1172–1181 *Nucleic Acids Research, 2010, Vol. 38, No. 4* *Published online 4 December 2009*
doi:10.1093/nar/gkp1094

The most frequent short sequences in non-coding DNA

Juan A. Subirana[1,*] and Xavier Messeguer[2]

[1]Departament d'Enginyeria Química, Universitat Politècnica de Catalunya, Av. Diagonal 647, E-08028, Barcelona and [2]Departament de Llenguatges I Sistemes Informàtics, Universitat Politècnica de Catalunya, C. Jordi Girona 1-3, E-08034, Barcelona, Spain

Received September 16, 2009; Revised November 4, 2009; Accepted November 6, 2009

ABSTRACT

The purpose of this work is to determine the most frequent short sequences in non-coding DNA. They may play a role in maintaining the structure and function of eukaryotic chromosomes. We present a simple method for the detection and analysis of such sequences in several genomes, including *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Homo sapiens*. We also study two chromosomes of man and mouse with a length similar to the whole genomes of the other species. We provide a list of the most common sequences of 9–14 bases in each genome. As expected, they are present in human *Alu* sequences. Our programs may also give a graph and a list of their position in the genome. Detection of clusters is also possible. In most cases, these sequences contain few alternating regions. Their intrinsic structure and their influence on nucleosome formation are not known. In particular, we have found new features of short sequences in *C. elegans*, which are distributed in heterogeneous clusters. They appear as punctuation marks in the chromosomes. Such clusters are not found in either *A. thaliana* or *D. melanogaster*. We discuss the possibility that they play a role in centromere function and homolog recognition in meiosis.

INTRODUCTION

For a long time it has been known that genomes contain a large amount of non-coding DNA. This DNA was suggested to be 'junk' or 'parasite DNA' with no specific function (1,2), although recent work has shown that a large fraction of this DNA is transcribed (3). The sequences present in non-coding DNA are complex and probably include features which are essential for chromosome structure, such as chromosome condensation, axis formation, homologous chromosome pairing in meiosis, etc. Our working hypothesis is that such features will be found in short sequences of DNA with a similar composition. The programs we have developed are aimed to find the eventual presence of such sequences. It should be taken into account that a large fraction of non-coding DNA is very variable in sequence, even when closely related species are compared. It could be involved in important structural/functional roles, which do not require exact sequence conservation (4). The recent availability of whole-genome DNA sequences allows a study of the length, composition and distribution of such short abundant sequences in the genome, as a prerequisite to determine their eventual function. Our results will be placed in the context of other non-coding DNA sequences which are present in the genome, such as microsatellites, longer satellites, SINEs, LINEs, etc. Microsatellites are long tandem repeats of short sequences, usually 1–6 bases in length. We have recently analyzed their distribution in different genomes (5).

A vast amount of data is available on repeated sequences in the genome. The main categories of repeats in vertebrates were already analyzed by Smit (6) in 1999. References to the methods used by different authors and the results obtained can be found in RepBase (7). Most of the methods available are aimed to find comparatively long sequences, which are repeated only a few times in the genome. For the results presented in this paper, we required a simpler method, suitable to discover and analyze short sequences (9–14 bases), which may be found many times in each genome. We have also developed a method to detect clusters of such sequences. We have studied the whole genomes of *Arabidopsis thaliana*, *Caenorhabditis elegans, Drosophila melanogaster* and *Homo sapiens* and one chromosome of mouse and man. Selected references to previous related studies with these species are given throughout the text. As examples of

*To whom correspondence should be addressed. Tel: +34 934016688; Fax: +34 934010978; Email: juan.a.subirana@upc.edu

© The Author(s) 2009. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

the application of our methodology, we discuss some features of selected sequences in man and mouse. We also describe the frequent sequences found in *C. elegans*. We compare our results with those obtained with other methods. In order to analyze the results obtained with different species, it should be noted that in vertebrates the average length of genome sequence, which corresponds to one protein-coding gene, is much greater than what is found in the other species studied here.

We believe that the abundant sequences that we have discovered deserve further studies with biophysical methods (X-ray diffraction, nuclear magnetic resonance, etc.), in order to determine if they present any specific structural features or may be targets for interaction with drugs and proteins. In particular, some of these sequences may influence nucleosome shape and position in the genome.

METHODS

The sequence data were downloaded from Genbank (http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi). They were analyzed with two different programs, which are described below.

MREPATT

This program has been described elsewhere (8) and is freely available at our website (http://alggen.lsi.upc.edu). It works with genome sequences already stored in our website or introduced by the user. It works with the standard bases A, T, C and G. It calculates the number and distribution of exact tandem repeats of a desired sequence and its complementary one. It also gives graphical representations. A whole chromosome can be viewed and zoomed on it down to fragments of 100 bp where the sequence is represented. The program can calculate several repeat sequences at the same time, but it can only visualize the distribution of individual sequences.

CONREPP

This program, also available at our website, is presented here for the first time. It allows the introduction of alternative base positions in any sequence by using the International Union of Pure and Applied Biophysics (IUPAB) symbols. It has several options. It may calculate the number and distribution of exact tandem repeats, but only on one strand of DNA. The program uses a different graphical representation, where several repeats can be represented in the same figure (Option 1). Palindromic repeats may also be calculated (Options 2–4). Clusters of short repeats can also be found (Option 6).

Another alternative of the program (Option 5) searches for the frequency of short sequences in whole genomes. The inputs of the program are the length of the sequences, one FASTA file, and the number r of most frequent sequences to be found. This program is the main tool used in this paper in order to determine the most frequent short sequences found in each genome. A list of the $r = 100$ most frequent sequences of 9–14 bases is given as part of the Supplementary Data. The list covers the whole genome of four species and two chromosomes with a similar size from man and mouse.

RESULTS

The most frequent short sequences

In the Supplementary Data, we not only give a list of the 100 most frequent sequences of 9–14 bases in the whole genomes of *C. elegans*, *A. thaliana, D. melanogaster* and *H. sapiens*, but also give the list of the most frequent sequences in chromosome 12 of both man and mouse. The latter chromosomes have been chosen because they have a size similar to the whole genomes of the other species; they are not evolutionary related. Note that the relative amount of non-coding DNA in man and mouse is much larger than in the other three species studied. This fact should be borne in mind in the analysis of the results we present below.

Inspection of the list of sequences shows that there are several classes of frequent sequences:

(i) microsatellite fragments;
(ii) microsatellite fragments with one or two bases changed; and
(iii) unique sequences.

The unique sequences are, in general, complex. They belong to two main classes: (i) the partial repeats of shorter sequences, for example, from the telomeres; and (ii) the partial sequences of longer frequent sequences, such as *Alus*.

In all cases, partial redundant sequences are found in the list. For example. in the case of the undecamers in man we find as frequent sequences TGGGATTACAG, CTGG GATTACA, GGGATTACAGG, etc., all of them overlapping fragments of the *Alu* sequence. Thus, the basic list has to be carefully analyzed in order to determine the eventual relationship among different abundant sequences.

We will first briefly discuss the distribution of microsatellite fragments and then present the unique short sequences present in the different species we have analyzed.

Microsatellite fragments

Microsatellites are repeated tandem sequences of a short motif which is 1–6 bases long. However, there is no sharp definition on the length of a tandem repeat in order to be considered a microsatellite. Usually, sequences of over 20–30 bases are considered as microsatellites. In our previous study (5), we defined tandem sequences, which were over 24 bases long, as microsatellites.

The distribution of microsatellite fragments with a short motif is given in Table 1. Microsatellites with a longer motif of three and more bases (data not shown in Table 1) are also very frequent in the mouse and in *D. melanogaster*. A few are also found in *A. thaliana*. A detailed analysis is presented in our previous study (5).

A peculiarity of the microsatellite distribution in *D. melanogaster* is the absence of the very abundant

Table 1. Microsatellite fragments of 11 bases in different genomes

Species	*Caenorhabditis elegans*	*Arabidopsis thaliana*	*Drosophila melanogaster*	*Mus musculus* chromosome 12	*Homo sapiens* chromosome 12
CG %	35.44	36.03	42.46	41.75	40.82
N_r bases	100 269 917	118 997 677	118 348 385	117 459 310	130 303 032
A/T	10 955	20 312	22 434	18 547	39 140
TA	1949	7633	6898	7450	5897
CA/TG	2038	1151	12 419	42 343	11 738
TC/GA	2739	5549	2920	22 731	5216
CG	153	6	18	512	119
G/C	1998	279	2635	3043	255

The number of repeats has been calculated with the CONREPP program. The repeats found in both strands have been added. Only microsatellites with a motif of 1–2 bases are shown.

Table 2. A selection of frequent tetradecamer sequences in the genome of *A. thaliana*

Identification number	Sequence	Numbers of times	Comments
1	$(AAACCCT)_2$	1235	Dimer of telomere sequence
2	TTGGTTAGTGTTTT	1160	Part of centromeric repeat
3	GATGTCATGTGTAT	1259	Part of centromeric repeat
4	AAAGCTTTGATGGT	1221	Part of centromeric repeat

motifs found in heterochromatin (9), which are not present in the euchromatin sequenced genome. Nevertheless, microsatellites with other motifs are very abundant. However, it does not present any unique sequence repeated a substantial number of times. Here, this species is not further analyzed.

Frequent short sequences in *A. thaliana*

In the case of *A. thaliana*, the most frequent sequences are usually strings of As or Ts interrupted with another base. However, there are also a substantial amount of abundant unique sequences distributed in the genome, mainly in the centromeric regions. The most abundant of them are given in Table 2. Sequence 1 is actually part of the telomeric $(AAACCCT)_n$ sequence which has been found to be frequent in some centromeric regions of Arabidopsis (10). The other sequences in the table are part of the centromeric repeat (11,12). Sequence 3 in Table 2 has the peculiarity of presenting a high degree of purine–pyrimidine alternation of sequence, which is found neither in the other sequences in Table 2 nor in those we have found in the other species that we describe next. As an example of the distribution of these sequences, in the Supplementary Data, we give a map of the centromeric region of one chromosome of *A. thaliana* (Supplementary Figure S4). Each chromosome in this species has a different pattern of short repeated sequences in the centromeric region, a question which deserves further studies.

Frequent short sequences in *C. elegans*

The genome of *C. elegans* has a high concentration of protein coding genes, about 1 gene/5 kb. In vertebrates, the density of genes is much lower, about 1 gene/100 kb. Thus, the *C. elegans* genome contains only a comparatively small amount of non-coding DNA. This feature indicates that *C. elegans* is a very appropriate organism in order to study the eventual structural role of non-coding DNA. Furthermore, *C. elegans* has holocentric chromosomes: during mitosis the centromere function is distributed along the whole condensed chromosomes. The centromere-specific histone variant CENP-A, for example, localizes along the whole length of mitotic chromosomes (13). However, no direct relationship has been established between the sequence and distribution of repeated sequences and the localization of centromere/kinetochore activity. Therefore, it appears of interest to study such sequences in greater detail.

Repetitive sequences in *C. elegans* have been studied by several authors (14–18). Some of those sequences were briefly described in the original paper presenting the full sequence of this organism (19). Most of these repeats have been stored either in the Repbase database (7) or in the website of the Sanger Institute (http://www.sanger.ac.uk/Projects/C_elegans/repeats). The data available in each site are complementary, different sequences may be found. Most of them are rather long, occur only a few times in the whole genome and show many mutations. Here, we apply the methods described above and find new features which complement previous results.

In Table 3, we present a list of some of the most frequent tetradecamer unique sequences found in *C. elegans*. A list with the location of sequences 1, 2 and 3 in chromosome I is given in the Supplementary Data.

Inspection of the tables given as Supplementary Data indicates that the ATTTGCCG octamer is very frequent. This sequence is part of 1a in Table 3. In the available databases of repeats mentioned above, we only detected one cluster of this sequence, described as HAT1_CE (7). Its sequence is analyzed in the Supplementary Data (Supplementary Figure S9). However, we have found that ATTTGCCG and related sequences are very abundant throughout the whole genome of *C. elegans*. Calculation with MREPATT shows that ATTTGCCG occurs 28 646 times in the genome, of which 2108 are

Table 3. A selection of frequent tetradecamer sequences in the genome of *C. elegans*

Identification number	Sequence	Number of times	Comments
1a	TTGCCGATTTGCCG	2777 + 2446	Partial repeat of ATTTGCCG
1b	TTTGCCGGAAATTT	2711 + 2328	Associated to ATTTGCCG
1c	AAATTGCCGGAATT	1449 + 1432	Associated to ATTTGCCG
2	TAGGCTTAGGCTTA	1688 + 1730	Telomere-like repeat of TTAGGC
3	GAAATTCAAATTTT	1485 + 1336	Found near *Helitron* transposons (13)
4	ACTACAAACTACAA	1670 + 1499	Dimer of ACTACAA
5	T/m/SA/n	2216	$m \geq 7$ and $n \geq 7$
6	T/m/SA/n	2863	$m \geq 6$ and $n \geq 6$

The first number corresponds to the sequence shown in the table, the second number to the complementary sequence. Sequences 5 and 6 are palindromic, they occur simultaneously in both strands.

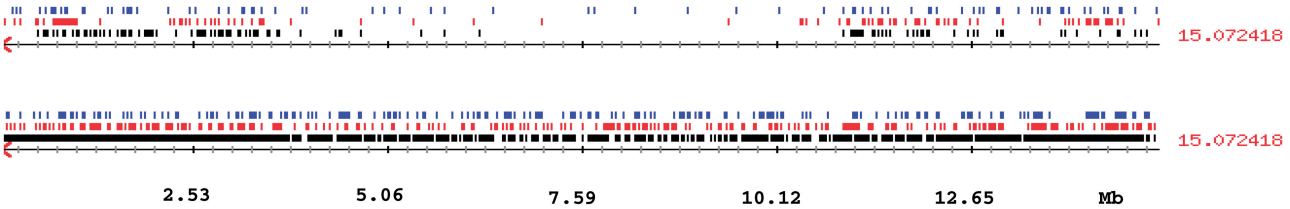

Figure 1. Representation of chromosome I of *C. elegans* obtained with the CONREPP program. In the upper frame are shown the following sequences, described in Table 3: $(ATTTGCCG)_2$ in black; a dimer of the telomere repeat $(TTAGGC)_2$ in red and GAAATTCAAATTTT in blue. In the figure, only the repeats found in the forward strand are presented. A similar representation is found in the complementary strand (data not shown). The sequence of a few repeats and numerical results of their distribution are given as Supplementary Data. In the lower frame are presented at the same scale the following sequences: $(AG)_5$ in blue, $(AT)_5$ in red and A_mT_n in black, with n, $m \geq 5$.

dimers. A few trimers (101 cases) and tetramers (two cases) are found as well. This octamer sequence does not show a significant frequency in the other species studied. Only in *D. melanogaster* a substantial number of cases is found (5090 cases), but only as individual, isolated sequences. In *C. elegans*, most of these octamers occur also as isolated sequences, but there is a significant number which appear associated in clusters together with other abundant sequences (1b and 1c), also shown in Table 3. Practically, all clusters of the latter sequences are absent in the central part of the chromosomes, as demonstrated in Figure 1. The exact sequence for some of them is given as Supplementary Data for chromosome I. In Figure 2, we give a detailed view of one fragment of chromosome I, where the position of some clusters can be viewed in detail. The figure illustrates that each cluster shows a different distribution of sequences, as it is also clear from the examples given as Supplementary Data. Clusters which also include sequence 1c (Table 3) are also common. It is difficult to give exact numbers for the number of clusters, since it will depend on the quantitative definition of a cluster. We can tentatively define a cluster as a segment of 200 bases, which contains at least five ATTTGCCG motifs or five complementary CGGCAAA T motifs. In general, both types of clusters show different positions. With the CONREPP program (Option 6), we have found the number and position of such clusters in all chromosomes, which are given in the Supplementary Data (Supplementary Table S5). We find a total of 1658 clusters, with an average density of 16.5 clusters/Mb. Chromosome X has a significantly lower density of only 2.7 clusters/Mb. However, these numbers are only approximate, since they do not include clusters in which motifs 1b or 1c may also be important. On the other hand, large clusters, such as those given in Supplementary Figures S7 and S8, are counted as several clusters. Examples are given in Figure 2 and in the Supplementary Data. Some of these clusters are very large and cover 1–3 kb. It is striking that each cluster has a very different structure. Comparison of the figures in the Supplementary Data demonstrates that some of them are very regular (Supplementary Figure S7). Other clusters appear to have evolved from a regular arrangement, but display many point mutations, insertions and deletions (Supplementary Figure S6). In very large clusters (Supplementary Figure S8), a complex organization is apparent, with internal long repeats. Different combinations of the 1a, 1b and 1c sequences, given in Table 3, are found in such cases. The other sequences given in Table 3 are not present in any of these clusters.

The other abundant tetradecamer sequences (Table 3) are also found at both ends of the chromosomes. Some of them are shown in Figure 1. For comparative purposes, we also present the distribution of two short microsatellites of 10 bases in Figure 1. They appear to be evenly distributed along the whole chromosome.

The telomere-like repeat (sequence 2 in Table 3) also occurs in clusters. However, the number of clusters found is smaller than those found with the ATTTGCCG motif. An example is given in Supplementary Figure S5. In this case, the original hexamer structure is clearly maintained, but with many frequent point mutations. The telomere-like repeats may be considered as heavily mutated microsatellites. Internal telomeric repeats have been described in *C. elegans* (20) and in other species (21,22); however, in *C. elegans* they are much more

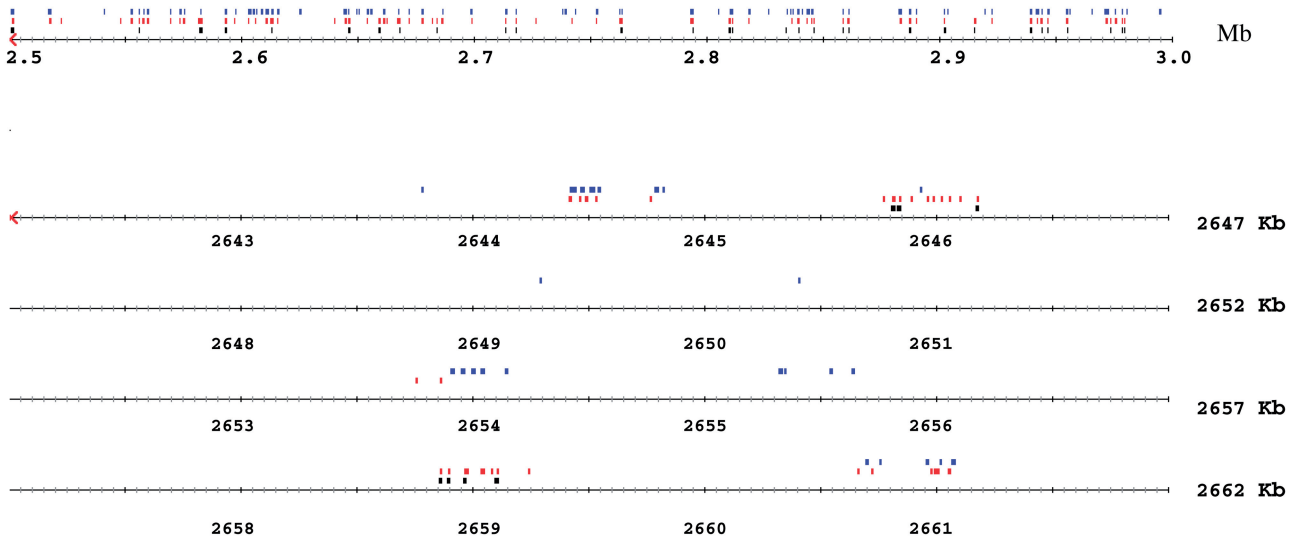

Figure 2. Example of the irregular size and distribution of clusters of repeated sequences. In the upper frame, a region of 500 kb of chromosome I of *C. elegans* is shown. Below is presented an enlargement of the region around 2.65 Mb covering 20 kb. Each cluster covers between 200 and 500 bases. The following sequences are shown: AAATTTCCGGCAAA (complementary to 1b in Table 3) in black; CGGCAAAT in red and its complementary sequence ATTTGCCG in blue. Note that the sequences presented here are different from those shown in Figure 1.

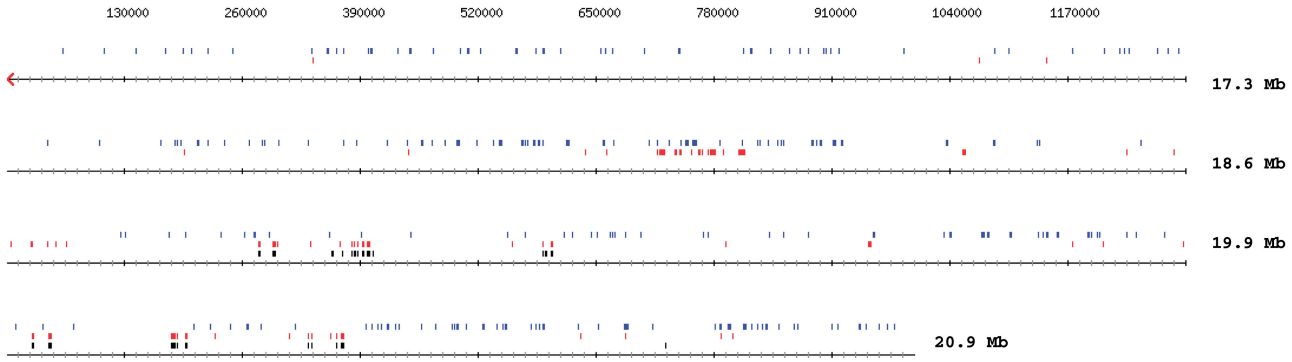

Figure 3. Distribution of different short sequences at the terminal end of chromosome V of *C. elegans* (16–20.9 Mb). The abundant dodecamer TTGGGCGCTGCT described by Phillips *et al.* (23) is shown in black. In red the repeat 4 (Table 3), which is associated with TTGGGCGCTGCT, also shows additional clusters. An example of the detailed sequence of several clusters of the latter two sequences is given in the Supplementary Data (Supplementary Figure S12). The abundant sequence ATTTGCCG (Part of 1a in Table 3) is shown in blue. The latter is excluded in the regions occupied by the other two sequences shown in the figure. Other repeated short sequences are also present in this region (Supplementary Figure S13).

abundant and longer. We have found about 900 clusters in the whole genome (data not shown).

The third sequence studied (Nr 3 in Table 3) is part of a longer sequence of 35 bases:

GAAATTCAAATTTTCAGTGAAAAAAATTTTGG CGG

This sequence occurs in all chromosomes, but it is particularly abundant in chromosomes IV (66 times) and V (59 times). It is rather rich in AT (71%), when compared with the other sequences given in Table 3. Point mutations are common. These sequences are usually associated in longer repeats of up to 17 times the unit of 35 bases. Thus, they may be considered as satellites with a repeat of 35 bases which shows a few mutations. This sequence has already been reported as associated to *Helitron* transposons (7,18). It is also most frequent in the terminal regions of chromosomes (Figure 1), although there are exceptions. In chromosome I, for example, there is a cluster of 22 repeats around position 7.741 Mb. (A list with their positions is given in the Supplementary Data.) In Figure 1, it appears as a single blue line due to the low magnification of the figure.

Sequence 4 has a more restricted distribution. It is mainly present in one end of chromosome I (Supplementary Figure S10), but occurs at one end of chromosome V as well (Figure 3). In the other chromosomes, it is less frequent.

An additional class of frequent sequences are A_mT_n and T_mA_n, which are much more frequent in *C. elegans* then in the other species studied (data not shown). For m, $n \geq 5$ there are about an equal number of A_mT_n and T_mA_n sequences (11 850 and 11 693). If m, $n \geq 7$, then the number of T_mA_n is very small, 98 cases, whereas the number of A_mT_n motifs is comparable to other tetradecamer sequences (Table 3). Some of them are presented in Figure 1. Related sequences are also common,

such as T_mSA_n (S = G or C), which occurs 12 549 times, with $m \geq 5$, $n \geq 5$. A_mST_n occurs less frequently. All these sequences are also more abundant at the ends of chromosomes, but do not tend to form clusters. Instead, they are found embedded in long regions which contain large amounts of adenine and thymine repeats. With regard to their eventual role, it should be noted that the A_mT_n sequence is expected to be rigid, whereas T_mSA_n always presents a flexible TG/CA step in the center. These differences in rigidity will certainly influence their position in nucleosomes.

A general feature of all the sequences which we have described, including most long microsatellites (data not shown), is their much lower frequency in the central part of all chromosomes. They tend to accumulate in regions which span about one-third of the chromosomes al each end. The central 40% part of each chromosome is, in general, depleted of such sequences. Another peculiarity is the lower frequency of all these repeated sequences in the X chromosome, as it is apparent by inspection of the results on clusters given in the Supplementary Data.

Regions of special interest in the chromosomes of *C. elegans* are the pairing centers, present at one end of each chromosome. These centers contain specific sequences (17,23) which appear to be a starting point for homologous pairing and synapsis during meiosis. They also associate with the nuclear envelope, in a way similar to the bouquet conformation described in many species (24,25). It is not clear (24,26) which is the relative importance of the telomere and pairing center sequences in order to establish the interaction with the nuclear membrane. Previous studies (17,23) have emphasized the distribution of some of these sequences, but other short sequences are also present in these regions (Supplementary Figure S13). The overall distribution is quite complex. Examples are given in Figure 3 and in the Supplementary Data (Supplementary Figures S10–S13). It is therefore not clear if the specific sequences described at the end of the *C. elegans* chromosomes (17, 23) are the only players in establishing the pairing center function of these regions.

As a general conclusion, our results show that the chromosomes of *C. elegans* are punctuated by a series of clusters which contain similar motifs, but have very different individual sequences. Such clusters are clear structural marks in the non-coding regions of the *C. elegans* genome. We consider their eventual role in the discussion.

The *Alu* sequence and related SINEs

Some of the most frequent unique sequences in the human chromosome 12 are given in Table 4. The list of common sequences in the whole human genome and in its chromosome 12 is practically identical, as shown in the Supplementary Data. In Table 4, we give one selected sequence in each of the five most conserved regions of the *Alu* sequence (Figure 4). The *Alu* sequence is known to occur about 1 million times in the human genome: it is the most frequent SINE. Inspection of the list of the 100 most frequent sequences in the human genome (given as Supplementary Data) shows that, as expected, most of them are either microsatellites or part of the *Alu* repeat. In Figure 4, we represent two members of the *Alu* family. The undecamers which appear more frequently in the whole human genome are highlighted in bold type. They overlap and cover five different regions in the whole *Alu* sequence. It is clear that there are several regions in the *Alu* sequences which are more conserved than the rest. The first and third regions correspond to an internal repeat, which explains the higher frequency of Alu-1 in Table 4. It is interesting to note that all of the latter sequences are present in the consensus sequence determined by Price *et al.* (27). Our results complement their observations, showing which regions of the sequences are more highly conserved throughout the human genome. These conserved regions appear to have a rigid conformation, since they contain many polypurine and polypyrimidine tracts (Figure 4).

Table 4. A selection of frequent tetradecamer sequences in chromosomes 12 of man and mouse

Alu-1	CCTGTAATCCCAGC	15 933 + 16 036
Alu-2	CTAAAAATACAAAA	9529 + 9653
Alu-4	TGCACTCCAGCCTG	10 765 + 11 000
Alu-5	TCTCAAAAAAAAAA	7257 + 7158
BI-1	CAGCCTGGTCTACA	2728 + 2807
BI-2	CTTTAATCCCAGCA	2582 + 2528

The Alu sequences reported in the table correspond to the different highly conserved regions shown in Figure 4. Alu-3 is not shown; it is an internal duplication of Alu-1.

It should be noted that the number of frequent sequences shown in Table 4 is significantly lower than the number of *Alu* sequences in chromosome 12, which is about 40 000. This is because point mutations, insertions or deletions have not been counted. Inspection of a few *Alu* sequences shows, for example, that the Alu-2 motif undergoes frequent changes, while it always conserves adenine tracts of variable length. It should be noted that in the study of Price *et al.* (27), insertions and deletions were not considered.

Another feature of the *Alu* sequence which we have noticed is the presence of a tract of adenines at the end of it, after the terminal TCTCA sequence. In fact, the TCT CA_{10}/T_{10}GAGA sequences occur 14 415 times in human chromosome 12 (Table 4). Thus, between the Alu-2 region and the end of the *Alu* sequence, there is a sequence of 149 bases which is limited by two adenine tracts, a fact which suggests that this part of the *Alu* sequence determines a well-defined nucleosome structure, according to current models of sequence features which define nucleosome position (28,29). In agreement with this suggestion, Englander and Howard (30; Figure 4) have reported that part of this region gives a strong rotational signal. The first part of the sequence before the Alu-2 motif is shorter. No well defined A-tract, has been detected ahead of it. Thus, this region of the *Alu* sequence does not appear to clearly define a nucleosome position.

In the mouse there is also a widespread SINE called B1 (31), although it occurs less frequently than the human *Alu*. The two most frequent unique repeats, which are part of this SINE sequence, are also given in Table 4.

>**AluSx:**

Alu-1

ggccgggcgcggtggctcac**GCCTGTAATCCCAGCACTTTGGGAGGC**cgaggcgggcggatcacctgaggt

Alu-2

caggagttcgagaccagcctggccaacatggtgaaaccccg**TCTCTACTAAAAATACAAAAA**ttagccggg

Alu-3

cgtggtggcgcgc**GCCTGTAATCCCAGCTACT**c**GGGAGGCTGAGGCAGGAG**aatcgcttgaacccgggaggc

Alu-4 **Alu-5**

ggagg**TTGCAGTGAGC**cgagatcgcg**CCACTGCACTCCAGCCTGGG**cgacagagcgagactccg**TCTCAAAAAAAAAA**

>**AluJo:**

ggccgggcgcggtggctcac**GCCTGTAATCCCAGCACTTTGGGAGGC**cgaggcgggaggattgcttgagcc

caggagttcgagaccagcctgggcaacatagcgagaccccgtctctaca**AAAAATACAAAAA**ttagccggg

cgtggtggcgcgcgcctgtag**TCCCAGCTACT**c**GGGAGGCTGAGGCAGGAG**gatcgcttgagcccaggagt

tcgaggctgcagtgagctatgatcgcg**CCACTGCACTCCAGCCTGGG**cgacagagcgagaccctgtctca

Figure 4. Sequence of two *Alu* repeat elements. J_0 is considered to be the oldest sequence and S_x the most abundant (27). Undecamer regions of the sequences in bold type are present over 400 000 times in the human genome. All of them are present in the consensus sequence determined by Price *et al*. (27). The names of the five most conserved regions correspond to those given in Table 4. The J_0 and S_x sequences have been downloaded from RepBase (7). A string of nine adenines was added at the end of the S_x sequence, since it was found to be also highly conserved. The only region which shows a highly alternating purine/pyrimidine sequence is underlined.

When mouse and man are compared, it is clear that microsatellites in the mouse are comparatively much more abundant (Table 1), a fact which is also apparent by inspection of the list of the 100 most frequent sequences given in the Supplementary Data.

It is interesting to note that two of the motifs (Alu-1 and BI-2) given in Table 4 are clearly related. The undecamer sequence TAATCCCAGCA is very abundant in both B1 and *Alu* sequences. Looking at the context of these sequences, the CTTT tetramer found at the start of BI-2 is also found next to the end of Alu-1 (Figure 4). The evolutionary conservation of this sequence might have a structural significance.

DISCUSSION

The sequences

Inspection of the most frequent sequences in Tables 2–4 shows that short clusters of purines/pyrimidines are frequent in all of them, such as AAA, GGAA, TTT, CCC, CTT, etc. Most of the 100 abundant sequences given in the Supplementary Data show this feature. The only exception is repeat 3 in *A. thaliana* (Table 2), which has a highly alternating purine–pyrimidine sequence. Another alternating region is found in the *Alu* sequence, which is indicated in Figure 4. Alternation is only favored in some microsatellites, such as $(TA)_n$ and $(CA)_n$. The latter is particularly abundant in *M. musculus* and *D. melanogaster*, as well as in *H. sapiens*, although in the latter species it has a lower relative frequency.

It has been shown (32) that such short clusters of purines/pyrimidines may decrease the flexibility of oligonucleotides, but there are no crystallographic data on sequences such as those given in Tables 2–4. Therefore, it appears of interest to study in greater detail the structural features of these abundant sequences, which show significant differences with the highly studied d(CGC GAATTCGCG). It should be noted that the latter sequence is very rare in most genomes. This fact is not surprising, since it contains four times the CG dimer, which is known to have a low probability of appearance. In particular, for the study of drugs which interact with DNA, sequences which are frequent in the target organism should be used in structural studies. In addition, their influence on nucleosome structure should be established. The methods that we have presented in this paper may be used to determine which sequences might be the most appropriate to study in each particular case.

Clusters of frequent sequences such as those presented in Figure 2 and in the Supplementary Data may also have a strong influence on the formation and position of nucleosomes (28,29). They may either induce variant nucleosome structures or prevent nucleosome formation. It is well known that DNA sequence has a strong influence on nucleosome structure, either by the formation of very stable nucleosomes (33–35) or by preventing their formation (36). In particular it has been shown (37, Figure 3) that GCCGG/CCGGC and TGCCG/CGGCA are not frequent within nucleosomes of *C. elegans*. The latter sequences are part of the frequent tetradecamers shown in Table 3. In fact, when the nucleosome prediction program of Kaplan *et al*. (37) is applied to the sequence shown in Figure 2, no nucleosome positions are predicted in the regions occupied by ATTTGCCG and related sequences. However, it cannot be excluded that they form unusual nucleosome structures (38,39). It is obvious that the role of these sequences in chromatin structure deserves a detailed study.

A particular class of sequences which deserves further studies are those found in telomeres. Practically, all eukaryotes (excluding yeasts and *Drosophila*) have a telomere consensus sequence of the type $(C)T_{2-4}(A)G_{2-4}$, which is also a repeat of alternating clusters of purines and pyrimidines. Again, very few structural data are available on DNAs with such sequences. For the biological function of telomeres, their DNA structure in the form of a duplex might be of significance. In fact, it is unlikely that the telomere sequence of *C. elegans* might be able to form guanine quadruplexes, since it contains only two guanines.

Clusters of non-coding sequences and meiosis

We may now ask if the frequent sequences we have found might have any structural role. It should be noted that exons in protein-coding genes should not be expected to play any structural role. They have evolved to optimize its coding function; they will normally form standard nucleosomes. On the other hand, some of the non-coding sequences may have evolved to play a structural role important for chromosome function (4).

In *A. thaliana*, the frequent sequences we have localized are all found in the pericentromeric regions. In *D. melanogaster*, we have not detected any short frequent sequence, but it is known that in the centromeres (9) they have abundant unique microsatellite and other sequences (40), which we have not found in the sequenced euchromatin of the genome. Microsatellites with a different sequence (5) are frequent in both species. A list of the simplest sequences is given in Table 1. They have a random distribution throughout the euchromatin region, they do not form clusters (data not shown), although isolated sequences of 10–20 bases are common. Longer sequences are occasionally found (5). Thus, both *A. thaliana* and *D. melanogaster* only have unique frequent short sequences in the pericentromeric regions.

A different situation is found in *C. elegans*, where clusters of repeated sequences such as those given in Table 3 are frequent all along the chromosomes. Examples are given in Figure 2 and in the Supplementary Data. Their eventual role in nucleosome formation has been discussed above. Although each cluster uses the same sequences, their local organization is different in each cluster. These abundant short sequences are distributed throughout the genome in a very unique manner. It appears that these clusters are structural punctuation marks throughout the genome. Although they do not have an identical sequence, they share features which may indicate a similar structural role. Since this organism has holocentric chromosomes, it is possible that some of these clusters have a centromere-like function, including those which have repeats of telomere sequences (41). The clusters of repeated sequences in *C. elegans* may thus be considered as spread out centromeric repeats. Our results suggest that they might be considered analogous to the pericentromeric regions in other organisms. They might also be involved in meiosis, as discussed below.

As found in practically all organisms which undergo meiosis, as reviewed by different authors (24–26,42–45), the first step in the premeiotic/leptotene stage is mutual recognition of identical DNA sequences. This recognition usually starts at special places in the chromosomes, such as telomeres (bouquet conformation), pairing centers, centromeres, heterochromatic regions, etc. Many models have been suggested to explain mutual recognition, as reviewed by Weiner and Kleckner (45). Some of them suggest a role for repetitive DNA. The availability of whole genome sequences allows a more detailed study of this issue. In particular, in *C. elegans*, pairing starts at specific regions at the end of chromosomes (25,46). In the Supplementary Data, we have given some examples of unique regions at the end of chromosomes (Supplementary Figures S10–S13). In *A. thaliana*, pairing starts at the centromeres (43), each of them has a unique distribution of repeated sequences. An example is given in Supplementary Figure S4. In *D. melanogaster*, paring also starts at the centromeres (42), each of which has a different distribution of satellite sequences (9).

Chromosome pairing requires additional regions throughout the chromosome, which will facilitate mutual recognition. If the usual starting region is damaged, chromosome pairing still occurs (26). Coding regions do not appear appropriate for such recognition, they are usually forming standard nucleosome structures which do not appear capable of recognizing homologous sequences. Non-coding sequences with unique structural features may be involved in homologous chromosome pairing, similar to the teeth in a zipper. However, homologous recognition requires that each tooth has a specific structure. The clusters of ATTTGCCG and related sequences (Table 3) may play such a role in homolog recognition. Each cluster will have a different structure depending on the exact combination of repeats. If these marks should allow recognition of homologous sequences, it must be assumed that they are able to recognize clusters with an identical sequence in the sister chromosome. Recognition could be either due to proteins which interact directly with the ATTTGCCG clusters or due to specific nucleosome–nucleosome interactions (47) promoted by changes in proteins in the nucleoplasm, which may create a favorable environment for homologous DNA recognition. Alternatively, a direct DNA–DNA interaction cannot be excluded. It is also possible that clusters of repeat regions may coalesce with their neighbors along the same chromosome and contribute to the formation of the axial core of chromosomes found in leptotene in many species (48,49).

The interaction between the silent mating-type regions HML and HMR in the yeast *Saccharomyces cerevisiae* suggests one of the possible models for the hypothetical interactions we have described. These two regions are found at both ends of chromosome III at a distance of about 280 kb. A nucleosome free region of each of these two sites is associated with a specific multiprotein complex (50). In spite of their distance in the chromosome, they may pair and form a loop (51). Such an interaction is conceptually similar to the recognition we have postulated between clusters of related sequences in homologous chromosomes in *C. elegans*. Another related model are the

CTCF proteins, which recognize multiple G-rich sequences in the human genome (52). The CTCF–DNA complexes can interact and also give rise to chromatin loops (53).

In *D. melanogaster,* homologous chromosomes are paired throughout mitosis and meiosis (42), also in giant chromosomes. In that case, pairing could be facilitated by clusters of microsatellite sequences. In this organism, homolog pairing also starts in the centromere (42), where microsatellites with a unique distribution are present (9,54). In the case of vertebrates, clusters of SINE, *Alu* and some microsatellite sequences could also be involved.

In order to demonstrate the hypothesis we have presented, it is first necessary to show that the abundant sequences we have found (such as ATTTGCCG and related repeats in *C. elegans*) do indeed determine special chromatin regions which might either exclude nucleosomes or form unique nucleosome structures, a question which deserves detailed studies. In any case, it is clear that these abundant short sequences provide unique structural marks. Unfortunately, the molecular target of the proteins (55–57), which participate in the specific structure of meiotic chromosomes, is not known.

CONCLUSIONS

The methods we have introduced show that some short nucleotide sequences are very abundant throughout genomes. They usually contain short clusters of contiguous purines/pyrimidines. It is clear that their structural properties and distribution in the genomes should be studied in greater detail, in order to determine their affinity for chromosome structural proteins, their influence in the formation of nucleosome structures, their potential for drug binding and other genome features. We have suggested a role for some of them in chromosome pairing in meiosis. As stated in a recent paper, 'the question of how chromosomes recognize their appropriate partners remains unclear' (23). In any case, it is clear that they provide structural marks in the genome, which may provide an explanation for the abundance of non-coding regions in eukaryotic genomes. Similar structural roles may hide under the variability in sequence of many non-coding sequences.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Drs F. Azorín, M. Blasco and M. Chiva for advice and discussions.

FUNDING

The Ministerio de Ciencia e Innovación, Spain (TIN2004-03382, TIN 2007-68093-C02-01 and BFU2006-04035); EU programme FEDER. Funding for open access charge: Departament d'enginyeria quimica, UPC, Barcelona and Ministerio de Ciencia e Innovacion, TIN 2007-68093-C02-01.

Conflict of interest statement. None declared.

REFERENCES

1. Doolittle,W.F. and Sapienza,C. (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature*, **284**, 601–603.
2. Orgel,L.E. and Crick,F.H. (1980) Selfish DNA: the ultimate parasite. *Nature*, **284**, 604–607.
3. Amaral,P.P., Dinger,M.E., Mercer,T.R. and Mattick,J.S. (2008) The eukaryotic genome as an RNA machine. *Science*, **319**, 1787–1789.
4. Shapiro,J.A. and Von Sternberg,R. (2005) Why repetitive DNA is essential to genome function. *Biol. Rev*., **80**, 1–24.
5. Subirana,J.A. and Messeguer,X. (2008) Structural families of genomic microsatellites. *Gene*, **408**, 124–132.
6. Smit,A.F.A. (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev*., **9**, 657–663.
7. Jurka,J. (2000) Repbase update a database and an electronic journal of repetitive elements. *Trends Genet*., **16**, 418–420.
8. Roset,R., Subirana,J.A. and Messeguer,X. (2003) MREPATT: detection and analysis of exact consecutive repeats in genomic sequences. *Bioinformatics*, **19**, 2475–2476.
9. Lohe,A.R., Hilliker,A.J. and Roberts,P.A. (1993) Mapping simple repeated DNA sequences in heterochromatin of *Drosophila melanogaster*. *Genetics*, **134**, 1149–1174.
10. Richards,E.J., Goodman,H.M. and Ausubel,F.M. (1991) The centromere region of *Arabidopsis thaliana* chromosome 1 contains telomere-similar sequences. *Nucleic Acids Res*., **19**, 3351–3357.
11. Murata,M., Ogura,Y. and Motoyoshi,F. (1994) Centromeric repetitive sequences in *Arabidopsis thaliana*. *Jpn. J. Genet*., **69**, 361–370.
12. Heslop -Harrison,J.S., Murata,M., Ogura,Y., Schwarzacher,T. and Motoyoshi,F. (1999) Polymorphisms and genomic organization of repetitive DNA from centromeric regions of Arabidopsis chromosomes. *Plant Cell*., **11**, 31–42.
13. Maddox,P.S., Oegema,K., Desai,A. and Cheeseman,I.M. (2004) "Holo"er than thou: chromosome segregation and kinetochore function in *C. elegans*. *Chromosome Res*., **12**, 641–653.
14. Nacleiro,G., Cangiano,G., Coulson,A., Levitt,A., Ruvolo,V. and La Volpe,A. (1992) Molecular and genomic organization of clusters of repetitive DNA sequences in *Caenorhabditis elegans*. *J. Mol. Biol*., **226**, 159–168.
15. Surzycki,S.A. and Belknap,W.R. (2000) Repetitive-DNA elements are similarly distributed on *Caenorhabditis elegans* autosomes. *Proc. Natl Acad. Sci. USA*, **97**, 245–249.
16. LeBlanc,M.D., Aspeslagh,G., Buggia,N.P. and Dyer,B.D. (2000) An annotated catalog of inverted repeats of *Caenorhabditis elegans* chromosomes III and X, with observations concerning odd/even biases and conserved motifs. *Genome Res*., **101**, 1381–1392.
17. Sanford,C. and Perry,M.D. (2001) Asymmetrically distributed oligonucleotide repeats in the *Caenorhabditis elegans* genome sequence that map to regions important for meiotic chromosome segregation. *Nucleic Acids Res*., **29**, 2920–2926.
18. Kapitonov,V.V. and Jurka,J. (2001) Rolling-circle transposons in eukaryotes. *Proc. Natl Acad. Sci. USA*, **98**, 8714–8719.
19. The *C.elegans* Sequencing Consortium. (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.
20. Cangiano,G. and La Volpe,A. (1993) Repetitive DNA sequences located in the terminal portion of the *Caenorhabditis elegans* chromosomes. *Nucleic Acids Res*., **5**, 1133–1139.
21. Regad,F., Lebas,M. and Lescure,B. (1994) Interstitial telomeric repeats within the *Arabidopsis thaliana* Genome. *J. Mol. Biol*., **239**, 163–169.
22. Cherry,J.M. and Blackburn,E.H. (1985) The internally located telomeric sequences in the germ-line chromosomes of Tetrahymena are at the ends of transposon-like elements. *Cell*, **43**, 747–758.

23. Phillips,C.M., Meng,X., Zhang,L., Chretien,J.H., Urnov,F.D. and Dernburg,A.F. (2009) Identification of chromosome sequence motifs that mediate meiotic pairing and synapsis in *C. elegans*. *Nat. Cell Biol.*, **11**, 934–943.
24. Bhalla,N. and Dernburg,A.F. (2008) Prelude to a division. *Annu. Rev. Cell Dev. Biol.*, **24**, 397–424.
25. Naranjo,T. and Corredor,E. (2008) Nuclear architecture and chromosome dynamics in the search of the pairing partner in meiosis in plants. *Cytogenet. Genome Res.*, **120**, 320–330.
26. Zetka,M. (2009) Homologue pairing, recombination and segregation in *Caenorhabditis elegans*. In Benavente,R. and Volff,J.N. (eds), *Meiosis Genome Dyn.*, Vol. 5. Karger, Basel, pp. 43–55.
27. Price,A.L., Eskin,E. and Pevzner,P.A. (2004) Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. *Genome Res.*, **14**, 2245–2252.
28. Segal,E. and Widom,J. (2009) Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.*, **19**, 65–71.
29. Segal,E. and Widom,J. (2009) What controls nucleosome positions? *Cell*, **25**, 335–343.
30. Englander,E.W. and Howard,B.H. (1995) Nucleosome positioning by human Alu elements in chromatin. *J. Biol. Chem.*, **270**, 10091–10096.
31. Krayev,A.S., Kramerov,D.A., Skryabin,K.G., Ryskov,A.P., Bayev,A.A. and Georgiev,G.P. (1980) The nucleotide sequence of the ubiquitous repetitive DNA sequence B1 complementary to the most abundant class of mouse fold-back RNA. *Nucleic Acids Res.*, **8**, 1201–1215.
32. Kahn,T.R., Fong,K.K., Jordan,B., Lek,J.C., Levitan,R., Mitchell,P.S., Wood,C. and Hatcher,M.E. (2009) An FTIR investigation of flanking sequence effects on the structure and flexibility of DNA binding sites. *Biochemistry*, **48**, 1315–1321.
33. Bussiek,M., Hoischen,C., Diekmann,S. and Bennink,M.L. (2009) Sequence-specific physical properties of African green monkey alpha-satellite DNA contribute to centromeric heterochromatin formation. *J. Struct. Biol.*, **167**, 36–46.
34. Widlund,H.R., Kuduvalli,P.N., Bengtsson,M., Cao,H., Tullius,T.D. and Kubista,M. (1999) Nucleosome structural features and intrinsic properties of the TATAAACGCC repeat sequence. *J. Biol. Chem.*, **274**, 31847–31852.
35. Widlund,H.R., Cao,H., Simonsson,S., Magnusson,E., Simonsson,T., Nielsen,P.E., Kahn,J.D., Crothers,D.M. and Kubista,M. (1997) Identification and characterization of genomic nucleosome-positioning sequences. *J. Mol. Biol.*, **267**, 807–817.
36. Cao,H., Widlund,H.R., Simonsson,T. and Kubista,M. (1998) TGGA repeats impair nucleosome formation. *J. Mol. Biol.*, **281**, 253–260.
37. Kaplan,N., Moore,I.K., Fondufe-Mitterndorf,Y., Gossett,A.J., Tillo,D., Field,Y., LeProust,E.M., Hughes,T.R., Lieb,J.D., Widom,J. *et al.* (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **458**, 362–366.
38. Zlatanova,J., Bishop,T.C., Victor,J.M., Jackson,V. and van Holde,K. (2009) The nucleosome family: dynamic and growing. *Structure*, **17**, 160–171.
39. Furuyama,T. and Henikoff,S. (2009) Centromeric nucleosomes induce positive DNA supercoils. *Cell*, **138**, 104–113.
40. Méndez-Lago,M., Wild,J., Whitehead,S.L., Tracey,A., de Pablos,B., Rogers,J., Szybalski,W. and Villasante,A. (2009) Novel sequencing strategy for repetitive DNA in a *Drosophila* BAC clone reveals that the centromeric region of the Y chromosome evolved from a telomere. *Nucleic Acids Res.*, **37**, 2264–2273.
41. Villasante,A., Méndez-Lago,M., Abad,J.P. and Montejo de Garcini,E. (2007) The birth of the centromere. *Cell Cycle*, **6**, 2872–2876.
42. McKee,B.D. (2004) Homologous pairing and chromosome dynamics in meiosis and mitosis. *Biochim. Biophys. Acta*, **1677**, 165–180.
43. Zickler,D. (2006) From early homologue recognition to synaptonemal complex formation. *Chromosoma*, **115**, 158–174.
44. Moore,G. and Shaw,P. (2009) Improving the chances of finding the right partner. *Curr. Opin. Struct. Biol.*, **19**, 99–104.
45. Weiner,B.M. and Kleckner,N. (1994) Chromosome pairing via multiple interstitial interactions before and during meiosis in yeast. *Cell*, **77**, 977–991.
46. Albertson,D.G., Rose,A.M. and Villeneuve,A.M. (1997) Chromosome organization in mitosis and meiosis. In Riddle,D.L., Blumenthal,T., Meyer,B.J. and Priess,J.R. (eds), C. elegans II. Cold Spring Harbor Laboratory Press, New York, pp. 47–78.
47. Chodaparambil,J.V., Barbera,A.J., Lu,X., Kaye,K.M., Hansen,J.C. and Luger,K. (2007) A charged and contoured surface on the nucleosome regulates chromatin compaction. *Nat. Struct. Mol. Biol.*, **14**, 1105–1107.
48. Moens,P.B. (1968) The structure and function of the synaptinemal complex in *Lilium longiflorum* sporocytes. *Chromosoma*, **23**, 418–451.
49. Comings,D.E. and Okada,T. (1970) Whole mount electron microscopy of meiotic chromosomes and the synaptonemal complex. *Chromosoma*, **30**, 269–286.
50. Fox,C.A. and McConnell,K.H. (2005) Toward biochemical understanding of a transcriptionally silenced chromosomal domain in Saccharomyces cerevisiae *J. Biol. Chem.*, **280**, 8629–8632.
51. Miele,A., Bystricky,K. and Dekker,J. (2009) Yeast silent mating type loci form heterochromatic clusters through silencer protein-dependent long-range interactions. *PLoS Genet.*, **5**, 1–18.
52. Ohlsson,R., Renkawitz,R. and Lobanenkov,V. (2001) CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Biochem. Sci.*, **9**, 520–522.
53. Phillips,J.E. and Corces,V.G. (2009) CTCF: master weaver of the genome. *Cell*, **137**, 1194–1211.
54. Sun,X., Hiep,D.L., Wahlstrom,J.M. and Karpen,G.H. (2003) Sequence analysis of a functional *Drosophila* centromere. *Genome Res.*, **13**, 182–194.
55. Goodyer,W., Kaitna,S., Couteau,F., Ward,J.D., Boulton,S.J. and Zetka,M. (2008) HTP-3 links DSB formation with homolog pairing and crossing over during *C. elegans meiosis*. *Dev. Cell*, **14**, 263–274.
56. Csankovszki,G., Collette,K., Spahl,K., Carey,J., Snyder,M., Petty,E., Patel,U., Tabuchi,T., Liu,H., McLeod,I. *et al.* (2009) Three distinct condensin complexes control *C. elegans* chromosome dynamics. *Curr. Biol.*, **19**, 9–19.
57. Severson,A.F., Ling,L., van Zuylen,V. and Meyer,B.J. (2009) The axial element protein HTP-3 promotes cohesion loading and meiotic axis assembly in *C. elegans* to implement the meiotic program of chromosome segregation. *Genes Dev.*, **23**, 1763–1778.