segregation in Saccharomyces cerevisiae**. *Mol Cell Biol* 1986, **6**(2):530-538.
21. Drechsler H, McAinsh AD: **Exotic mitotic mechanisms**. *Open biology* 2012, **2**(12):120140.
22. Black BE, Cleveland DW: **Epigenetic Centromere Propagation and the Nature of CENP-A Nucleosomes**. *Cell* 2011, **144**(4):471-479.
23. Camahort R, Shivaraju M, Mattingly M, Li B, Nakanishi S, Zhu D, Shilatifard A, Workman JL, Gerton JL: **Cse4 is part of an octameric nucleosome in budding yeast**. *Mol Cell* 2009, **35**(6):794-805.
24. Sekulic N, Bassett EA, Rogers DJ, Black BE: **The structure of (CENP-A-H4)(2) reveals physical features that mark centromeres**. *Nature* 2010, **467**(7313):347-351.
25. Hasson D, Panchenko T, Salimian KJ, Salman MU, Sekulic N, Alonso A, Warburton PE, Black BE: **The octamer is the major form of CENP-A nucleosomes at human centromeres**. *Nat Struct Mol Biol* 2013, **20**(6):687-695.
26. Lochmann B, Ivanov D: **Histone H3 localizes to the centromeric DNA in budding yeast**. *PLoS Genet* 2012, **8**(5):e1002739.
27. Furuyama T, Henikoff S: **Centromeric nucleosomes induce positive DNA supercoils**. *Cell* 2009, **138**(1):104-113.
28. Dalal Y, Wang H, Lindsay S, Henikoff S: **Tetrameric structure of centromeric nucleosomes in interphase Drosophila cells**. *PLoS Biol* 2007, **5**(8):e218.
29. Aravamudhan P, Felzer-Kim I, Joglekar AP: **The budding yeast point centromere associates with two Cse4 molecules during mitosis**. *Curr Biol* 2013, **23**(9):770-774.
30. Mizuguchi G, Xiao H, Wisniewski J, Smith MM, Wu C: **Nonhistone Scm3 and histones CenH3-H4 assemble the core of centromere-specific nucleosomes**. *Cell* 2007, **129**(6):1153-1164.
31. Xiao H, Mizuguchi G, Wisniewski J, Huang Y, Wei D, Wu C: **Nonhistone Scm3 binds to AT-rich DNA to organize atypical centromeric nucleosome of budding yeast**. *Mol Cell* 2011, **43**(3):369-380.
32. Bloom K, Joglekar A: **Towards building a chromosome segregation machine**. *Nature* 2010, **463**(7280):446-456.
33. Lyubchenko YL: **Centromere chromatin: a loose grip on the nucleosome?** *Nat Struct Mol Biol* 2014, **21**(1):8.
34. Furuyama T, Codomo CA, Henikoff S: **Reconstitution of hemisomes on budding yeast centromeric DNA**. *Nucleic Acids Res* 2013, **41**(11):5769-5783.
35. Brutlag DL: **Molecular arrangement and evolution of heterochromatic DNA**. *Annu Rev Genet* 1980, **14**:121-144.
36. Pidoux AL, Allshire RC: **The role of heterochromatin in centromere function**. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 2005, **360**(1455):569-579.
37. Bannister AJ, Zegerman P, Partridge JF, Miska EA, Thomas JO, Allshire RC, Kouzarides T: **Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain**. *Nature* 2001, **410**(6824):120-124.
38. Kohne DE: **Evolution of higher-organism DNA**. *Quarterly reviews of biophysics* 1970, **3**(3):327-375.
39. Corneo G, Ginelli E, Polli E: **Repeated sequences in human DNA**. *J Mol Biol* 1970, **48**(2):319-327.

40. Miller JR, Koren S, Sutton G: **Assembly algorithms for next-generation sequencing data**. *Genomics* 2010, **95**(6):315-327.
41. Heitz E: **Die somatische Heteropyknose bei Drosophila melanogaster und ihre genetische Bedeutung (Cytologische Untersuchungen an Dipteren1, 11)**. *Z Zellforsch mikrosk Anat* 1934, **20**:237-287.
42. Waring M, Britten RJ: **Nucleotide sequence repetition: a rapidly reassociating fraction of mouse DNA**. *Science* 1966, **154**(3750):791-794.
43. Britten RJ, Kohne DE: **Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms**. *Science* 1968, **161**(3841):529-540.
44. Laird CD, McCarthy BJ: **Molecular characterization of the Drosophila genome**. *Genetics* 1969, **63**(4):865-882.
45. Dickson E, Boyd JB, Laird CD: **Sequence diversity of polytene chromosome DNA from Drosophila hydei**. *J Mol Biol* 1971, **61**(3):615-627.
46. Laird CD, McCarthy BJ: **Magnitude of interspecific nucleotide sequence variability in Drosophila**. *Genetics* 1968, **60**(2):303-322.
47. Lohe AR, Brutlag DL: **Multiplicity of satellite DNA sequences in Drosophila melanogaster**. *Proc Natl Acad Sci U S A* 1986, **83**(3):696-700.
48. Mackay TF, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM *et al*: **The Drosophila melanogaster Genetic Reference Panel**. *Nature* 2012, **482**(7384):173-178.
49. Fondon JW, 3rd, Martin A, Richards S, Gibbs RA, Mittelman D: **Analysis of microsatellite variation in Drosophila melanogaster with population-scale genome sequencing**. *PLoS One* 2012, **7**(3):e33036.
50. Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T *et al*: **Comprehensive analysis of the chromatin landscape in Drosophila melanogaster**. *Nature* 2011, **471**(7339):480-485.
51. Riddle NC, Minoda A, Kharchenko PV, Alekseyenko AA, Schwartz YB, Tolstorukov MY, Gorchakov AA, Jaffe JD, Kennedy C, Linder-Basso D *et al*: **Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin**. *Genome Res* 2011, **21**(2):147-163.
52. Malik HS, Henikoff S: **Major evolutionary transitions in centromere complexity**. *Cell* 2009, **138**(6):1067-1082.
53. Rudd MK, Willard HF: **Analysis of the centromeric regions of the human genome assembly**. *Trends Genet* 2004, **20**(11):529-533.
54. Warburton PE: **Chromosomal dynamics of human neocentromere formation**. *Chromosome Res* 2004, **12**(6):617-626.
55. WC DWPE: **The Kinetochore**. *Springer, Berlin* 2009.
56. Hu H, Liu Y, Wang M, Fang J, Huang H, Yang N, Li Y, Wang J, Yao X, Shi Y *et al*: **Structure of a CENP-A-histone H4 heterodimer in complex with chaperone HJURP**. *Genes Dev* 2011.
57. Bui M, Dimitriadis EK, Hoischen C, An E, Quenet D, Giebe S, Nita-Lazar A, Diekmann S, Dalal Y: **Cell-cycle-dependent structural transitions in the human CENP-A nucleosome in vivo**. *Cell* 2012, **150**(2):317-326.
58. Dimitriadis EK, Weber C, Gill RK, Diekmann S, Dalal Y: **Tetrameric organization of vertebrate centromeric nucleosomes**. *Proc Natl Acad Sci U S A* 2010, **107**(47):20317-20322.
59. Tachiwana H, Kagawa W, Kurumizaka H: **Comparison between the CENP-A and histone H3 structures in nucleosomes**. *Nucleus* 2012, **3**(1).

60. Bancaud A, Wagner G, Conde ESN, Lavelle C, Wong H, Mozziconacci J, Barbi M, Sivolob A, Le Cam E, Mouawad L *et al*: **Nucleosome chiral transition under positive torsional stress in single chromatin fibers**. *Mol Cell* 2007, **27**(1):135-147.
61. Dechassa ML, Wyns K, Li M, Hall MA, Wang MD, Luger K: **Structure and Scm3-mediated assembly of budding yeast centromeric nucleosomes**. *Nat Commun* 2011, **2**:313.
62. Armache KJ, Garlick JD, Canzio D, Narlikar GJ, Kingston RE: **Structural basis of silencing: Sir3 BAH domain in complex with a nucleosome at 3.0 A resolution**. *Science* 2011, **334**(6058):977-982.
63. Henikoff JG, Belsky JA, Krassovsky K, MacAlpine DM, Henikoff S: **Epigenome characterization at single base-pair resolution**. *Proc Natl Acad Sci U S A* 2011, **108**(45):18318-18323.
64. Densmore L, Payne WE, Fitzgerald-Hayes M: **In vivo genomic footprint of a yeast centromere**. *Mol Cell Biol* 1991, **11**(1):154-165.
65. Dalal Y, Furuyama T, Vermaak D, Henikoff S: **Structure, dynamics, and evolution of centromeric nucleosomes**. *Proc Natl Acad Sci U S A* 2007, **104**(41):15974-15981.
66. Lefrancois P, Euskirchen GM, Auerbach RK, Rozowsky J, Gibson T, Yellman CM, Gerstein M, Snyder M: **Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing**. *BMC Genomics* 2009, **10**:37.
67. Hemmerich P, Weidtkamp-Peters S, Hoischen C, Schmiedeberg L, Erliandri I, Diekmann S: **Dynamics of inner kinetochore assembly and maintenance in living cells**. *J Cell Biol* 2008, **180**(6):1101-1114.
68. Kent NA, Adams S, Moorhouse A, Paszkiewicz K: **Chromatin particle spectrum analysis: a method for comparative chromatin structure analysis using paired-end mode next-generation DNA sequencing**. *Nucleic Acids Res* 2011, **39**(5):e26.
69. Hemmerich P, Stoyan T, Wieland G, Koch M, Lechner J, Diekmann S: **Interaction of yeast kinetochore proteins with centromere-protein/transcription factor Cbf1**. *Proc Natl Acad Sci U S A* 2000, **97**(23):12583-12588.
70. Niedenthal RK, Sen-Gupta M, Wilmen A, Hegemann JH: **Cpf1 protein induced bending of yeast centromere DNA element I**. *Nucleic Acids Res* 1993, **21**(20):4726-4733.
71. Pietrasanta LI, Thrower D, Hsieh W, Rao S, Stemmann O, Lechner J, Carbon J, Hansma H: **Probing the Saccharomyces cerevisiae centromeric DNA (CEN DNA)-binding factor 3 (CBF3) kinetochore complex by using atomic force microscopy**. *Proc Natl Acad Sci U S A* 1999, **96**(7):3757-3762.
72. MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E: **An improved map of conserved regulatory sites for Saccharomyces cerevisiae**. *BMC bioinformatics* 2006, **7**:113.
73. Bloom KS, Amaya E, Carbon J, Clarke L, Hill A, Yeh E: **Chromatin conformation of yeast centromeres**. *J Cell Biol* 1984, **99**(5):1559-1568.
74. Hsieh TF, Ibarra CA, Silva P, Zemach A, Eshed-Williams L, Fischer RL, Zilberman D: **Genome-wide demethylation of Arabidopsis endosperm**. *Science* 2009, **324**(5933):1451-1454.
75. Floer M, Wang X, Prabhu V, Berrozpe G, Narayan S, Spagna D, Alvarez D, Kendall J, Krasnitz A, Stepansky A *et al*: **A RSC/nucleosome complex determines chromatin architecture and facilitates activator binding**. *Cell* 2010, **141**(3):407-418.
76. Dion MF, Kaplan T, Kim M, Buratowski S, Friedman N, Rando OJ: **Dynamics of replication-independent histone turnover in budding yeast**. *Science* 2007, **315**(5817):1405-1408.
77. da Rosa JL, Holik J, Green EM, Rando OJ, Kaufman PD: **Overlapping regulation of CenH3 localization and histone H3 turnover by CAF-1 and HIR proteins in Saccharomyces cerevisiae**. *Genetics* 2011, **187**(1):9-19.
78. Meluh PB, Koshland D: **Budding yeast centromere composition and assembly as revealed by in vivo cross-linking**. *Genes Dev* 1997, **11**(24):3401-3412.

79. Sandman K, Reeve JN: **Archaeal histones and the origin of the histone fold**. *Curr Opin Microbiol* 2006, **9**(5):520-525.
80. Luk E, Ranjan A, Fitzgerald PC, Mizuguchi G, Huang Y, Wei D, Wu C: **Stepwise histone replacement by SWR1 requires dual activation with histone H2A.Z and canonical nucleosome**. *Cell* 2010, **143**(5):725-736.
81. Takahashi K, Murakami S, Chikashige Y, Funabiki H, Niwa O, Yanagida M: **A low copy number central sequence with strict symmetry and unusual chromatin structure in fission yeast centromere**. *Mol Biol Cell* 1992, **3**(7):819-835.
82. Erhardt S, Mellone BG, Betts CM, Zhang W, Karpen GH, Straight AF: **Genome-wide analysis reveals a cell cycle-dependent mechanism controlling centromere propagation**. *J Cell Biol* 2008, **183**(5):805-818.
83. Chen XF, Kuryan B, Kitada T, Tran N, Li JY, Kurdistani S, Grunstein M, Li B, Carey M: **The Rpd3 core complex is a chromatin stabilization module**. *Curr Biol* 2012, **22**(1):56-63.
84. Ortiz J, Stemmann O, Rank S, Lechner J: **A putative protein complex consisting of Ctf19, Mcm21, and Okp1 represents a missing link in the budding yeast kinetochore**. *Genes Dev* 1999, **13**(9):1140-1155.
85. Dalal Y, Bui M: **Down the rabbit hole of centromere assembly and dynamics**. *Curr Opin Cell Biol* 2010, **22**(3):392-402.
86. Yunis JJ, Yasmineh WG: **Heterochromatin, satellite DNA, and cell function. Structural DNA of eucaryotes may support and protect genes and aid in speciation**. *Science* 1971, **174**(4015):1200-1209.
87. Goldberg AD, Banaszynski LA, Noh KM, Lewis PW, Elsaesser SJ, Stadler S, Dewell S, Law M, Guo X, Li X *et al*: **Distinct factors control histone variant H3.3 localization at specific genomic regions**. *Cell* 2010, **140**(5):678-691.
88. Nekrasov M, Amrichova J, Parker BJ, Soboleva TA, Jack C, Williams R, Huttley GA, Tremethick DJ: **Histone H2A.Z inheritance during the cell cycle and its impact on promoter organization and dynamics**. *Nat Struct Mol Biol* 2012.
89. Yamamoto M, Miklos GL: **Genetic studies on heterochromatin in Drosophila melanogaster and their implications for the functions of satellite DNA**. *Chromosoma* 1978, **66**(1):71-98.
90. Hayden KE, Willard HF: **Composition and organization of active centromere sequences in complex genomes**. *BMC Genomics* 2012, **13**:324.
91. Brutlag D, Appels R, Dennis ES, Peacock WJ: **Highly repeated DNA in Drosophila melanogaster**. *J Mol Biol* 1977, **112**(1):31-47.
92. Peacock WJ, Brutlag D, Goldring E, Appels R, Hinton CW, Lindsley DL: **The organization of highly repeated DNA sequences in Drosophila melanogaster chromosomes**. *Cold Spring Harb Symp Quant Biol* 1974, **38**:405-416.
93. Hoskins RA, Smith CD, Carlson JW, Carvalho AB, Halpern A, Kaminker JS, Kennedy C, Mungall CJ, Sullivan BA, Sutton GG *et al*: **Heterochromatic sequences in a Drosophila whole-genome shotgun assembly**. *Genome Biol* 2002, **3**(12):RESEARCH0085.
94. Yasuhara JC, Wakimoto BT: **Molecular landscape of modified histones in Drosophila heterochromatic genes and euchromatin-heterochromatin transition zones**. *PLoS Genet* 2008, **4**(1):e16.
95. Ruthenburg AJ, Allis CD, Wysocka J: **Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark**. *Mol Cell* 2007, **25**(1):15-30.
96. Ebert A, Lein S, Schotta G, Reuter G: **Histone modification and the control of heterochromatic gene silencing in Drosophila**. *Chromosome Res* 2006, **14**(4):377-392.

97. Kuhn GC, Kuttler H, Moreira-Filho O, Heslop-Harrison JS: **The 1.688 repetitive DNA of Drosophila: concerted evolution at different genomic scales and association with genes**. *Molecular biology and evolution* 2012, **29**(1):7-11.
98. Gall JG, Cohen EH, Polan ML: **Reptitive DNA sequences in drosophila**. *Chromosoma* 1971, **33**(3):319-344.
99. Hammond MP, Laird CD: **Chromosome structure and DNA replication in nurse and follicle cells of Drosophila melanogaster**. *Chromosoma* 1985, **91**(3-4):267-278.
100. Bate M, Martinez Arias A: **The Development of Drosophila melanogaster**. Plainview, N.Y.: Cold Spring Harbor Laboratory Press; 1993.
101. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A: **Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries**. *Genome Biol* 2011, **12**(2):R18.
102. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.
103. Dechering KJ, Cuelenaere K, Konings RN, Leunissen JA: **Distinct frequency-distributions of homopolymeric DNA tracts in different genomes**. *Nucleic Acids Res* 1998, **26**(17):4056-4062.
104. Struhl K, Segal E: **Determinants of nucleosome positioning**. *Nat Struct Mol Biol* 2013, **20**(3):267-273.
105. Pinheiro I, Margueron R, Shukeir N, Eisold M, Fritzsch C, Richter FM, Mittler G, Genoud C, Goyama S, Kurokawa M *et al*: **Prdm3 and Prdm16 are H3K9me1 methyltransferases required for mammalian heterochromatin integrity**. *Cell* 2012, **150**(5):948-960.
106. Veiseth SV, Rahman MA, Yap KL, Fischer A, Egge-Jacobsen W, Reuter G, Zhou MM, Aalen RB, Thorstensen T: **The SUVR4 histone lysine methyltransferase binds ubiquitin and converts H3K9me1 to H3K9me3 on transposon chromatin in Arabidopsis**. *PLoS Genet* 2011, **7**(3):e1001325.
107. Lachner M, O'Carroll D, Rea S, Mechtler K, Jenuwein T: **Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins**. *Nature* 2001, **410**(6824):116-120.
108. Smothers JF, Henikoff S: **The hinge and chromo shadow domain impart distinct targeting of HP1-like proteins**. *Mol Cell Biol* 2001, **21**(7):2555-2569.
109. Blattes R, Monod C, Susbielle G, Cuvier O, Wu JH, Hsieh TS, Laemmli UK, Kas E: **Displacement of D1, HP1 and topoisomerase II from satellite heterochromatin by a specific polyamide**. *EMBO J* 2006, **25**(11):2397-2408.
110. Perrini B, Piacentini L, Fanti L, Altieri F, Chichiarelli S, Berloco M, Turano C, Ferraro A, Pimpinelli S: **HP1 controls telomere capping, telomere elongation, and telomere silencing by two different mechanisms in Drosophila**. *Mol Cell* 2004, **15**(3):467-476.
111. Iyer V, Struhl K: **Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure**. *EMBO J* 1995, **14**(11):2570-2579.
112. Vlahovicek K, Kajan L, Pongor S: **DNA analysis servers: plot.it, bend.it, model.it and IS**. *Nucleic Acids Res* 2003, **31**(13):3686-3687.
113. Levinger L, Varshavsky A: **Protein D1 preferentially binds A + T-rich DNA in vitro and is a component of Drosophila melanogaster nucleosomes containing A + T-rich satellite DNA**. *Proc Natl Acad Sci U S A* 1982, **79**(23):7152-7156.
114. Levinger L: **Nucleosomal structure of two Drosophila melanogaster simple satellites**. *J Biol Chem* 1985, **260**(21):11799-11804.
115. Marcais G, Kingsford C: **A fast, lock-free approach for efficient parallel counting of occurrences of k-mers**. *Bioinformatics* 2011, **27**(6):764-770.
116. Schmieder R, Edwards R: **Quality control and preprocessing of metagenomic datasets**. *Bioinformatics* 2011, **27**(6):863-864.

117. Shivaraju M, Unruh JR, Slaughter BD, Mattingly M, Berman J, Gerton JL: **Cell-cycle-coupled structural oscillation of centromeric nucleosomes in yeast**. *Cell* 2012, **150**(2):304-316.
118. Padeganeh A, Ryan J, Boisvert J, Ladouceur AM, Dorn JF, Maddox PS: **Octameric CENP-A nucleosomes are present at human centromeres throughout the cell cycle**. *Curr Biol* 2013, **23**(9):764-769.
119. Miell MD, Fuller CJ, Guse A, Barysz HM, Downes A, Owen-Hughes T, Rappsilber J, Straight AF, Allshire RC: **CENP-A confers a reduction in height on octameric nucleosomes**. *Nat Struct Mol Biol* 2013, **20**(6):763-765.
120. Walkiewicz MP, Dimitriadis EK, Dalal Y: **CENP-A octamers do not confer a reduction in nucleosome height by AFM**. *Nat Struct Mol Biol* 2014, **21**(1):2-3.
121. Codomo CA, Furuyama T, Henikoff S: **CENP-A octamers do not confer a reduction in nucleosome height by AFM**. *Nat Struct Mol Biol* 2014, **21**(1):4-5.
122. Miell MD, Straight AF, Allshire RC: **Reply to "CENP-A octamers do not confer a reduction in nucleosome height by AFM"**. *Nat Struct Mol Biol* 2014, **21**(1):5-8.
123. Henikoff S, Ramachandran S, Krassovsky K, Bryson TD, Codomo CA, Brogaard K, Widom J, Wang JP, Henikoff JG: **The budding yeast Centromere DNA Element II wraps a stable Cse4 hemisome in either orientation in vivo**. *eLife* 2014, **3**:e01861.
124. Consortium EP, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M: **An integrated encyclopedia of DNA elements in the human genome**. *Nature* 2012, **489**(7414):57-74.
125. 17363976, Feschotte C, Pritham EJ: **DNA transposons and the evolution of eukaryotic genomes**. *Annu Rev Genet* 2007, **41**:331-368.
126. Slotkin RK, Martienssen R: **Transposable elements and the epigenetic regulation of the genome**. *Nat Rev Genet* 2007, **8**(4):272-285.
127. Cummings CJ, Zoghbi HY: **Trinucleotide repeats: mechanisms and pathophysiology**. *Annu Rev Genomics Hum Genet* 2000, **1**:281-328.
128. Levinson G, Gutman GA: **Slipped-strand mispairing: a major mechanism for DNA sequence evolution**. *Molecular biology and evolution* 1987, **4**(3):203-221.
129. Bassett AR, Tibbit C, Ponting CP, Liu JL: **Mutagenesis and homologous recombination in Drosophila cell lines using CRISPR/Cas9**. *Biology open* 2014, **3**(1):42-49.