Strong nucleosomes of mouse genome in recovered centromeric sequences

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

[a] Genome Diversity Center, Institute of Evolution, University of Haifa, Mount Carmel, Haifa 31905, Israel

[b] Department of Computer Science, University of Haifa, Mount Carmel, Haifa 31905, Israel

[c] German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

[*] Corresponding author. Tel.: + 972 54 2396075. E-mail address: bsaleh@campus.haifa.ac.il

Abstract

Recently discovered strong nucleosomes (SNs) characterized by visibly periodical DNA sequences have been found to concentrate in centromeres of *A. thaliana* and in transient meiotic centromeres of *C. 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. We recovered those missing sequences by using the data from MNase-seq experiments in mouse embryonic stem cells, where the sequence of DNA inside nucleosomes, including un-annotated 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 (Salih &

Trifonov, 2013; Salih & Trifonov, 2014) seems to be an important clue as to the functionality of nucleosomes in general and of SNs in particular.

In this work, we analyze SNs in mouse as it was done before with *A. thaliana* (Salih & Trifonov, 2013) and *C. elegans* (Salih & Trifonov, 2014). Unfortunately, most of the sequenced genomes of multicellular eukaryotes, as of today, lack the centromeric sequences 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 significant parts of 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 annotation. 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*.

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 E heterochromatin band of X chromosome.

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; Salih & Trifonov, 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 still not sequenced. The terminal non-sequenced regions (about 3Mbase each) further referred to as 'gaps' 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 by 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).

[Figure 1]

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

SN is defined as a DNA sequence of size 115 bp (114 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 DNA reads generated by MNase digestion (Teif *et al.*, 2012) (about 108 million sequences). These are nucleosomal DNA sequences of average size ~160 bases, uniformly collected from the whole mouse genome. From this data-set we generated the pair-ends database of the nucleosome DNA sequences, representing, presumably, the whole 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). The projection of this 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, largely centromeric parts of the genome (centromeres occupy ~ 80% of the gap regions), as summarized in Table 1. The density of SNs in gap regions is ~ 3.9 times (.252/.064) higher than in the non-gap regions.

[Table 1]

Figure 2 shows SN distribution in all mouse chromosomes. The small gap regions 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).

[Figure 2]

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

[Table 2]

2.4. No correlation between SNs and heterochromatin.

Heterochromatin is known to contain a 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 normally 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).

[Table 3]

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.

[Figure 3]

2.5. Congestion of SNs in heterochromatin region E of X chromosome

Contrary to other heterochromatin regions, the E-region of chromosome X 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 4a). Each compact group (7 to 58 Kb) contains from 5 to 13 SNs (Figure 4b). 16 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 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 4b) – groups G, J, M, O (signature ghijklm) and groups H, I, K, N, P, Q (signature hknol).

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*. The annotated NCBI database does not report any peculiar information regarding the sequence environment of these SNs. The function of this region is uncertain as well.

[Figure 4a, Figure 4b]

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.

[Table 4]

2.7. Strong nucleosomes 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 96099457 to 96100825) 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 114 dinucleotide long (115 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 2 or more, up to 6 elementary chromatin-units each (see Table 5). (Note that the statistics does not include recovered SNs of centromeres).

[Table 5]

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 5a we see an example of nucleosome mapping, corresponding to a characteristic solitary SN. The Figures 5b, 5c, and 5e are examples of SN clusters forming columnar structures (in-phase nucleosomes) accommodating 2, 3, and 6 SNs, respectively. While Figure 5d shows a non-columnar cluster of 4 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 114. Note that in the examples of Figure 5 the amplitudes do not exceed ~80.

[Figure 5]

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 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 *et al.*, 2014).

3. Conclusions

The fact that both plant centromere (*A. thaliana*) and transient meiotic nematode centromere (*C. elegans*) share the property of harboring SNs seems now to be also true 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 synapsis 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 4a suggests a 'bar-code' for such interaction.

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 *et al.*, 2013; Salih & Trifonov, 2013; Salih & Trifonov, 2014; 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 8 or 4 histones (Miell, Straight, & Allshire, 2014; Codomo, Furuyama, & Henikoff, 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 dataset (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 an 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 strong nucleosomes'. We use, therefore, the following algorithm: 1) Generate many random sequences (say 100 sequences of 1 million bases each) according to some base composition distribution, 2) For each sequence, independently, find the highest scoring fragment (i.e. a 115 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.

5. References

Beshnova, D. A., Cherstvy, A. G., & Teif V. B. (2014). Genetic and epigenetic determinants of the nucleosome repeat length. *PLoS Computational Biology* (under review).

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. *American Journal of Human Genetics, 61(6),* 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(5532),* 1098-1102.

Macnab, S., & Whitehouse, A. (2009). Progress and prospects: human artificial chromosomes. *Gene Therapy, 16(10),* 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(12),* 2202-2213.

Rosenfeld, J. A., Wang, Z., Schones, D., Zhao, K., Desalle, R., & Zhang, M. Q. (2009). Determination of enriched histone modifications in non-genic portions of the human genome. *BMC Genomics 10:143*. doi:10.1186/1471-2164-10-143

Salih, B., & Trifonov, E. N. (2013). Strong nucleosomes of *A. thaliana* concentrate in centromere regions. *Journal of Biomolecular Structure and Dynamics*. 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., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. V., & Ren, B. (2012). A map of the cis-regulatory sequences in the mouse genome. *Nature, 488*, 116-120.

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.

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

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

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

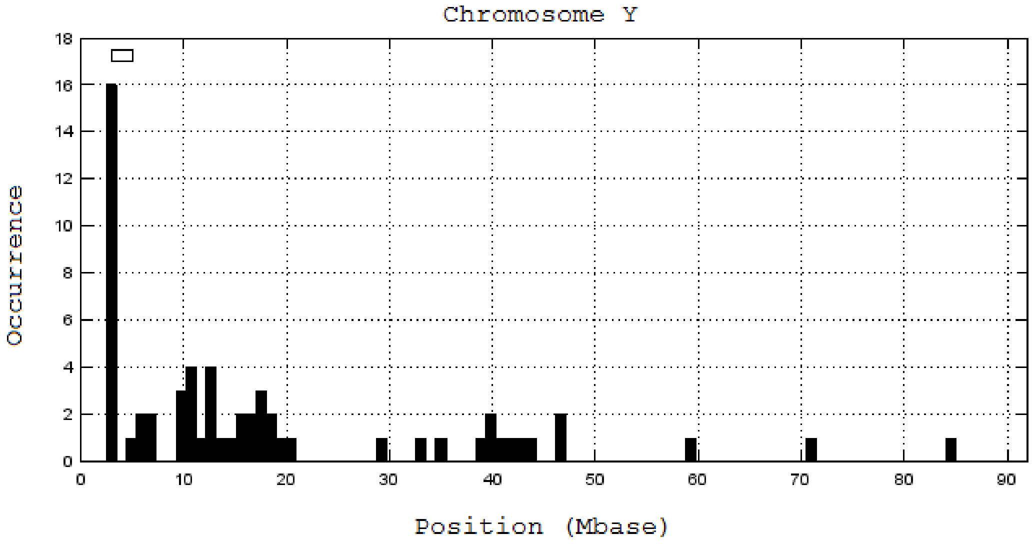
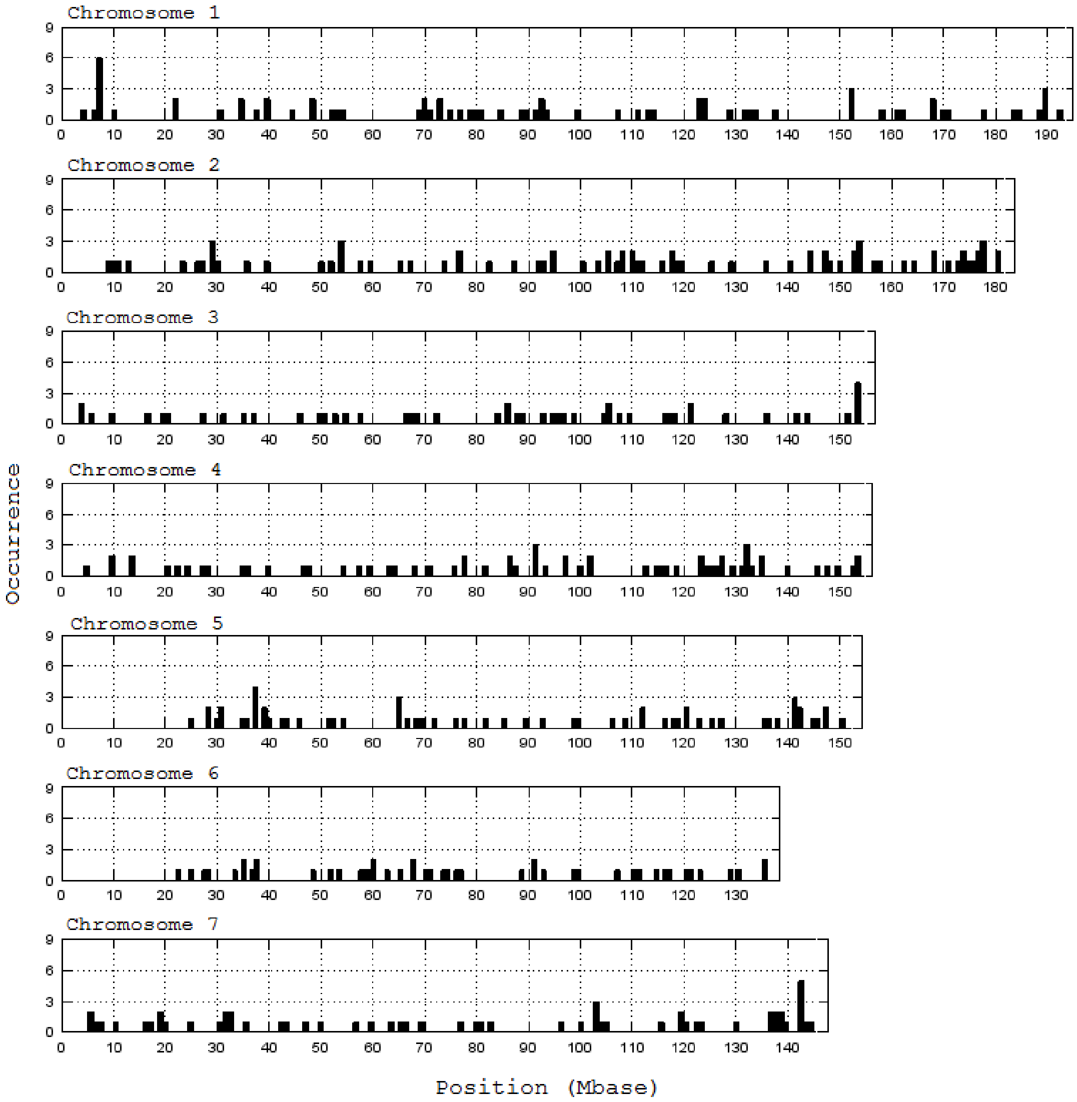
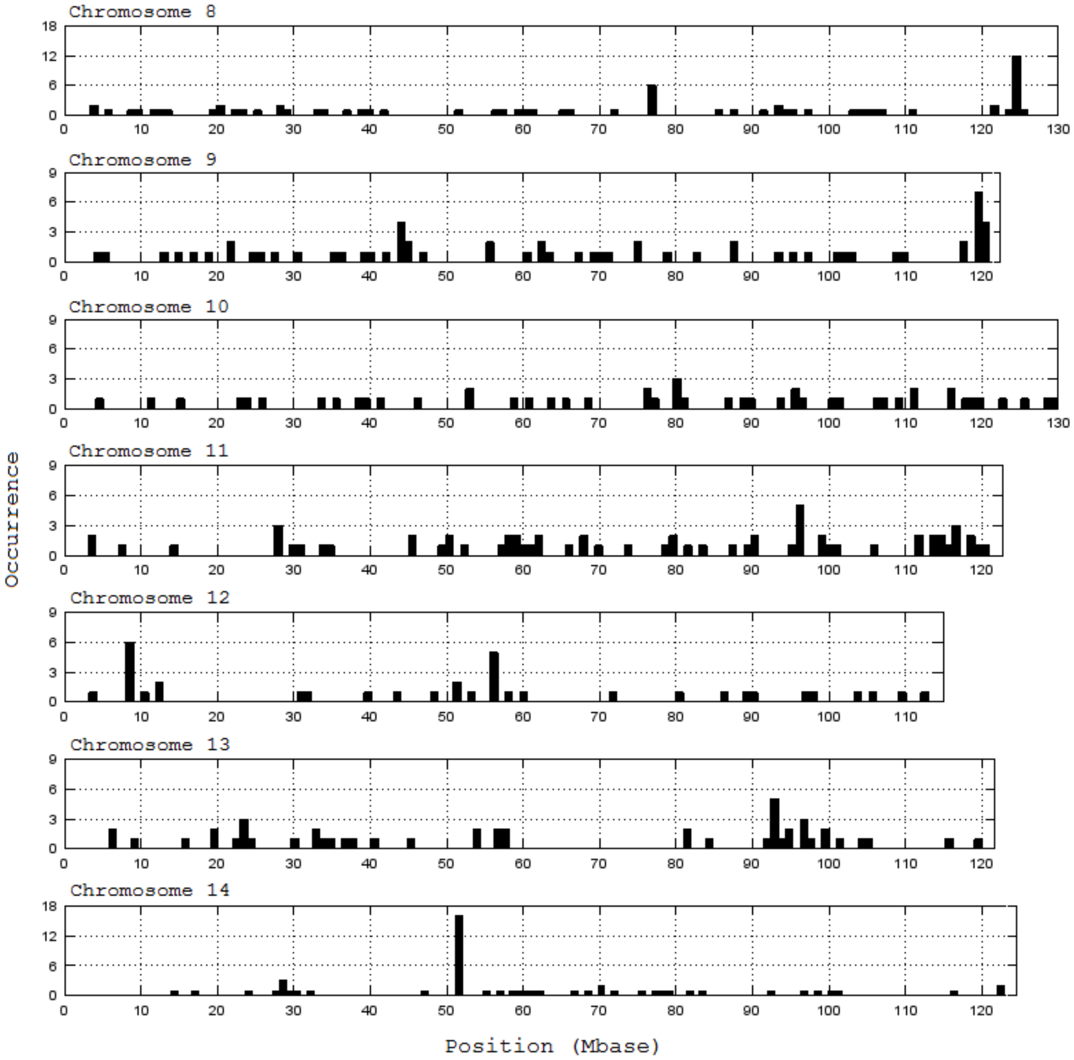


Figure 1. Distribution of strong nucleosomes along the sequenced mouse chromosome Y, including the centromere region (leftmost). The white rectangle (3-5 Mbase, according to Pertile *et al.*, 2009) indicates the approximate centromere position. The SN sequences of the first peak do not overlap with minor satellite repeats of the centromere (ibid). The bins of the histogram are of 1 Mbase width.


Chromosome 1
Chromosome 2
Chromosome 3
Chromosome 4
Chromosome 5
Chromosome 6
Chromosome 7
Occurrence
0
3
6
9
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
Position (Mbase)


Chromosome 8
Chromosome 9
Chromosome 10
Chromosome 11
Chromosome 12
Chromosome 13
Chromosome 14
Occurrence
Position (Mbase)
0
3
6
9
12
18
10
20
30
40
50
60
70
80
90
100
110
120
130

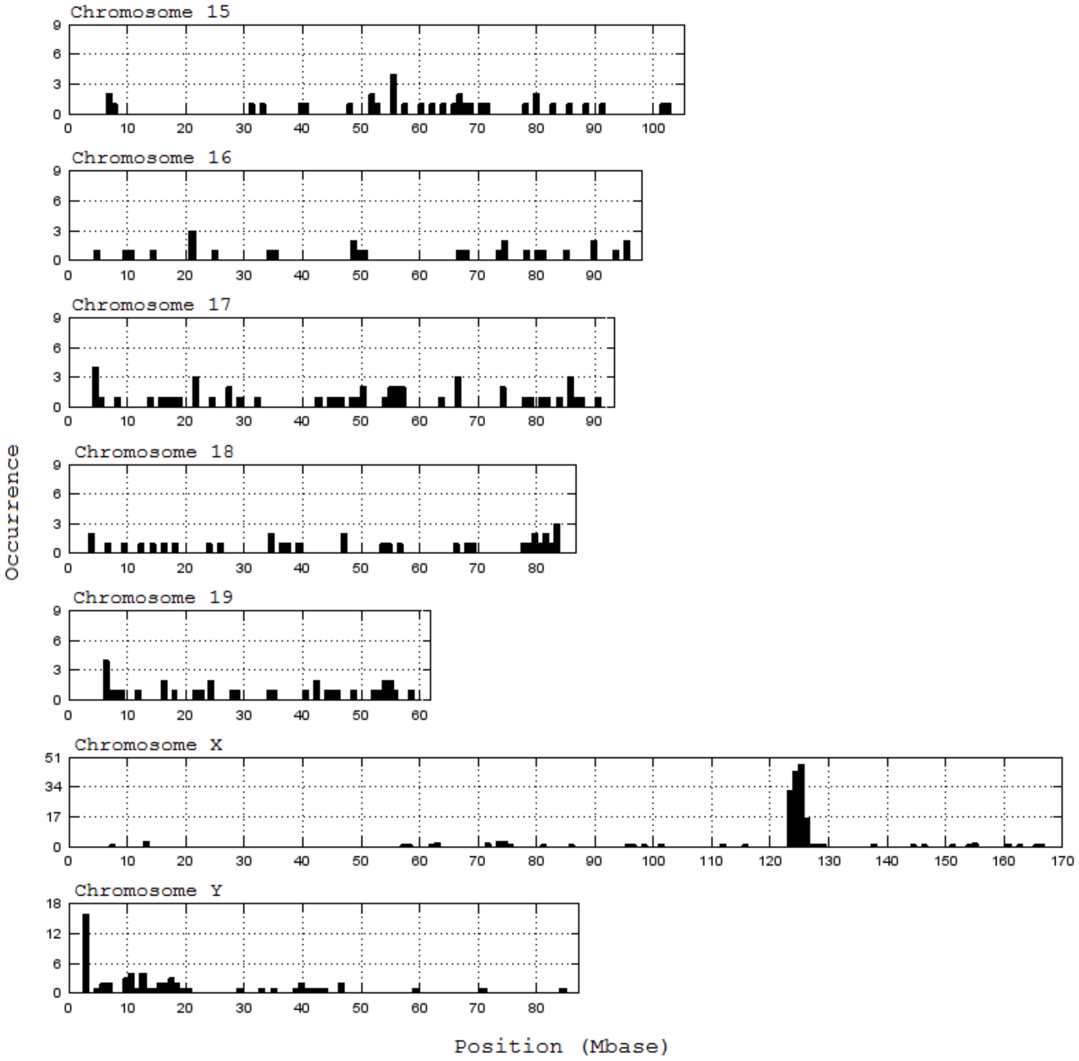


Figure 2. Strong nucleosome distribution for all mouse chromosomes. Note the differences in Y-scales.

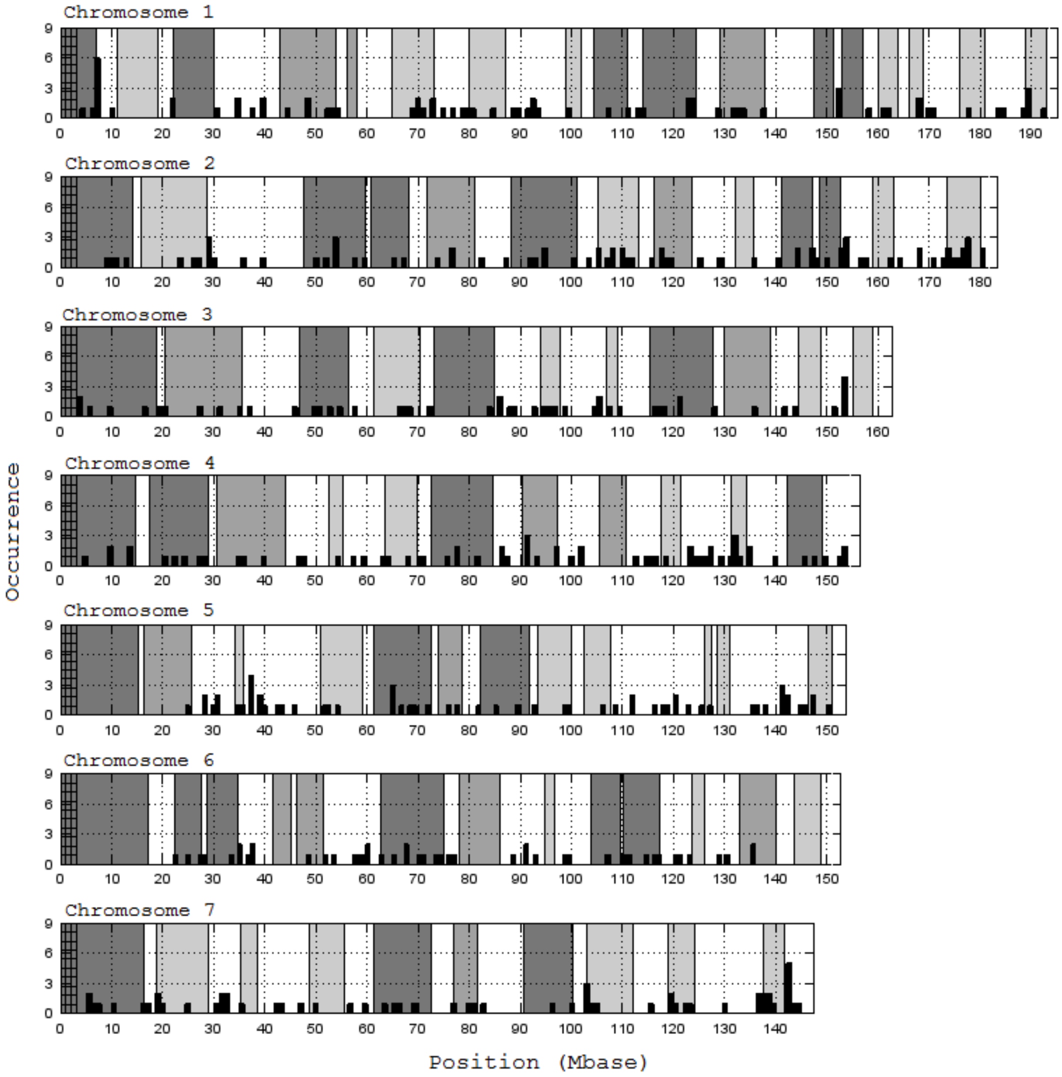


Figure 3. SN Distribution of strong nucleosomes in heterochromatin (with 3 intensity levels of gray) and euchromatin regions of chromosomes 1 to 7. Gap (centromere) regions at the beginning of each chromosome, 3Mb each, are checkered.

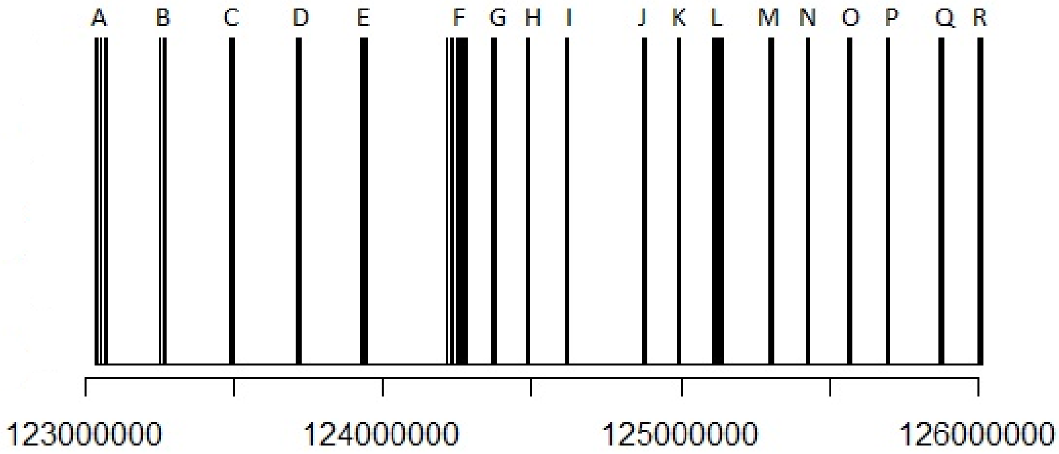


Figure 4a. Distribution of the SNs in the SN congestion region of chromosome X. 18 SN groups containing 5-13 SNs each are labeled from A to R. Individual SNs (thin vertical bars) are seen in A, B, and F, and are not resolved in other groups, fusing in the thicker bars.

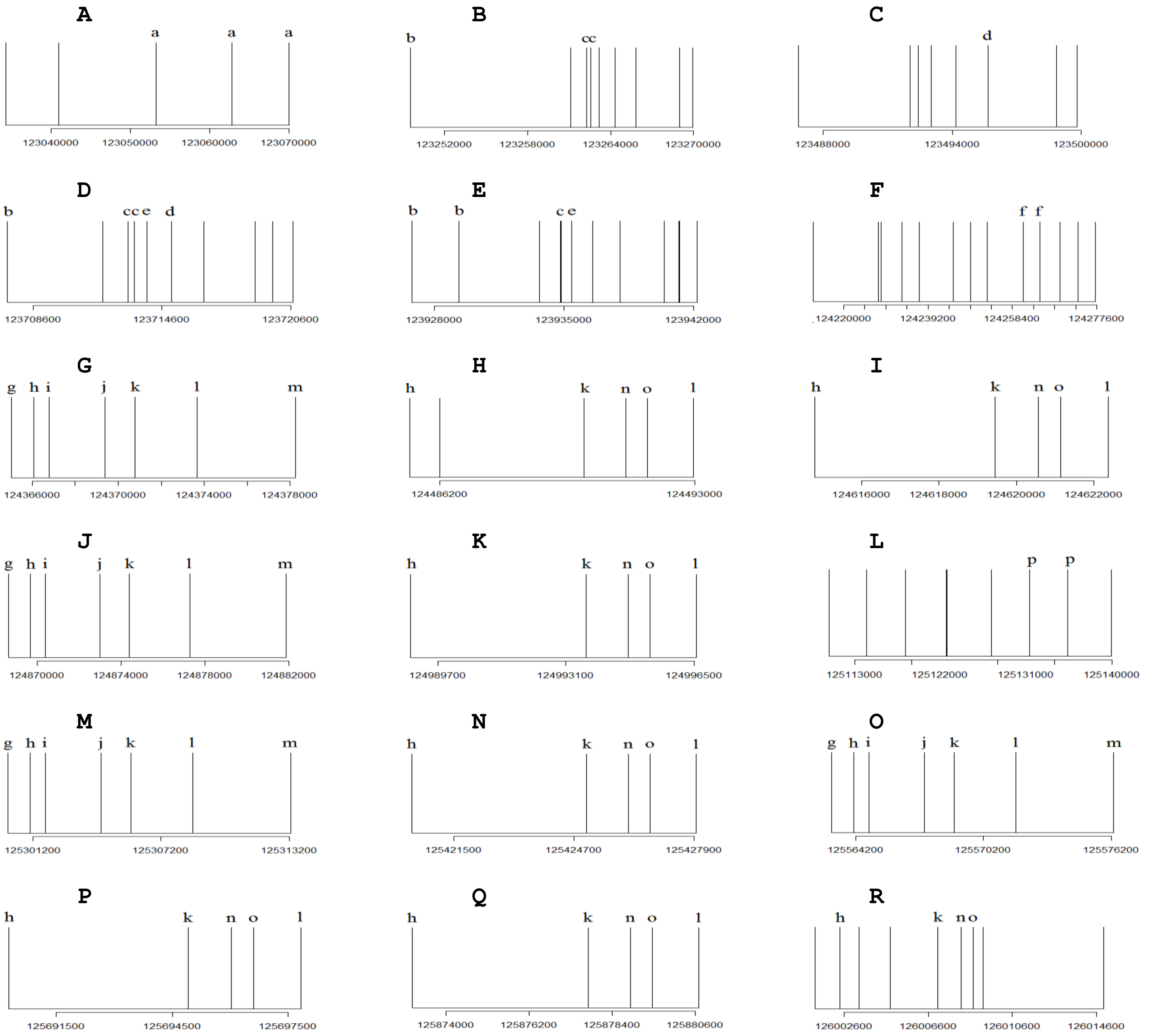


Figure 4b. Individual SN groups of the SN congestion of chromosome X. Identical or nearly identical SN sequences in locations marked by vertical bars are labeled by lowercase letters. Note identical signatures for groups G, J, M, and O, and for groups H, I, K, N, P, and Q.

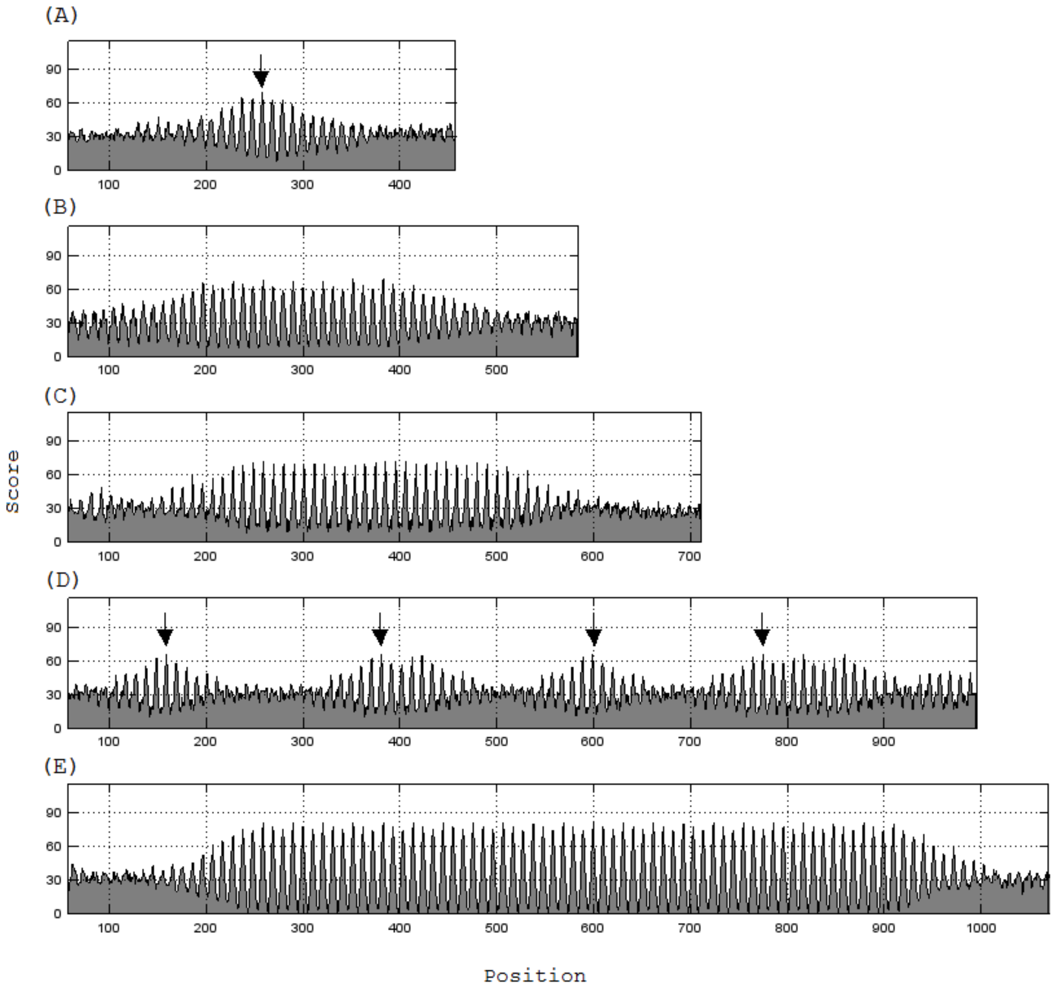


Figure 5. Examples of SN maps of mouse genome calculated with $(R_5Y_5)_{11}$ probe (Tripathi *et al.*, 2014). (A) Solitary SN from chr1, centered at 74905011. (B), (C), and (E) Examples of columnar structures potentially accommodating 2, 3, and 6 SNs, respectively. Approximate starting coordinates of the columns: 81431793 (B, chr13), 141210334 (C, chr5), and 77221117 (E, chr8). (D) A cluster of 4 SNs from chr8, centered at 125021424, 125021646, 125021864, and 125022040.

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(10.4x5) bases. Note the periodically appearing runs of purines (bold) alternating with pyrimidine runs.

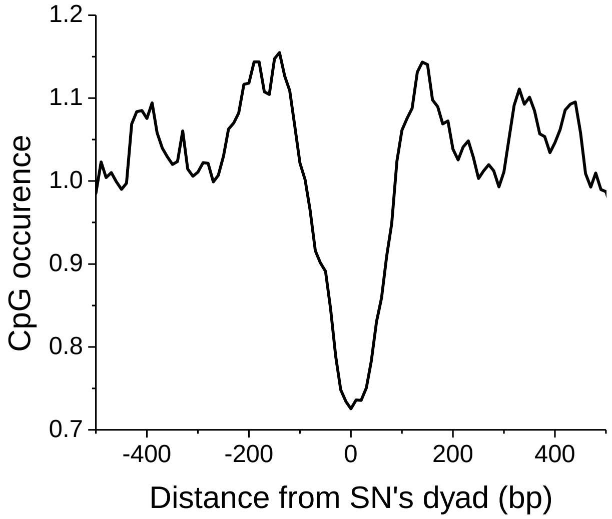


Figure 7. CpG profile averaged over all SNs in the annotated mouse genome showing the CpG depletion centered at the SN.

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

	Gap regions	Sequenced regions
Length (Mbase)	79.3	2,719.48
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 strong nucleosomes together is about the same in both sequence types, i.e., ~ 1 nucleosome per 150-200 base pairs.

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.

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

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

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