*Journal of Biomolecular Structure and Dynamics*, 2014
http://dx.doi.org/10.1080/07391102.2014.938700

# Strong nucleosomes of mouse genome including recovered centromeric sequences

Bilal F. Salih[a,b]*, Vladimir B. Teif[c], Vijay Tripathi[a] and Edward N. Trifonov[a]

[a]*Genome Diversity Center, Institute of Evolution, University of Haifa, Mount Carmel, Haifa31905, Israel;* [b]*Department of Computer Science, University of Haifa, Mount Carmel, Haifa 31905, Israel;* [c]*German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, Heidelberg 69120, Germany*

Communicated by Ramaswamy H. Sarma

(*Received 26 April 2014; accepted 23 June 2014*)

Recently discovered strong nucleosomes (SNs) characterized by visibly periodical DNA sequences have been found to concentrate in centromeres of *Arabidopsis thaliana* and in transient meiotic centromeres of *Caenorhabditis elegans*. To find out whether such affiliation of SNs to centromeres is a more general phenomenon, we studied SNs of the *Mus musculus.* The publicly available genome sequences of mouse, as well as of practically all other eukaryotes do not include the centromere regions which are difficult to assemble because of a large amount of repeat sequences in the centromeres and pericentromeric regions. We recovered those missing sequences using the data from MNase-seq experiments in mouse embryonic stem cells, where the sequence of DNA inside nucleosomes, including missing regions, was determined by 100-bp paired-end sequencing. Those nucleosome sequences, which are not matching to the published genome sequence, would largely belong to the centromeres. By evaluating SN densities in centromeres and in non-centromeric regions, we conclude that mouse SNs concentrate in the centromeres of telocentric mouse chromosomes, with ~3.9 times excess compared to their density in the rest of the genome. The remaining non-centromeric SNs are harbored mainly by introns and intergenic regions, by retro-transposons, in particular. The centromeric involvement of the SNs opens new horizons for the chromosome and centromere structure studies.

**Keywords:** strong nucleosome; chromatin; centromere; retro-tranposon; mouse

## 1. Introduction

The discovery of strong nucleosomes (SNs) (Salih, Tripathi, & Trifonov, 2013) has opened new vistas in the chromatin research field and in cytogenetics. The correlation between SNs and centromeres, which has been demonstrated recently for *Arabidopsis thaliana* (Salih & Trifonov, 2013) on the centromere sequences, and *Caenorhabditis elegans* (Salih & Trifonov, 2014), on transient centromeres which appear in spermatocytes during meiosis (Albertson & Thomson, 1993), provides an important clue to the functionality of nucleosomes in general and of SNs in particular.

To increase the spectrum of species, we turned to the mouse genome although almost none of sequenced chromosomes of mouse contains the centromere sequences, as well as most of the sequenced genomes of multicellular eukaryotes, due to technical difficulties in assembling highly repeating DNA segments comprising the centromere regions. In the mouse genome, chromosome Y is the only one which is almost completely sequenced (including its centromere region). As anticipated, the SN distribution of this chromosome showed a clear peak at one end where the centromere of this telocentric chromosome is located. As to other chromosomes, we found the way around the issue of the missing centromere regions. The idea is to use the unassembled nucleosome reads from MNase-seq experiments in mouse embryonic stem cells (ESCs), where 100 bps of DNA wrapped around the histone octamer were sequenced from both ends of the nucleosome (Teif et al., 2012) for the estimation of SN density ratio in gap regions (mainly centromeres) and sequenced regions. The calculations show significantly higher concentration of SNs in centromeric regions over non-centromeric ones, similar to the cases of *A. thaliana* and *C. elegans*.

The study has an additional motivation in view of the early work (Widlund et al., 1997) where unusually stable nucleosomes have been experimentally discovered in the mouse chromatin (by DNA/octamer binding competition), with similar sequence properties to our SNs, and located within centromeres (by FISH).

Analysis of the sequence environment of SNs in mouse shows that SNs are predominantly harbored by intergenic sequences, introns, and retrotransposons (LINE, LTR). SNs are found to have no special affinity neither to heterochromatin nor to euchromatin regions of the genome. One interesting exception is a congestion of the SNs in the E heterochromatin band of X chromosome, which might play a role in the X inactivation.

*Corresponding author. Email: bsaleh@campus.haifa.ac.il

Sequence-directed mapping of the SNs along the chromosomes shows the same features as in *A. thaliana* and *C. elegans* – solitary SNs and columnar structures (Salih & Trifonov, 2013, 2014).

## 2. Results and discussion

### 2.1. *SNs of chromosome Y concentrate in the centromere region*

The mouse genome is almost completely sequenced (approximately, 97% of its full size, http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/mouse/data/). However, 3% of it is still not sequenced. The missing sequences are further referred as 'gaps.' The terminal non-sequenced regions (3Mbase each) are located at one end of each of the mouse telocentric chromosomes, except for chromosome Y, which is practically fully sequenced. In Figure 1, the map of SNs along the Y chromosome is shown, calculated using the universal RR/YY nucleosome positioning probe (Tripathi, Salih, & Trifonov, 2014). This procedure is equivalent (Salih & Trifonov, 2014) to the original 'magic distances' algorithm described in (Salih et al., 2013). The SNs of chromosome Y are scattered all along, but they are clearly concentrated at the centromere end (Figure 1).

### 2.2. *Estimating SN density in centromeres and non-centromere regions*

SN is defined as a DNA sequence of size 116 bp (115 dinucleotides) with significant match to the 10.4 base periodical $(RRRRRYYYYY)_{11}$ probe representing idealized (strongest) nucleosome DNA sequence (Tripathi et al., 2014). With the match higher than ~66 (of maximal 115) the sequences display a clearly visible 10–11 base periodicity (Salih & Trifonov, 2014), while, typically, the nucleosome DNA sequences reveal the (hidden) periodicity only after one or another kind of sequence analysis is applied. The calculation of SN densities in centromeric and in non-centromeric regions is straightforward – by scoring the sequence segments with the match above threshold. To overcome the problem of mouse centromere sequences missing in public databases, we used the data-set of nucleosomal DNA reads generated by MNase digestion followed by paired-end sequencing of 100 bp from each nucleosome end (Teif et al., 2012) (about 108 million sequences). These are nucleosomal DNA sequences of average size ~160 bases, uniformly collected from the mouse genome. From this data-set, we generated the database of the nucleosome DNA sequences, presumably, equally representing all parts of the genome, centromeres included (see Materials and Methods). By applying the universal nucleosome positioning RR/YY probe, we collected all SNs from the experimentally determined nucleosome sequences, ending with total 195 SNs (after filtering the duplicates). It is worth to note that the MNase digestion of chromatin has not been geared for genome assembly purposes, and is far from being complete. Correspondingly, it is not expected to generate sequences of all nucleosomes. The projection of the derived set of SNs on the published full genome sequence of mouse finds 175 SNs belonging to the sequenced regions, while remaining 20 SNs are not found there and, thus, belong to the non-sequenced gaps, largely centromeric parts of the genome (according to the genome sequence source http://www.ncbi.nlm.nih.gov/genome/52, centromeres occupy ~76% of the gap regions). From this value and from Table 1, it follows that the density of SNs in centromere regions is at least ~3.9 times (.252/.064) higher than in the non-gap regions. This is to compare with the only fully sequenced chromosome Y, where this ratio shows an impressive 10.5 ratio (5.7/.54) as it follows from Figure 1. Alignment of the centromere SNs of this chromosome did not reveal any homology, so that this high ratio is not a result of high copy number of some subset of SNs. Generally, however, the 3.9 value may well be an underestimation due to possible repetition of

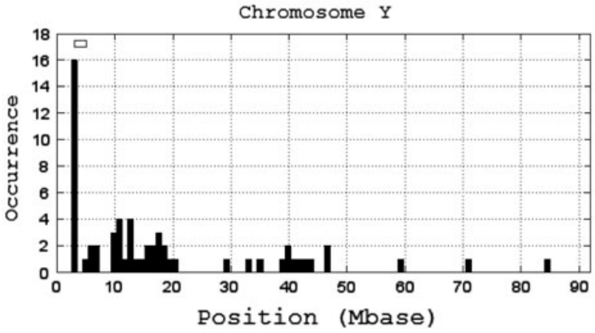

Figure 1. Distribution of SNs along the sequenced mouse chromosome Y, including the centromere region (leftmost). The white rectangle (3–5 Mbase, according to Pertile, Graham, Choo, & Kalitsis, 2009) indicates the approximate centromere position. The SN sequences of the first peak do not overlap with minor satellite repeats of the centromere (ibid). The bins of the histogram are of 1 Mbase width.

Table 1. SN density in gap regions and sequenced regions (calculated from pair-ends data-set).

| | Gap regions | Sequenced regions |
|---|---|---|
| Length (Mbase) | 79.3 | 2719.48 |
| Length (%) | 2.83% | 97.17% |
| Number of SNs | 20 | 175 |
| SN density* | 0.252/Mb | 0.064/Mb |

*SN densities are calculated on the assumption that density of ordinary and SNs together is about the same in both sequence types, i.e. ~1 nucleosome per 150–200 base pairs.

some of the centromere SNs which are removed by filtering (see Section 4.2).

Figure 2 shows SN distribution in all mouse chromosomes. The gap regions and telocentric centromeres, first three Mbases in each chromosome, are not indicated. The SNs are, essentially, scattered all along except for chromosome Y (as described above) and chromosome X which shows a conspicuous condensed region of SNs (coordinates 123,000,000 – 126,000,000) within XE heterochromatin region (see the X chromosome section below).

### *2.3. SN densities in other species*

In Table 2, actual ratios of SN densities (in centromeres vs. non-centromeric regions) in *A. thaliana* and *C. elegans* are presented. In the chromosomes of *A. thaliana*, the number of SNs in centromeres is 184 (the total centromere regions size is approximately 10 Mbase), while the number of SNs in non-centromeric regions of the same genome is 538, that is the SNs concentration (per unit length) in centromeres is 3.7 times higher than in non-centromeric regions. Same analysis for *C. elegans* genome yields the ratio 3.3.

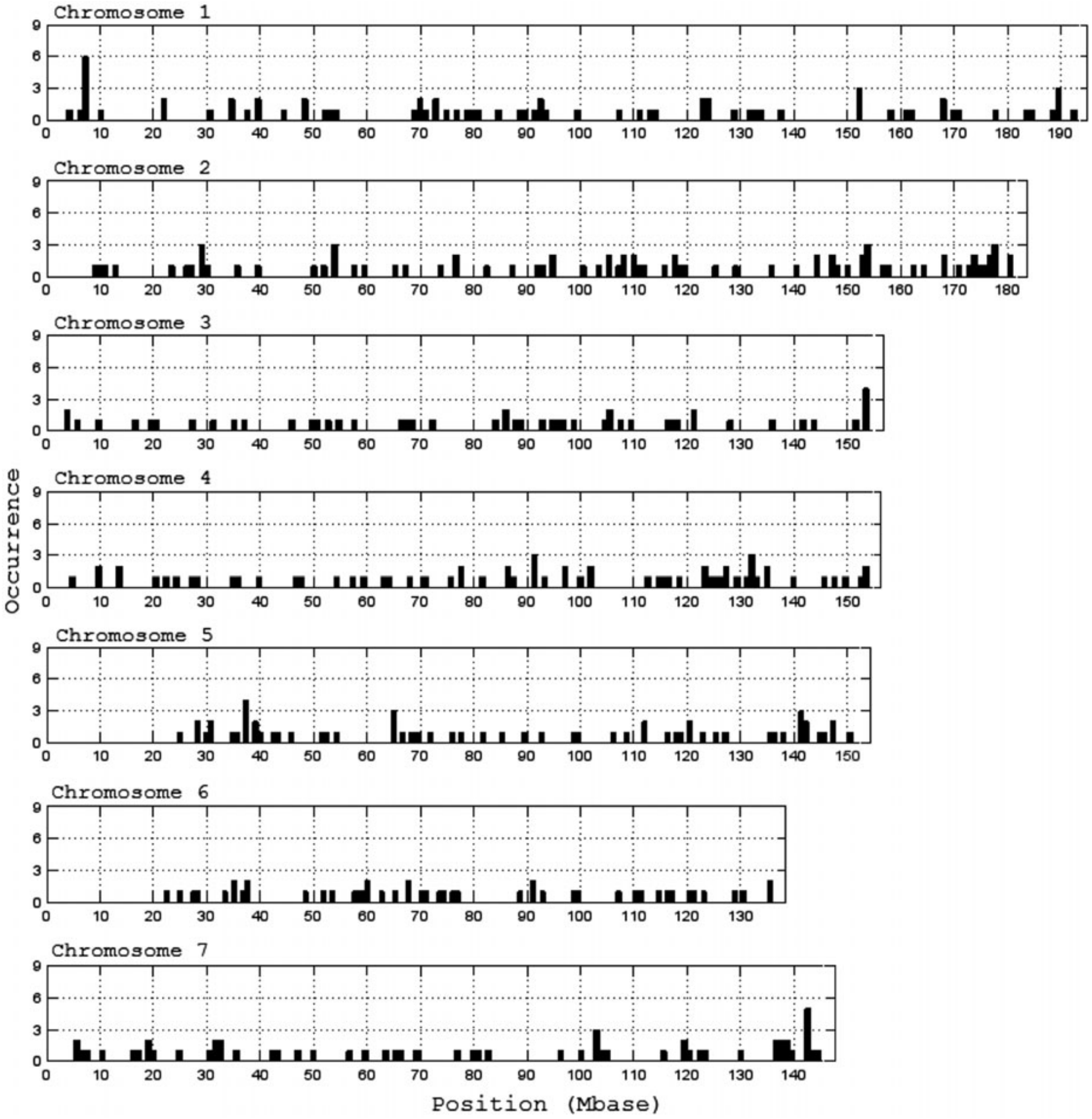

Figure 2. Strong nucleosome distribution for all mouse chromosomes. Note the differences in Y scales. Centromere regions are not highlighted.

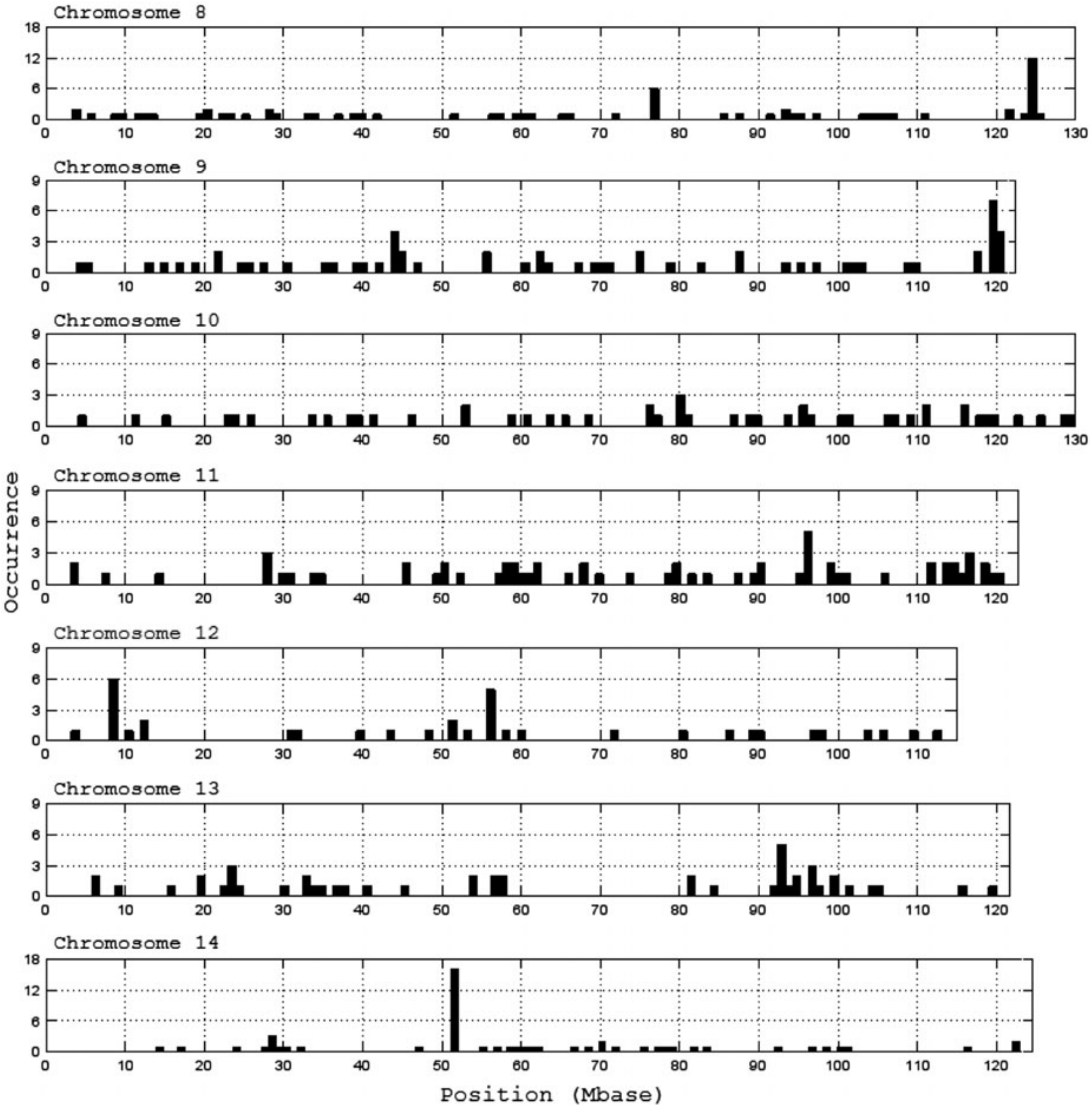

Figure 2. (*Continued*).

These ratios are comparable with the value estimated above for the mouse genome, ~3.9.

### 2.4. *No correlation between SNs and heterochromatin*

Heterochromatin is known to contain tightly packed DNA. It comes in different varieties between dense ‘constitutive’ heterochromatin and more diffuse ‘facultative’ heterochromatin. The constitutive heterochromatin is usually repetitive, forms centromeres, telomeres, and mostly does not contain genes. Facultative heterochromatin is less repetitive and is usually gene rich. Facultative heterochromatin can, under specific conditions, lose its condensed structure and become transcriptionally active (Oberdoerffer & Sinclair, 2007). A natural question would be: is there any correlation between tight SNs and dense heterochromatin? Table 3 lists SN densities in heterochromatin vs. euchromatin regions for chromosomes 1–7 separately, and for all chromosomes together (not including SNs from gaps). The numbers certify that SNs are evenly distributed between heterochromatin and euchromatin with only one remarkable exception – the chromosome X (see below). We have also checked that the typical heterochromatic mark H3K9me3 determined by ChIP-seq in mouse ESCs (Teif et al., 2014) is not enriched around SNs (data not shown).

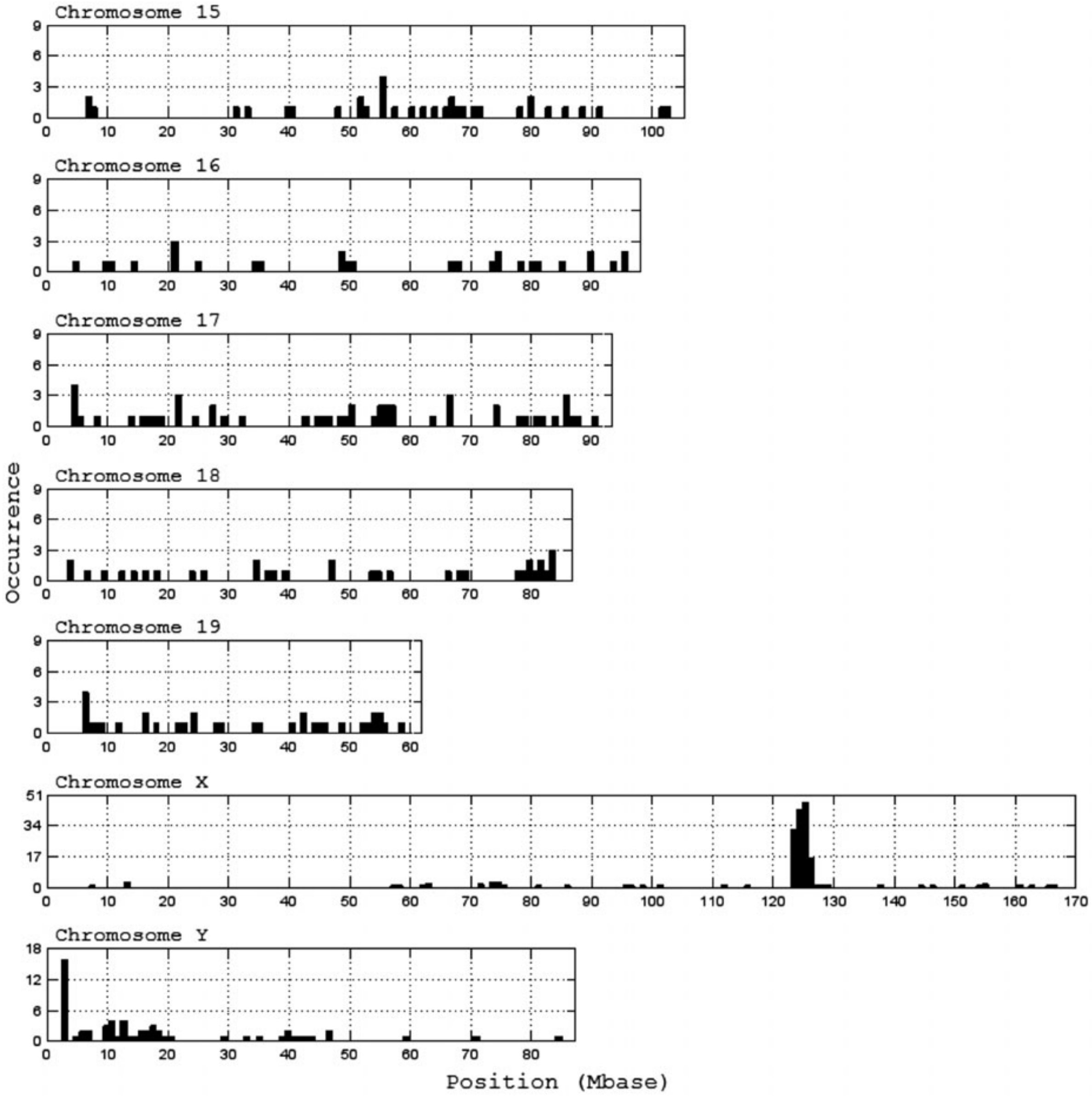

Figure 2. (*Continued*).

In Figure 3, a graphical illustration of SN distribution through the heterochromatin and euchromatin regions is shown for chromosomes 1–7. The results, thus, demonstrate that SNs do not have any special affinity to heterochromatin. However, they do have preference to centromeres and, consequentially, to the centromere heterochromatin.

### 2.5. *Congestion of SNs in the heterochromatin band E of X chromosome*

Contrary to other heterochromatin regions, the E band of chromosome X (http://www.ncbi.nlm.nih.gov/genome/52) contains conspicuously large number (131) of SNs within sequence coordinates 123 to 127 Mb (Figure 2). The SNs are distributed in 18 groups, often separated by 210–230 or 120–130 Kb from one another (Figure 4(a)). Each compact group (7 to 58 Kb) contains from 5 to 13 SNs (Figure 4(b)). Sixteen of SN sequences of the congestion region appear there more than once, from 2 to 11 times, in various groups. They are labeled in the Figure 4(b) by, respectively, different lowercase letters. This obvious structural regularity is further illustrated by apparent close similarity if not identity of some groups, containing SNs with the same sequences (Figure 4(b)) – groups G, J, M, and O (signature ghijklm) and groups H, I, K, N, P, and Q (signature hknol).

Table 2. SN densities in centromere/non-centromere regions of *A. thaliana* and *C. elegans*.

| | *A. thaliana* | *C. elegans* |
|---|---|---|
| SNs in centromere regions | 184 | 615 |
| SNs in non-centromere regions | 538 | 1381 |
| Centromeres sizes (Mbase) | ~10 | ~12 |
| Non-centromere size (Mbase) | 109.160 | 88.3 |
| SN density in CENs (per Mbase) | 18.4 | 51.3 |
| SN density in non-CENs (per Mbase) | 4.9 | 15.6 |
| SN density ratio | 3.7 | 3.3 |

Table 3. SN densities in heterochromatin/euchromatin regions of mouse chromosomes.

| | SN density* in heterochromatin regions (per Mbase) | SN density* in euchromatin regions (per Mbase) |
|---|---|---|
| Chrom. 1 | 0.318 | 0.433 |
| Chrom. 2 | 0.489 | 0.380 |
| Chrom. 3 | 0.260 | 0.399 |
| Chrom. 4 | 0.369 | 0.469 |
| Chrom. 5 | 0.274 | 0.542 |
| Chrom. 6 | 0.219 | 0.442 |
| Chrom. 7 | 0.397 | 0.418 |
| All (Chrom. 1–19, X, Y) | 0.459 | 0.445 |

*The densities do not include SNs from gap regions.

Although clusters of SNs of various sizes are found, typically, all along chromosomes, not just in centromeres (Salih & Trifonov, 2013, 2014), such large congestion as in XE heterochromatin is highly unusual. We have no explanation for this observation. All these congested SNs appear as solitary ones, neither in clusters, nor as part of columnar structures, as in *C. elegans*. Previously, it was reported that the E band of the inactive X chromosome is peculiar in several respects: this late-replicating band is devoid of Xist RNA and ubiquitinated H2A histones, in contrast to the yearly replicating D and F bands (Smith, Byron, Clemson, & Lawrence, 2004). It is also known that the E band has ~50% lower than the average X chromosome gene density, while 14% higher density of L1 repeats bands (Smith et al., 2004). All these features make E band special in terms of the X inactivation, which suggests that SN enrichment of E band might play a role in this process.

### 2.6. *Non-centromeric SNs are found primarily within introns and intergenic regions*

To find out which are particular sequence types where the SNs are located, we inspected the NCBI annotations of the mouse sequences surrounding the SNs. The data are presented in Table 4. Of 1238, SNs 805 are found within intergenic sequences, and 412 within introns, often within intronic and intergenic retrotransposons (270 cases). These are LINEs (mainly L1 type) and LTR transposons of subtypes ERVK, ERV1, and ERVL-MaLR. It, thus, appears that the SNs are located almost exclusively in non-coding regions. Of the 1238 cases scrutinized, only 21 SNs are found within exons, of which 11 – in protein-coding exons and 10 – within non-coding exons. We also found that SNs, according to annotations, do not belong to any satellite.

### 2.7. *SNs residing in exon (coding) sequences*

Eleven solitary SNs are found within exons of genes *Dst* and *Cenpf* (chr. 1), *Defb26* (chr. 2), *Iqgap3* (chr. 3), *Mllt3* (chr. 4), *Ccdc70* (chr. 8), *Homer1* (chr. 13), *Lrfn2* (chr. 17), and *Crem* (chr. 18). The SNs which would contain short exon sequences are not found. In chromosome 11, the 3rd exon (946 bases, positions 96,099,457 to 96,100,825) of gene *Calcoco2* encodes a columnar structure of size sufficient to accommodate 3 to 4 SNs (333 bases between last and first peaks corresponding to potential nucleosome centers on the map). The gene *Calcoco2* encodes the calcium binding and coiled-coil domain-2 protein. The coding sequence involved in the column is built of imperfect tandem repeats with consensus AAGGCCTCCTGGGAGGAAGAG (Crick strand) encoding amino acid repeat KASWEEE. The sequence of SN within gene Ccdc70 contains very similar repeat AAAACTTTCTGGGAAGAAGAG (Watson strand) encoding amino acid repeat KTFWEEE. SN of yet another exon, in gene of special interest, for *Cenpf* (centromere protein) has unrelated repeated sequence AGAAGTTCTGAGGATAATCAG (Crick strand) corresponding to consensus amino acid repeat RSSEDNQ.

### 2.8. *Clusters of SNs*

The tight clusters of SNs are observed in mouse as well as in *A. thaliana* (Salih & Trifonov, 2013) and *C. elegans* (Salih & Trifonov, 2014). This is seen in Table 5, where the occurrences of clusters of various sizes in the whole genome are presented. The cluster is understood as a group of 115 dinucleotide long (116 bases) SN DNA sequence fragments corresponding to DNA of elementary chromatin units (Trifonov, 2011) – separated one from another by not more than one unit (center-to-center distance 228). Majority of SNs appear as single isolated strongly periodical sequence segments accommodating only one (strong) nucleosome each. However, more than 6% of the SNs belong to clusters of two or more, up to six elementary chromatin units each (see Table 5). (Note that the statistics does not include recovered SNs of centromeres).

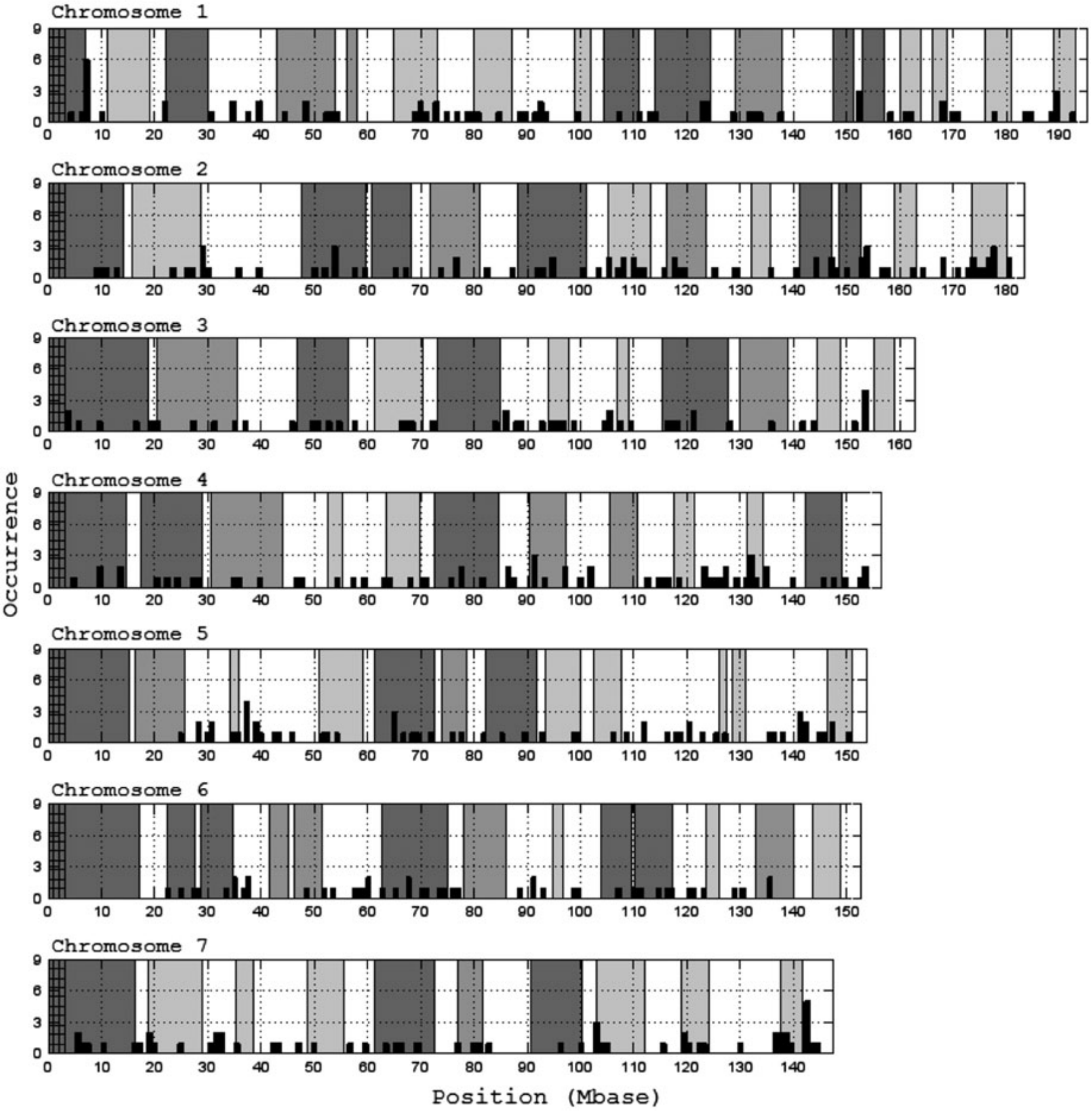

Figure 3. SN Distribution of SNs in heterochromatin (with three intensity levels of gray) and euchromatin regions of chromosomes 1–7. Gap (centromere) regions at the beginning of each chromosome, 3 Mb each, are checkered. The sequence coordinates of the heterochromatin regions are read from the map viewers in http://www.ncbi.nlm.nih.gov/genome/52.

Within the clusters, the SNs appear at short distances from one another, often following one right after another, in the same 10.4 base repeat phase, as it was also observed in *A. thaliana* (Salih & Trifonov, 2013) and *C. elegans* (Salih & Trifonov, 2014). In Figure 5(a), we see an example of nucleosome mapping corresponding to a characteristic solitary SN. The Figure 5(b), (c), and (e) are examples of SN clusters forming columnar structures (in-phase nucleosomes) accommodating two, three, and six SNs, respectively. While Figure 5(d) shows a cluster of four solitary SNs. Figure 6 provides an example of exceptionally strong nucleosome DNA sequence, corresponding to the nucleosome strength 96 (match to RR/YY probe), of maximal possible match 115. Note that in the examples of Figure 5, the amplitudes do not exceed ~80.

### 2.9. *SNs in insulatory chromatin regions*

Our analysis has revealed that at least 39 SNs are located within 500 bp from the sites bound by the insulatory protein CTCF in ESCs. Furthermore, at least 291 SNs (24% of all non-centromeric SNs) are located within 10,000 bp of CTCF sites bound in ESCs. CTCF demarcates active

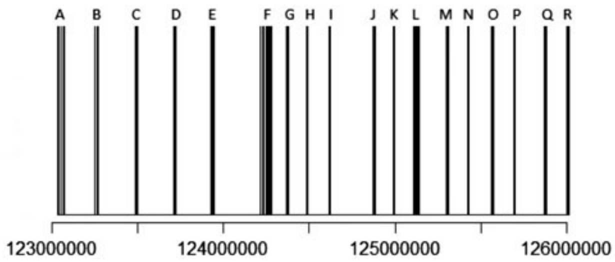

Figure 4(a). Distribution of the SNs in the SN congestion region of chromosome X. Eighteen SN groups containing 5–13 SNs each are labeled from A to R. Individual SNs (thin vertical bars) are seen in A, B, and F, and are not resolved in other groups, fusing in the thicker bars.

Table 4. Sequences containing SNs (1238 with strength above 65).

| Sequence type | Occurrence |
|---|---|
| *Intergenic* | *805* |
| LINE (96% L1, 4% L2) | 105 |
| LTR (48% ERVK, 32% ERVL-MaLR, 19% ERV1) | 83 |
| SINE (56% B2, 25% Alu, 18% B4) | 16 |
| *Intron* | *412* |
| LINE (90% L1, 3% L2) | 40 |
| LTR (50% ERVL-MaLR, 39% ERVK, 11% ERV1) | 18 |
| SINE (75% B2, 12% B4, 12% Alu) | 8 |
| *Exon* | *21* |
| LINE (L1) | 1 |
| LTR | 0 |
| SINE | 0 |

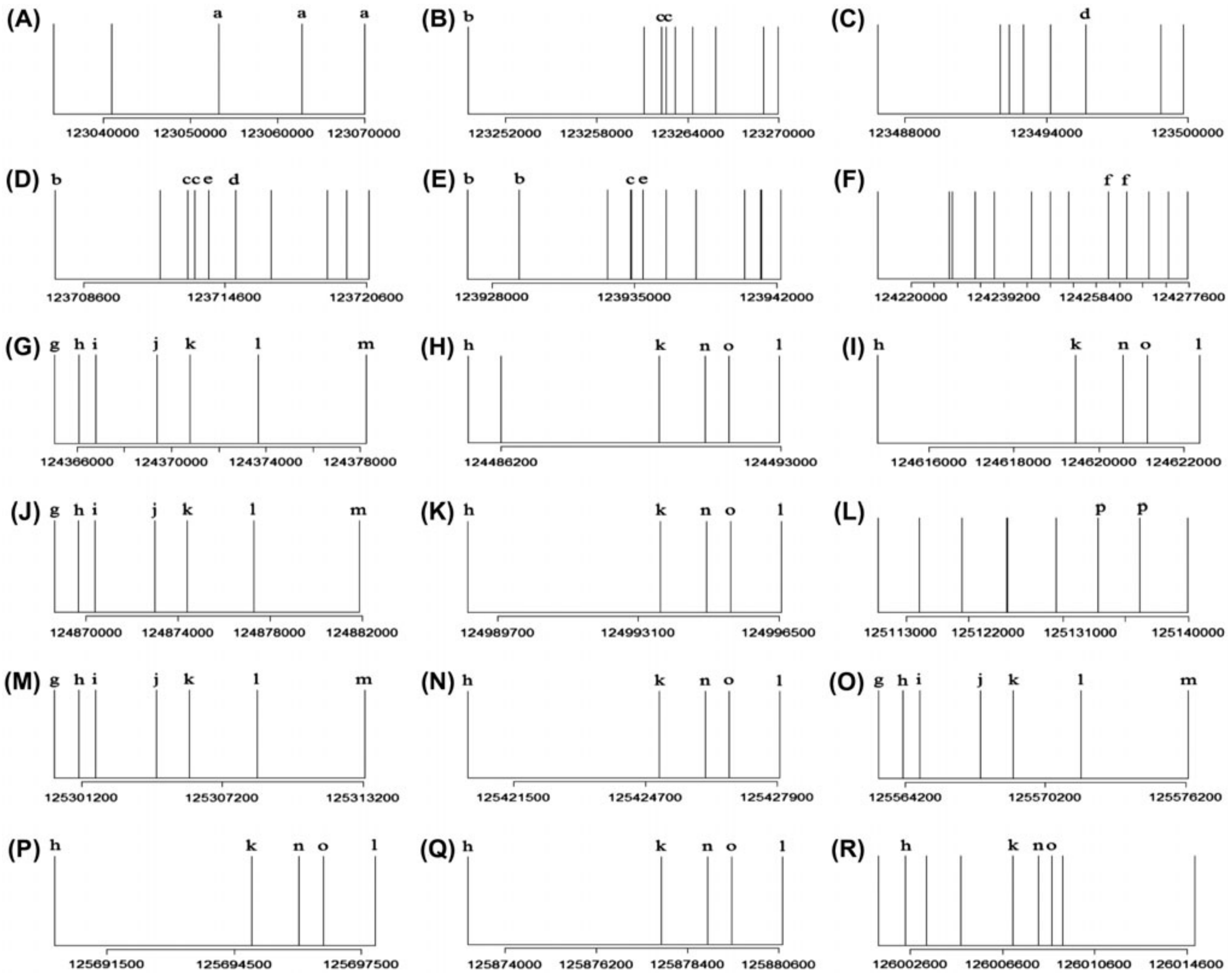

Figure 4(b). Individual SN groups of the SN congestion of chromosome X. Identical or nearly identical SN sequences in locations marked by vertical bars are labeled by lowercase letters. Note identical signatures for groups G, J, M, and O, and for groups H, I, K, N, P, and Q.

Table 5. Occurrence of isolated and clustered SNs in mouse chromosomes.

| Cluster size | Number of clusters |
|---|---|
| 1 | 1153 |
| 2 | 26 |
| 3 | 6 |
| 4 | 1 |
| 5 | 1 |
| 6 | 1 |

Note: The clusters are defined as those with distances <115 bases between the SNs of the clusters. Not including clusters from gap regions.

and inactive chromatin regions and plays a structural role by maintaining loops between distant chromatin regions. The positions of the boundaries set by CTCF change during the cell development. One aspect of this chromatin change by differential CTCF binding is through the regulation by DNA methylation and nucleosomes (Teif et al., 2014). CTCF sites are strongly enriched with CpGs (which can be either methylated or not, depending on the cell state). Interestingly, however, SNs located near CTCF are significantly depleted of CpGs (Figure 7). Importantly, SN arrangement near CTCF might have implications for the overall nucleosome arrangement in the insulatory regions (Beshnova, Cherstvy, Vainshtein, & Teif, 2014).

### *2.10. Comparison with experimental data on high affinity nucleosomes*

In the work of Widlund et al. (1997), the high affinity (HA) nucleosomes are defined as those, which are formed in experiments with competition between various

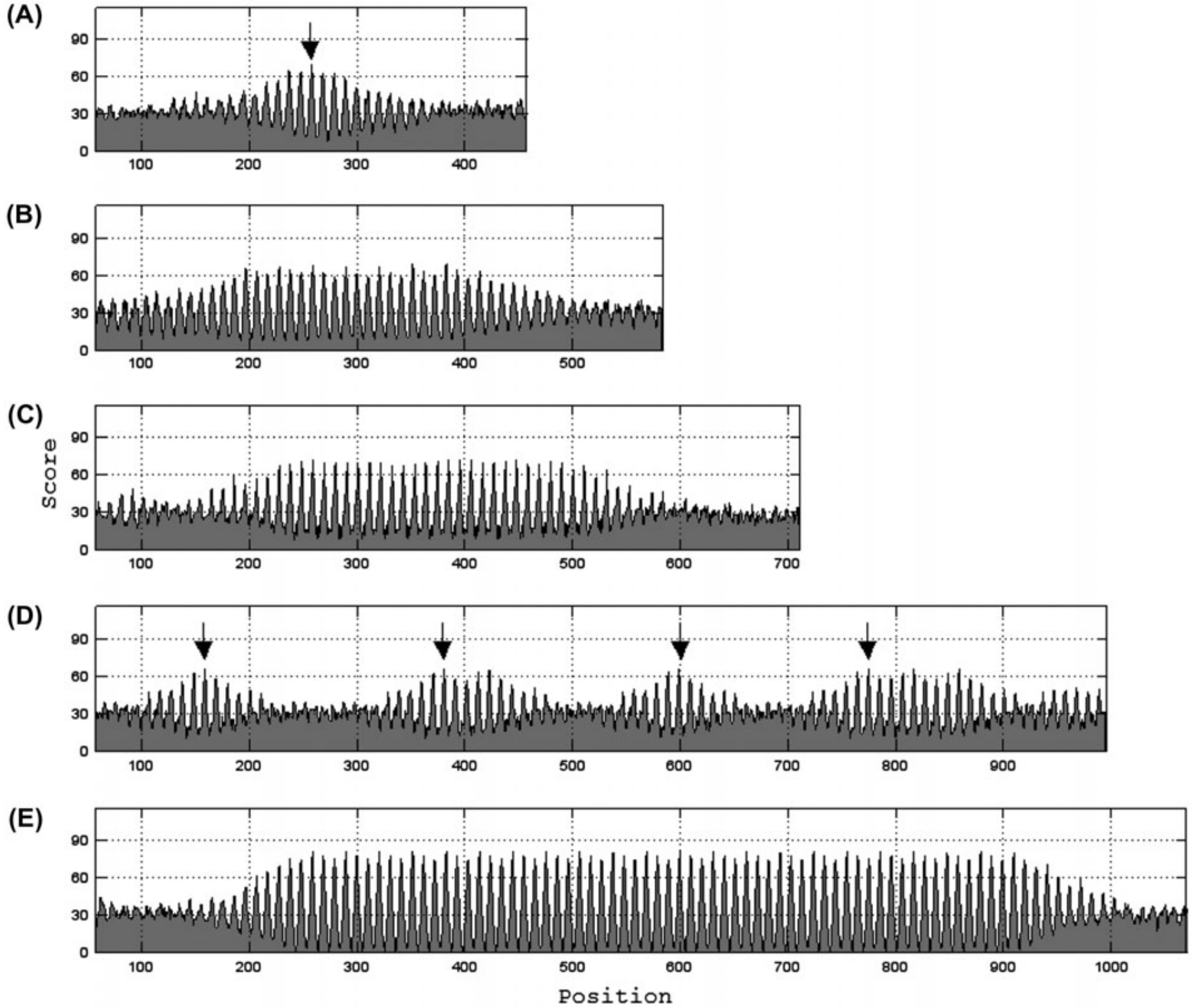

Figure 5. Examples of SN maps of mouse genome calculated with $(R_5Y_5)_{11}$ probe (Tripathi et al., 2014). (a) Solitary SN from chr1, centered at 74,905,011. (b), (c), and (e) Examples of columnar structures potentially accommodating 2, 3, and 6 SNs, respectively. Approximate starting coordinates of the columns: 81,431,793 (B, chr13), 141,210,334 (C, chr5), and 77,221,117 (E, chr8). (d) A cluster of four SNs from chr8, centered at 125,021,424, 125,021,646, 125,021,864, and 125,022,040.

C**AGGGAA**CCTCT**GGGGA**CCTC**AGGGGA**CCTCT**GGAGGA**CCTC**AGGGAA**CCTC
T**GGGGA**CCTC**AGGGGA**CCTCC**AGGGAG**CCTCC**AGAAAAA**TTT**AGGGGA**CCTC
C**AGAGA**TCTC**AG**

Figure 6. Sequence of the strong nucleosome with the highest for mouse genome score 96 detected within an intron in chromosome 5 at starting position 120,478,305. The sequence line size, for the purpose of illustration, is chosen equal 52(10.4 × 5) bases. Note the periodically appearing runs of purines (bold) alternating with pyrimidine runs.

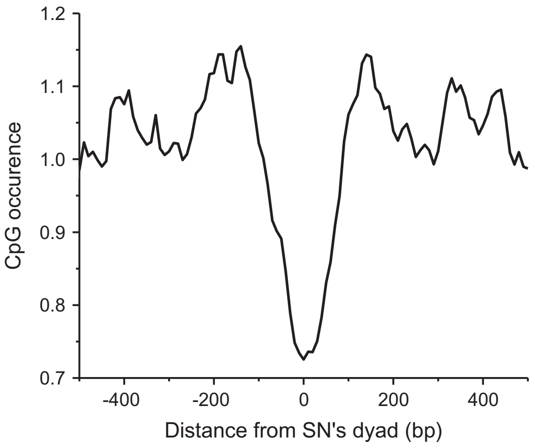

Figure 7. CpG profile averaged over all SNs in the annotated mouse genome showing the CpG depletion centered at the SN.

nucleosome DNA sequences for binding with histone octamers. Since the sequence definition of SNs is different (visible sequence periodicity, match to standard RR/YY nucleosome probe more than 56.5%, i.e. 66 or more 115 dinucleotide sequence probe), the sequences of the selected clones in (Widlund et al., 1997) should be compared to the standard SN probe. Such a comparison (not shown) reveals that of 12 sequences presented in (ibid), only one (TATA tetrads) satisfies the SN sequence definition. It consists of repeating 10-mers with consensus AAACGTCTAT (RRRyRYYYaY) which is 60% dinucleotide match to the (RRRRRYYYYY) standard. Other clones, strictly speaking, are not SNs, but their sequence strengths are higher than those for bulk nucleosome average. The experimental HA nucleosomes, thus, are weaker than SNs, which, perhaps, is explained by the low concentration of SNs in the competition experiments. From the data presented in previous sections, it follows that the proportion of SNs in the mouse genome is ~1 per 10,000 ordinary nucleosomes. Naturally, the clones in (Widlund et al., 1997) should be dominated by the ordinary or moderately SNs even after selection by competition.

The fact that the photolabeled clones are found to have collectively strong attraction to centromeres (ibid) suggests that not only exceptionally strong SNs but the stronger than average nucleosomes are associated with centromeres as well.

## 3. Conclusions

The fact that both plant centromere (*A. thaliana*) and transient meiotic nematode centromere (*C. elegans*) share the property of harboring SNs now seems to be true also for the telocentric chromosomes of mouse. This is a further confirmation that SNs are important structural elements of centromeres. Occurrence of SNs in other parts of the chromosomes as well suggests that they may play a similar role(s). One likely involvement is securing exact structural match during homologous pairing of chromatids, probably, being an integral part of the synaptonemal complexes. The match could be a specific interaction, either direct or via intermediates, between homologous SNs of the contacting chromatids. Figure 4(a) suggests a 'barcode' for such interaction (Ishiguro, Jihye, Fujiyama-Nakamura, Kato, & Watanabe, 2011), not unlike precise match of the barcode (Silver & Woodland, 1952) on a purchased article in supermarket to the barcode stored in the checking device.

Of course, these observations should be eventually extended to other species as well. However, even the limited data obtained already warrant further studies on the structure of the runs of SNs and on details of their distributions along chromosomes. The high-resolution computational sequence-directed tools for the nucleosomes' characterization, as in this work, open a whole new playground for the studies linking classical cytogenetics with modern genomics. The immediate experimental approaches are suggested as well, such as extraction and characterization of the tight SN aggregates (columns), and their possible crystallization. The columnar structures of the SNs, as they appear in the opening papers of a series on the subject (Salih & Trifonov, 2013, 2014; Salih et al., 2013; this work), seem to represent first well-defined natural elements of higher order structure of chromatin – perhaps, a first step towards its long-awaited high-resolution characterization.

The studies on the structure and function of centromeres, and on the role of SNs, in particular, are important for cytogenetics in general and for applications, especially in the field of artificial therapeutic chromosome design (Macnab & Whitehouse, 2009). SNs can be a part of solution of the CEN-DNA paradox, i.e. lack of sequence conservation in the highly conserved chromosome segregation structures, centromeres (Henikoff, Ahmad, & Malik, 2001). SNs may or may not be a universal signature of the centromeres, obligatory or dispensable, like the alpha satellites in human centromeres vs. nonalphoic neocentromeres (Choo, 1997). It is believed that the inheritance mechanism for centromeres involves chromatin (Henikoff et al., 2001). Centromeric nucleosomes have peculiar properties stemming in part from their specific histone composition. For example, heated discussions in recent high-profile publications have addressed the question of whether centromeric nucleosome contains eight or four histones (Codomo, Furuyama, & Henikoff, 2014; Miell, Straight, & Allshire, 2014). In addition, several hundreds of centromeric nucleosomes contain CENP-A histone variant (Burrack & Berman, 2012). Do centromeric SNs belong to CENP-A nucleosomes? This question remains to be addressed in the future, as well as many other interesting questions related to the role of SNs.

SNs, with their exceptional properties and affinity to centromeres, seem to have a significant role in the function of centromeres. The discovery of the SNs opens new prospects in both computational and experimental studies of chromatin, of chromosome structure, and of transposable elements.

## 4. Materials and methods

### 4.1. *DNA sequences*

Throughout this study, we used the mm10 genome assembly of *Mus musculus*. The DNA sequences of chromosomes 1–19, X, Y were downloaded from http://www.ncbi.nlm.nih.gov/genome/52. Experimental nucleosome positions in ESCs (Teif et al., 2012) were downloaded from the SRA archive (SRR572706.SRA). Experimental CTCF positions in ESCs (Shen et al., 2012) were obtained from the GEO archive (GSM918743).

### 4.2. *Post-processing of the DNA reads generated by MNase digestion*

The MNase-seq nucleosome data-set (SRR572706.SRA) contains 199,337,332 pairs of DNA reads (100 bases each). By merging the ends (up to reverse complement and 0% letter mismatch), we obtain 108,847,403 valid DNA sequences of average length ~160 bp. Then, we apply the $(R_5Y_5)_{11}$ nucleosome probe to the sequences to pick up SNs (those with score above 65), ending with 714 SNs. Finally, we filter duplicates or overlapping SNs based on sequence similarity, ending with 195 SNs (two SNs are considered duplicates or overlapping if they have overlapping sub-sequences – up to 7% letter mismatch – of length at least 60 bp). It is important to note that the total number of the filtered pair-end nucleosomes in the resulting database, though using a whole genome reads, may be rather small, depending on the sequence similarity thresholds. The rigorous filtering used, however, is not discriminating against any class of the nucleosomes, so that the resulting 175 + 20 SNs should adequately reflect their occurrence in the sequenced and centromeric regions.

### 4.3. *$(R_5Y_5)_{11}$ nucleosome mapping probe*

For the mapping of the nucleosomes, we used the $(R_5Y_5)_{11}$ probe (see Tripathi et al., 2014), or its earlier version, with negligible influence on results.

### 4.4. *Determination of strong nucleosome's cut-off threshold*

Using random sequences, appropriately generated, one can evaluate the score cut-off threshold. In this context, the null hypothesis, $H_0$, would be that 'Random sequences of base composition similar to those of the DNA sequence in question do not contain SNs.' We use, therefore, the following algorithm: (1) generate many random sequences (say 100 sequences of one million bases each) according to some base composition distribution; (2) for each sequence, independently, find the highest scoring fragment (i.e. a 116-bp-long fragment with highest match to the $(R_5Y_5)_{11}$ mapping probe); and (3) choose the maximum score of the highest scoring fragments over all sequences to be the cut-off threshold.

The estimated threshold for *M. musculus* genome is 66 (>65) (with significance level 0.01). This threshold separates fairly well the sequences with visible sequence periodicity from ordinary nucleosome DNA sequences.

## Supplementary material

The list of recovered centromeric SNs. The supplementary material for this paper is available online at http://dx.doi.10.1080/07391102.2014.938700.

## Acknowledgments

V.T. acknowledges the support from the Heidelberg Center for Modeling and Simulation in the Biosciences (BIOMS) and a DKFZ Intramural Grant.

## References

Albertson, D. G., & Thomson, J. N. (1993). Segregation of holocentric chromosomes at meiosis in the nematode, *Caenorhabditis elegans*. *Chromosome Research, 1*, 15–26.

Beshnova, D. A., Cherstvy, A. G., Vainshtein, Y., & Teif, V. B. (2014). Regulation of the nucleosome repeat length in vivo by the DNA sequence, protein concentrations and long-range interactions. *PLoS Computational Biology, 10*, e1003698. doi:10.1371/journal.pcbi.1003698.

Burrack, L. S., & Berman, J. (2012). Flexibility of centromere and kinetochore structures. *Trends in Genetics, 28*, 204–212.

Choo, K. H. (1997). Centromere DNA dynamics: Latent centromeres and neocentromere formation. *The American Journal of Human Genetics, 61*, 1225–1233.

Codomo, C. A., Furuyama, T., & Henikoff, S. (2014). CENP-A octamers do not confer a reduction in nucleosome height by AFM. *Nature Structural Molecular Biology, 21*, 4–5.

Henikoff, S., Ahmad, K., & Malik, H. S. (2001). The centromere paradox: Stable inheritance with rapidly evolving DNA. *Science, 293*, 1098–1102.

Ishiguro, K.-I., Jihye, K., Fujiyama-Nakamura, S., Kato, S., & Watanabe, Y. (2011). A new meiosis-specific cohesin complex implicated in the cohesin code for homologous pairing. *EMBO Reports, 12*, 267–275.

Macnab, S., & Whitehouse, A. (2009). Progress and prospects: Human artificial chromosomes. *Gene Therapy, 16*, 1180–1188.

Miell, M. D., Straight, A. F., & Allshire, R. C. (2014). Reply to 'CENP-A octamers do not confer a reduction in nucleosome height by AFM'. *Nature Structural Molecular Biology, 21*, 5–8.

Oberdoerffer, P., & Sinclair, D. (2007). The role of nuclear architecture in genomic instability and ageing. *Nature Reviews Molecular Cell Biology, 8*, 692–702.

Pertile, M. D., Graham, A. N., Choo, K. H., & Kalitsis, P. (2009). Rapid evolution of mouse Y centromere repeat DNA belies recent sequence stability. *Genome Research, 19*, 2202–2213.

Salih, B., & Trifonov, E. N. (2013). Strong nucleosomes of *A. thaliana* concentrate in centromere regions. *Journal of Biomolecular Structure and Dynamicsonline*. Advance Online Publication. doi:10.1080/07391102.2013.860624.

Salih, B., & Trifonov, E. N. (2014). Strong nucleosomes reside in meiotic centromeres of *C. elegans*. *Journal of Biomolecular Structure and Dynamics*. Advance Online Publication. doi:10.1080/07391102.2013.879263.

Salih, B., Tripathi, V., & Trifonov, E. N. (2013). Visible periodicity of strong nucleosome DNA sequences. *Journal of Biomolecular Structure and Dynamics*. Advance Online Publication. doi:10.1080/07391102.2013.855143.

Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., …, Ren, B. (2012). A map of the cis-regulatory sequences in the mouse genome. *Nature, 488*, 116–120.

Silver, B., & Woodland, N. J. (1952). *US Patent No. 2,612,994*. Washington, DC: U.S. Patent and Trademark Office.

Smith, K. P., Byron, M., Clemson, C. M., & Lawrence, J. B. (2004). Ubiquitinated proteins including uH2A on the human and mouse inactive X chromosome: Enrichment in gene rich bands. *Chromosoma, 113*, 324–335.

Teif, V. B., Beshnova, D. A., Vainshtein, Y., Marth, C., Mallm, J.-P., Höfer, T., & Rippe, K. (2014). Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development. *Genome Research, 24*, 1285–1295.

Teif, V. B., Vainstein, E., Marth, K., Mallm, J.-P., Caudron-Herger, M., Höfer, T., & Rippe, K. (2012). Genome-wide nucleosome positioning during embryonic stem cell development. *Nature Structural Molecular Biology, 19*, 1185–1192.

Trifonov, E. N. (2011). Cracking the chromatin code: Precise rule of nucleosome positioning. *Physics of Life Reviews, 8*, 39–50.

Tripathi, V., Salih, B., & Trifonov, E. N. (2014). Universal full length nucleosome mapping sequence probe. *Journal of Biomolecular Structure and Dynamics*. Advance Online Publication. doi:10.1080/07391102.2014.891262.

Widlund, H. R., Cao, H., Simonsson, S., Magnusson, E., Simonsson, T., Nielsen, P. E., Kahn, J. D., Crothers, D. M., & Kubista, M. (1997). Identification and characterization of genomic nucleosome-positioning sequences. *Journal of Molecular Biology, 267*, 807–817.