Chromosome Res (2013) 21:461–474
DOI 10.1007/s10577-013-9371-y

# Notable clustering of transcription-factor-binding motifs in human pericentric regions and its biological significance

**Yuki Iwasaki · Kennosuke Wada · Yoshiko Wada · Takashi Abe · Toshimichi Ikemura**

Received: 10 May 2013 / Revised: 14 June 2013 / Accepted: 14 June 2013 / Published online: 30 July 2013


**Abstract** Since oligonucleotide composition in the genome sequence varies significantly among species even among those possessing the same genome G+C%, the composition has been used to distinguish a wide range of genomes and called as "genome signature". Oligonucleotides often represent motif sequences responsible for sequence-specific protein binding (e.g., transcription-factor binding). Occurrences of such motif oligonucleotides in the genome should be biased compared to those observed in random sequences and may differ among genomes and genomic portions. Self-Organizing Map (SOM) is a powerful tool for clustering high-dimensional data such as oligonucleotide composition on one plane. We previously modified the conventional SOM for genome informatics to batch learning SOM or "BLSOM". When we constructed BLSOMs to analyze pentanucleotide composition in 20-, 50-, and 100-kb sequences derived from the human genome, BLSOMs did not classify human sequences according to chromosome but revealed several specific zones composed primarily of sequences derived from pericentric regions. Interestingly, various transcription-factor-binding motifs were characteristically overrepresented in pericentric regions but underrepresented in most genomic sequences. When we focused on much shorter sequences (e.g., 1 kb), the clustering of transcription-factor-binding motifs was evident in pericentric, subtelomeric and sex chromosome pseudoautosomal regions. The biological significance of the clustering in these regions was discussed in connection with cell-type and -stage-dependent chromocenter formation and nuclear organization.

**Keywords** Pericentric region · Centromeric heterochromatin · Chromocenter · Transcription-factor-binding · Oligonucleotide composition · Self-organizing map

Responsible Editor: Tatsuo Fukagawa.

**Electronic supplementary material** The online version of this article (doi:10.1007/s10577-013-9371-y) contains supplementary material, which is available to authorized users.

Y. Iwasaki · K. Wada · Y. Wada · T. Ikemura (✉)
Department of Bioscience, Nagahama Institute of Bio-Science and Technology,
Nagahama-shi, Shiga-ken 526-0829, Japan
e-mail: t_ikemura@nagahama-i-bio.ac.jp

Y. Iwasaki
Japan Society for the Promotion of Science,
Chiyoda-ku, Tokyo, Japan

T. Abe
Institute of Science and Technology, Department of Information Engineering, Faculty of Engineering,
Niigata University,
Niigata-ken 950-2181, Japan

## Introduction

The chromosome research in the post genome sequencing era needs a bioinformatics method that can analyze even a big sequence data, in order to connect a wide variety of chromosome characteristics with chromosome-level features of genomic sequences. G+C% has long been used as

a fundamental value for classifying not only inter- but also intra-genomic characteristics, and actually genomes of warm-blooded vertebrates are known to be composed of a long-range segmental G+C% distribution designated as "isochore", which has been connected with chromosomal bands (Bernardi et al. 1985; Bernardi 2004; Ikemura 1985; Ikemura and Aota 1998). The G+C%, however, is apparently too simple a parameter to differentiate a wide variety of genome characteristics, but oligonucleotide composition can distinguish the genomes even with the same GC% because the oligonucleotide composition varies significantly among the genomes and is called "genome signature" (Karlin et al. 1998; Gentles and Karlin 2001).

The self-organizing map (SOM) is a powerful tool for clustering and visualizing high-dimensional complex data on one plane (Kohonen et al. 1996). For the codon and oligonucleotide composition handled as high-dimensional data, we modified the conventional SOM to batch learning SOM or "BLSOM" (Kanaya et al. 2001; Abe et al. 2003). On BLSOMs for oligonucleotide compositions, genomic sequence fragments (e.g., 10 and 100 kb) derived from a wide range of species were clustered (self-organized) according to species without information regarding species (Abe et al. 2005 and 2006).

By constructing pentanucleotide BLSOMs for all 20-, 50-, and 100-kb sequences derived from the human genome, this study examined the usefulness of BLSOM in unveiling sequence characteristics hidden within a single genome and revealed several specific zones on the BLSOM, which were composed primarily of sequences derived from pericentric regions. Importantly, in sequences belonging to the specific zones, various transcription-factor-binding (TFB) consensus motifs were characteristically enriched.

## Materials and methods

### BLSOM

Multidimensional vectorial data can be used to represent complex data numerically. When high-dimensional data (e.g., 1,024 dimensions for the pentanucleotide composition) are analyzed through linear projection onto a two-dimensional map such as done in the principal component analysis (PCA), only a minor portion of characteristics of the multidimensional data can be extracted in most cases: a low cumulative contribution rate of the first two dimensions in PCA. SOM is a clustering method to solve this issue and an unsupervised algorithm that projects non-linearly the high-dimensional data onto a two-dimensional plane (Kohonen et al. 1996). We modified the conventional SOM for genome informatics to make the learning process and resulting map independent of the order of data input, on the basis of BLSOM (Kanaya et al. 2001). In the original Kohonen's SOM the initial vectorial data were set by random values (Kohonen et al. 1996), but in the BLSOM the initial vectors were initialized by PCA (Kanaya et al. 2001). BLSOM for oligonucleotide composition was conducted as described previously (Abe et al. 2003). The clustering ability of BLSOM for four vertebrate genomes and the basal features of BLSOM patterns are presented in Supplementary Figs. S1 and S2.

Visualization of diagnostic oligonucleotides for the category separation was conducted as described previously (Abe et al. 2005). BLSOM program was obtained from UNTROD Inc. (y_wada@nagahama-i-bio.ac.jp). Distances of weight vectors between neighboring lattice points on BLSOM can be visualized as black levels with a U-matrix method (Ultsch 1993), and this provides information regarding similarity of oligonucleotide composition in local areas on BLSOM.

### Human genome sequence and generation of random sequences

Genome sequence of *Homo sapiens* (GRCh37) was obtained from NCBI ftp site (http://www.ncbi.nlm.nih.gov/genomes/). For each 50-kb sequence derived from the human genome, a random sequence that had nearly the same oligonucleotide composition as the respective sequence was generated by the Markov chain model (Abe et al. 2009).

## Results

### BLSOM for human genomic sequences

To study the usefulness of oligonucleotide-BLSOMs in examining intra-genomic difference, we analyzed 20-, 50- and 100-kb sequences derived from the human genome by using BLSOM for pentanucleotide composition (abbreviated as Penta-BLSOM in Fig. 1a for 50 kb and in Supplementary Fig. S3A for 20 and 100 kb); the reason why the Penta-BLSOM was chosen

was explained in Fig. S1 and discussed later. Lattice points containing sequences from a single chromosome were indicated in a color specifying the chromosome and those containing sequences from multiple chromosomes were indicated in black. Most lattice points were presented in black, showing that the BLSOMs did not separate most human sequences according to chromosome. However, several characteristic zones with colored or white lattice points were observed and designated as specific zones (Sz); similar observations were found on 20- and 100-kb BLSOMs (Fig. S3A). Lattice points with no genomic sequence after the BLSOM calculation were left white, and our previous study (Abe et al. 2003) showed that lattice points containing sequences with the oligonucleotide compositions very distinct from others were often surrounded by the white lattice points. Because colored lattice points in Sz were often surrounded by white lattice points, the Sz sequences were thought to have peculiar pentanucleotide compositions distinct from most other human sequences.

In DNA databases, only one strand of each pair of complementary sequences is registered. Our previous analysis of a wide variety of species (Abe et al. 2003) revealed that sequences from a single genome were often split vertically into two territories according to the transcriptional direction of the genes present in the fragment. To study general characteristics of genomic sequences, differences in oligonucleotide composition between two complementary strands are not important. To obtain a simplified BLSOM pattern, we constructed another BLSOM in which the frequencies of a pair of complementary pentanucleotides (e.g., AAAAC and GTTTT) in each fragment were summed. This BLSOM (DegePenta-BLSOM in Fig. 1b) yielded a simpler pattern than that of Penta-BLSOM because the mirror-symmetric split in the vertical direction disappeared.

## Addition of computer-generated random sequences

To understand the biological significance of the Sz sequences, it is useful to know their chromosomal locations. Therefore, we attempted to select Sz sequences, but it was rather difficult to unambiguously define Sz borders. To overcome this problem, we added computer-generated random sequences as described by Abe et al. (2009). To obtain the random sequences, we generated 50-kb random sequences with nearly the same mono-, di- , tri-, and tetranucleotide compositions as individual 50-kb human sequences and designated the random-sequence set as Mono-, Di-, Tri-, and TetraRan. Then, we constructed Penta- and DegePenta-BLSOMs with 50-kb human sequences plus one of the four random-sequence sets. After addition of Mono- and DiRan, real and random sequences were clearly separated from each other, and Sz sequences were located within the human territory (Fig. S3B); lattice points with only random sequences or with no sequences were left white. In contrast, the addition of TetraRan sequences formed specific zones composed of colored lattice points outside the major human territory and thus within TetraRan territories (white zones) on the Penta- and DegePenta-BLSOMs (Fig. 1c, d). The DegePenta-BLSOM yielded a much simpler pattern than the Penta-BLSOM, and five clear specific zones (Sz1-5) were found within the TetraRan territory in the DegePenta-BLSOM, and the Sz sequences thus unambiguously defined were used in the following analyses.

## Chromosomal locations of Sz sequences

To clarify characteristics of Sz sequences, we investigated their chromosomal locations, and Fig. 2 plotted the numbers of 50-kb Sz sequences per 500 kb on ten chromosomes with colored symbols distinguishing Sz. Edges of the centromeric heterochromatin region "band p11 (or p11.1)–q11 (or q11.1)", which was designated here as the centromeric band region, obtained from the UCSC Genome Browser (http://genome.hmgc.mcw.edu/) were marked with two brown arrows. On these and almost all human chromosomes, with the clear exception of chrY, the highest peak for one or a few types of Sz were located not only within but also near the centromeric band region, as far as judged on the UCSC Genome Browser. Although precise assignment of centromeric bands at a sequence level should be difficult in principle, we concluded that the Sz sequences located within and near the centromeric region and thus in pericentric regions.

The Sz that produced the highest peak in pericentric regions differed among chromosomes, and the less evident enrichment was observed in restricted regions far away from the pericentric regions on some chromosomes; e.g., the subtelomeric region in chr11. In the case of chrY, the Sz-sequence clusters were observed primarily outside the pericentric regions. This exceptional case may relate to the fact that a large portion of chrY consists of repetitive sequences to form heterochromatin.

## Oligonucleotides diagnostic for Sz

The notable clustering of Sz sequences in pericentric regions may provide novel knowledge concerning their functions because the clustering was observed on almost all chromosomes with a clear exception of chrY. Thus, we searched for pentanucleotides specifically enriched in Sz sequences, using a BLSOM ability to visualize diagnostic oligonucleotides for category-specific clustering (Abe et al. 2005); refer to examples listed in Figs. S1C and S2B for the species-specific clustering. In Fig. 3B, after calculating the expected frequency of each pentanucleotide from the mononucleotide composition at each lattice point, the observed/expected ratio for the pentanucleotide was indicated as follows: red (overrepresented), blue (underrepresented), and white (moderately represented). This observed/expected ratio was shown to be useful in unveiling genome signatures because the oligonucleotide composition could be analyzed independently of a simplex effect reflecting the mononucleotide composition of genomic sequences (Abe et al. 2006).

The overrepresentation of a certain oligonucleotide only in a restricted genomic portion should yield useful information for understating the biological significance of the sequence, especially when a biological function of the respective oligonucleotide is known. During the search on this assumption, we unexpectedly found that various pentanucleotides corresponded to TFB-consensus motifs or their closely related sequences were evidently enriched in the Sz sequences when compared to other 50-kb human sequences; five examples of TFB-consensus motifs were listed in Fig. 3B and other TFB pentanucleotides enriched specifically in Sz were listed in Fig. S4. For example, the TFB-consensus motif AGATA/TATCT is specifically enriched (red) in Sz2 but underrepresented (blue) in the major portion of the human territory and evidently underrepresented (dark blue) in Sz3. The finding that these TFB-consensus motifs were evidently overrepresented specifically in Sz was rather unexpected, because a major portion of Sz sequences were located in pericentric regions, which are rather poor in protein-coding genes.

## Distribution of TFB-consensus pentanucleotides on each chromosome

When we searched matrices of TFBSs through JASPAR Core Vertebrata (http://jaspar.cgb.ki.se/), approximately 130 matrices were found for vertebrate genomes. While a major portion of consensus cores of the TFBSs were longer than pentanucleotides, we here focused on the case where the consensus core was a contiguous pentanucleotide, as explained in Discussion. By searching also through TRANSFAC Public (http://www.gene-regulation.com/pub/databases.html), a total of ten pentanucleotides were obtained as human TFB-consensus motif pentanucleotides (abbreviated as TFB pentanucleotides) and listed in the first row of Table 1. Because relation to transcriptional direction was not concerned in this study, a pair of complementary pentanucleotides was assumed to be one and the same, and therefore, 20 pentanucleotides were concerned. Figure 4 plotted the occurrence of each TFB pentanucleotide per 100 kb for individual chromosomes. Interestingly, the highest occurrence was found in pericentric regions not only on chromosomes listed in Fig. 4 but also on almost all chromosomes, although the pentanucleotides that produced the highest peak differed among chromosomes. TFB pentanucleotides were also enriched in subtelomeric regions and/or internal restricted regions, but their occurrences were primarily lower than those found in pericentric regions. It should be noted here that in Fig. 4 and in subsequent analyses, AAATA/TATTT was not included because this TFB pentanucleotide was enriched neither in Sz nor in pericentric regions and thus was out of scope of this study. The finding that nine out of ten human TFB pentanucleotides were evidently enriched in pericentric regions indicated that their evident enrichment in pericentric regions should have biological significances.

"Homo sapiens high coverage assembly GRCh37" reported by the Genome Reference Consortium (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/) does not include highly repetitive centromeric sequences (e.g., alphoid repeats), because unique sequences required for sequence assembly are absent there. Sizes of unsequenced regions (blank regions marked by brown arrows in Fig. 4) differed between chromosomes primarily due to the varying difficulty of contig formation. Thus, only positive results (i.e. the presence of sequences), not negative results (i.e. the absence of sequences), must be considered. It should also be mentioned that the evident enrichment of TFB pentanucleotides was observed both within and near the centromeric band region.

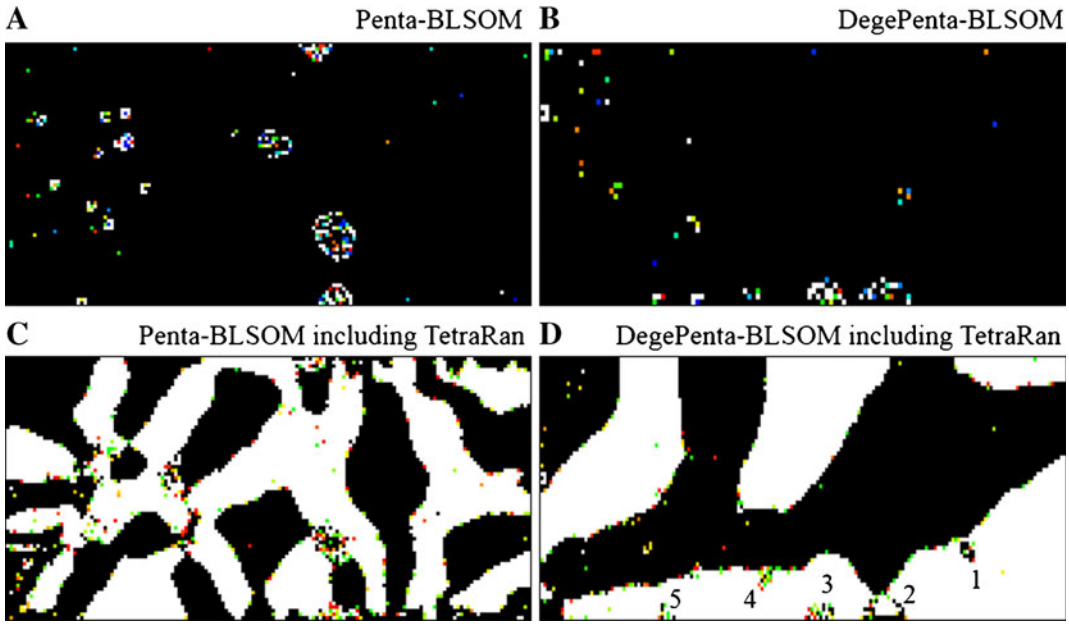


**Fig. 1** Penta- and DegePanta-BLSOMs for 50-kb sequences derived from the human genome (**a**) and for those including TetraRan sequences (**b**). When we construct BLSOM for a certain window size (e.g., 50 kb), we can set a certain sliding step. While the clustering patterns obtained with and without the sliding step resembled to each other in this study, the pattern listed was the 50-kb BLSOM with a 25-kb step. Examples without a sliding step were listed in Fig. S3A. The map size was chosen as an average number of data was ten at one lattice point. Lattice points containing sequences from multiple chromosomes are indicated in black and those containing sequences from a single chromosome are indicated in color: chr1 (■), chr2 (■), chr3 (■), chr4 (■), chr5 (■), chr6 (■), chr7 (■), chr8 (■), chr9 (■), chr10 (■), chr11 (■), chr12 (■), chr13 (■), chr14 (■), chr15 (■), chr16 (■), chr17 (■), chr18 (■), chr19 (■), chr20 (■), chr21 (■), chr22 (■), chrX (■), and chrY (■). Lattice points that contain only random sequences or no sequences are indicated in *white blank*. While difference in color specifying 24 chromosomes is not clear, this is of no importance here because colored lattice points shows only that the lattice points contain sequences derived from a single chromosome

## BLSOM for 1-kb sequences derived from Sz

Analyzing the pentanucleotide composition with 50-kb BLSOMs (Fig. 1) and the occurrence of each TFB pentanucleotide per 100 kb (Fig. 4) was suitable for ascertaining a chromosome-level distribution of the pentanucleotide, but analyzing much shorter sequences was required to understand characteristics of TFB-motif-rich sequences at a primary sequence level; e.g., whether the TFB pentanucleotides were arranged dispersively or locally within each 50-kb sequence. We thus constructed a DegePenta-BLSOM for 1-kb sequences derived only from 50-kb Sz sequences. In Fig. 5A, lattice points containing 1-kb sequences derived from the same Sz were colored to specify the Sz and those containing sequences from multiple Sz were indicated in black. The large extended black region, which contained sequences derived from multiple Sz and designated here as "nonspecific zone", was surrounded by colored territories containing sequences derived from one Sz. After this segmentation into 1 kb, Sz-core sequences were clearly separated from nonspecific sequences, showing the local clustering of TFB pentanucleotides within each 50-kb sequence. While human centromeric regions are known to contain a large amount of alpha (alphoid), beta and gamma repeats, the present Sz-core sequences differed from these repeats and also from Alu; the 1-kb sequences derived from LINE/L1 in 50-kb Sz sequences were located in the aforementioned nonspecific zone but not in Sz-core territories (data not shown).

## Diagnostic pentanucleotides for Sz-specific cores

Similarity in oligonucleotide composition between neighboring lattice points in BLSOM (and thus between sequences belonging to neighboring lattice points) can be visualized using a U-matrix (Ultsch

1993) with a level of blackness (Fig. 5B); the areas containing sequences with similar or distinct oligonucleotide composition were recognized as low or high black level. On the U-matrix, borders between different Sz-core territories were accurately visualized as heavy black lines. In the large Sz2- and 3-core territories (pink and light brown in Fig. 5A), thin black lines were observed, showing lesser but significant differences in pentanucleotide composition within a single Sz-core territory. In Fig. 5Ci–iv, diagnostic pentanucleotides representing Sz cores were examined by focusing on TFB pentanucleotides. Combinatorial high occurrences of several (not a single) TFB pentanucleotides were observed for individual Sz cores. Importantly, preference of the TFB pentanucleotides in individual SZ was primarily similar between 50- and 1-kb BLSOMs, showing that 1-kb Sz-core sequences were the good represents for the 50-kb Sz sequences.

### Occurrences of TFB pentanucleotides in 1-kb sequences plotted on each chromosome

To compare TFB pentanucleotide occurrence in pericentric regions with other chromosomal regions, we simply counted occurrences of individual TFB pentanucleotides in all 1-kb sequences derived from all chromosomes (approximately 3million sequences). When listing the distribution patterns for all nine TFB pentanucleotides simultaneously, it was difficult to understand characteristics of their distributions in one figure, except for sex chromosomes. In the case of sex chromosomes, characteristic occurrences could be easily detected even for the 1-kb sequences, because eight TFB pentanucleotides were enriched to the highest degree in approximately 2.5 Mb of the short-arm tips of chrY and X (Figs. 6 and S5, respectively). Distribution patterns in their tips turned out to be the same

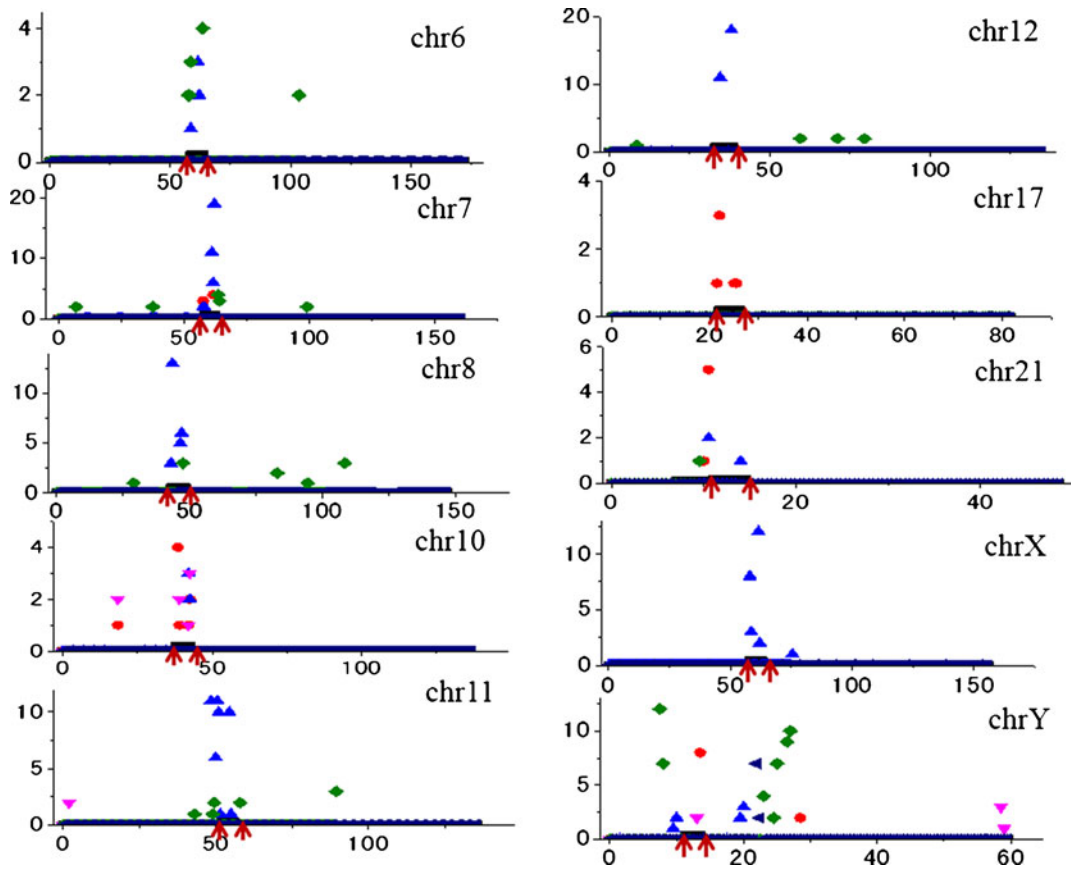


**Fig. 2** Distribution of Sz sequences on individual chromosomes. Numbers of sZ sequences per 500-kb were plotted with colored symbols distinguishing sZ: Sz1 (●), Sz2 (▲), Sz3 (▼), Sz4 (◆), and Sz5 (◂). Centromeric and pericentric heterochromatin regions were marked with *horizontal bars* just above the *X*-axis and also with *two brown arrows*

between the two chromosomes, since these regions corresponded to pseudoautosomal regions (PARs) with the same sequence, which play crucial roles in pairing between two sex chromosomes during meiosis. The evident enrichment of many TFB motifs in PARs is of particular interest when considering the biological significance of TFB-motif-rich sequences.

For other chromosomes, the occurrence of one or two TFB pentanucleotides was presented in order to show their characteristics distribution clearly (Figs. 6 and S5). AATCA/TGATT and/or GCCAA/TTGGC were evidently enriched in pericentric regions and produced the highest peak composed of many clustered dots, indicating the large-scale clustering of TFB pentanucleotide-rich 1-kb sequences. TFB pentanucleotides were also enriched in subtelomeric regions, primarily producing the second highest peak composed of multiple dots. Orphan high dots in internal positions corresponded to orphan 1-kb sequences rich in the respective pentanucleotides. Because the cluster size for PAR, in the subtelomeric and internal regions was smaller than for the pericentric regions, the enrichment in the former regions was not significant in the 100-kb analysis in Fig. 4.

Occurrences of TFB pentanucleotides in 1-kb sequences were summarized for all chromosomes in Table 1. When the highest occurrence of a certain TFB pentanucleotide was found in a pericentric or subtelomeric region on a certain chromosome, "CEN" or "TEL" was listed; when the second or third highest occurrence was found, "cen" or "tel" was listed. For sex chromosomes, when the highest (the second or third highest) occurrence was found in PAR, "PAR" ("par") was listed. The evident enrichment of TFB motifs in pericentric and/or subtelomeric regions was observed for 21 out of 24 chromosomes, although the enriched pentanucleotides differed among chromosomes.

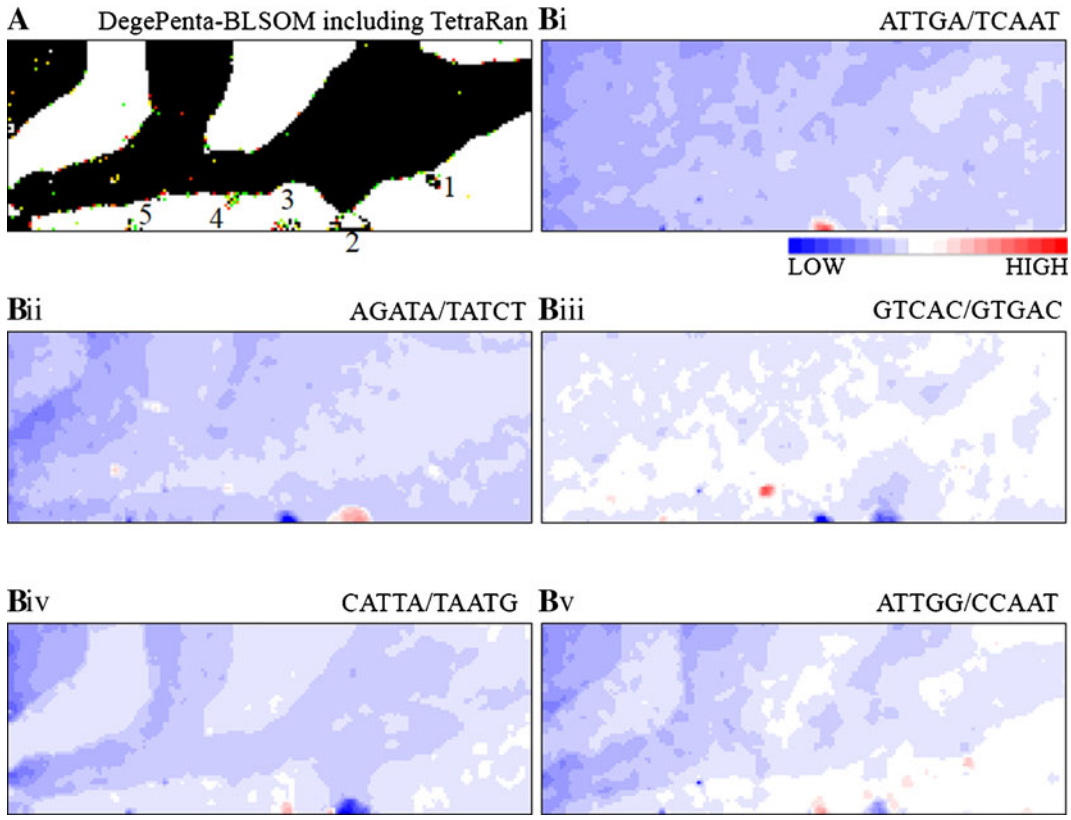


**Fig. 3** Pentanucleotides enriched in Sz. (*A*) DegePenta-BLSOM listed in Fig. 1d. (*Bi–v*) TFB pentanucleotides specifically enriched in Sz. After calculating the expected frequency of each pentanucleotide from the mononucleotide composition at each lattice point, the observed/expected ratio for the pentanucleotide was indicated in color presented under the *Bi* panel

**Table 1** High occurrence of TFB pentanucleotide in 1-kb sequence

| Chr | AATCA/ TGATT | GCCAA/ TTGGC | ACCAC/ GTGGT | TATCA/ TGATA | ATTGG/ CCAAT | AATCT/ AGATT | CTATC/ GATAG | CTTCC GGAAG/ | AAATA/ TATTT | AGATA/ TATCT |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | CEN, tel | | tel | | | | tel | | | tel |
| 2 | CEN, tel | CEN, tel | | tel | | | tel | tel | | |
| 3 | | | tel | tel | | | tel | TEL | | |
| 4 | tel | | TEL | | | | | | | |
| 5 | | | | cen | | | | | | |
| 6 | | | tel | | | | | TEL | | |
| 7 | cen, TEL | | TEL | | | | | TEL | | |
| 8 | | | TEL | | | | | TEL | | tel |
| 9 | | | TEL | | | | | | | |
| 10 | CEN | | | | | | tel | | | |
| 11 | | | TEL | | | | tel | tel | | |
| 12 | TEL | | | | | tel | TEL | TEL | | TEL |
| 13 | | | cen | | | | | | | |
| 14 | | | | | | | | | | |
| 15 | | CEN | | | | | | | | |
| 16 | CEN | CEN, tel | | | | | | | | |
| 17 | | | | | | | | TEL | | |
| 18 | | | tel | | | | | | | tel |
| 19 | | | | | | | | | | |
| 20 | | | TEL | | | | | CEN | | |
| 21 | | | cen | | | | cen | | | |
| 22 | CEN | CEN | | | | | | | | |
| X | | par | | par | PAR | par | PAR | PAR | PAR | PAR |
| Y | CEN, par | PAR | | PAR | PAR | PAR, cen | PAR | PAR | PAR | PAR |

## Examples of actual sequences evidently rich in TFB pentanucleotides

If pentanucleotides are randomly used, the average occurrence of a certain pentanucleotide per 1 kb should be approximately one, since the number of pentanucleotide types is 1,024 ($=4^5$). Taking this into account, we examined actual 1-kb sequences rich in TFB pentanucleotides. A main purpose of the present study is to investigate a possible function of the clustered occurrence of TFB motifs that differs from the conventional transcriptional regulation, and therefore, we focused on sequences within the centromeric region that were located far away from the neighboring band. In the case of chr22, a large portion of q11 was sequenced and two examples of 22q11 sequences were selected because neither protein nor RNA coding gene was observed around these sequences. In the case of the first sequence "chr22;17048001–17049000" (Fig. S6), GCCAA was used 67 times, showing approximately 70 times more occurrence than the random occurrence; more than one third of the 1-kb sequence was occupied by this TFB pentanucleotide (67×5=335 nt). The Ensembl-Blat analysis (http://asia.ensembl.org/Homo_sapiens/blastview) showed that homologous sequences existed also in pericentric regions of chr2, 15 and 16 (Fig. S6). In the case of another 22q11 sequence "chr22: 16,855,001–16,856,000" (Fig. S7), AATCA was used 48 times and both TATCA and ATTGG were used three times. Homologous sequences to this 1-kb sequence exist both in pericentric regions of chr2, 7, 10, 16, and Y and in internal regions of chr1 and 16.

## Bindings of TFs in pericentric regions

Because notable clustering of TFB pentanucleotides was found in pericentric regions, an important question

of interest is to know whether TFs actually bind to the pericentric Sz sequences, especially in the regions absent of protein-coding genes. ChIP-Seq Data for TFs (abbreviated as TfChIP) and Histone H3K27Ac (often found near active regulatory elements; abbreviated as AcChIP) obtained by the ENCODE Project, were available from the UCSC Genome Browser. We next examined whether these marks were observed in Sz sequences in the centromeric regions where protein-coding genes were very rare. This reduced significantly the number of Sz sequences for the analysis because a large portion of Sz sequences were located near but outside of the centromeric regions. As easily expected, occurrence of these marks in the regions was evidently lower than in euchromatic regions.

In all of the 12 Sz4 sequences, TfChIP and/or clear AcChIP were observed; 7 had both TfChIP and AcChIP (one example is listed in Fig. 7a); 4 had only TfChIP; 1 had only AcChIP (Fig. S8A). Six noncoding RNAs (but no protein-coding gene) were observed; TfChIP in individual Sz4 sequences primarily contained more than ten different types of TFBSs (Fig. 7a) and often included sites for pol2 and/or TBP. These findings indicate that at least a portion of TFBSs in the centromeric Sz4 sequences may relate to transcriptional regulation of noncoding RNAs.

Other Sz sequences, however, had different characteristics. Out of 98 Sz2 sequences, 39 had only TfChIP (Fig. 7b); 10 had only AcChIP (Fig. S9A); 9 had both TfChIP and AcChIP; one noncoding and one protein-coding gene were observed. After removing these two sequences having genes, 56 (out of 98) sequences had TfChIP and/or AcChIP, and TfChIP in these sequences was primarily composed of a few types of TFBSs (Fig. 7b) and did not contain pol2, indicating that a major portion of these TfChIP and/or AcChIP may not relate to the conventional transcriptional regulation.

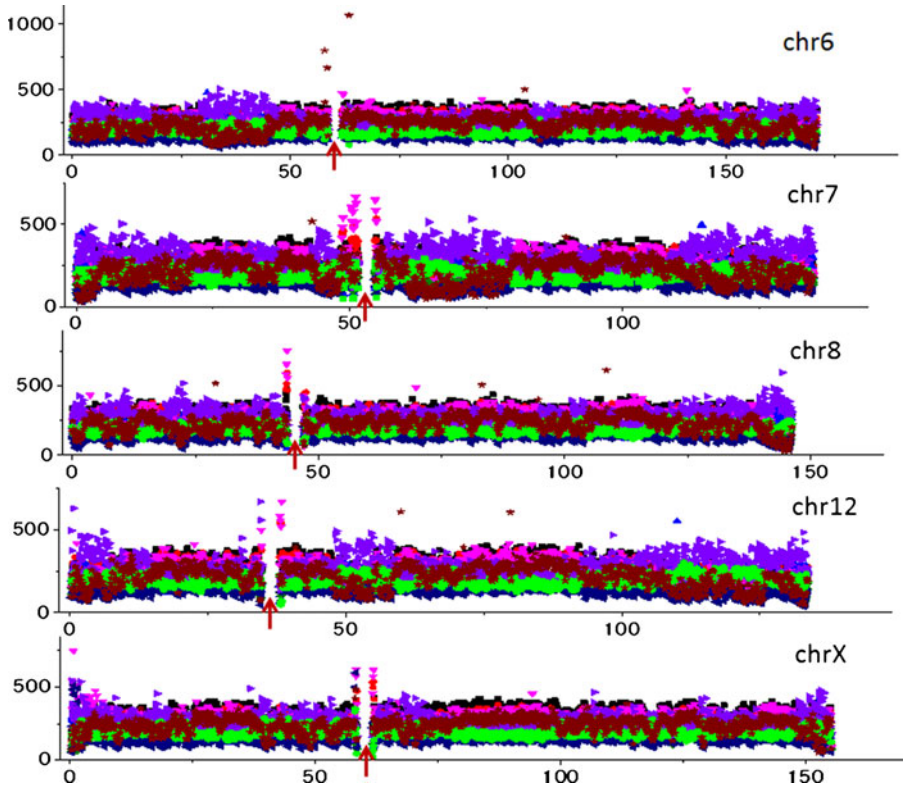


**Fig. 4** Distribution of TFB pentanucleotides on individual chromosomes. Numbers of TFB pentanucleotides per 100 kb were plotted with *colored symbols* distinguishing TFB pentanucleotide: AATCA/TGATT (■), AATCT/AGATT (●), ACCAC/GTGGT (▲), AGATA/TATCT (▼), ATTGG/CCAAT (◆), CTATC/GATAG (◂), CTTCC/GGAAG (▸), GCCAA/TTGGC (■), and TATCA/TGATA (★)

A major portion of the centromeric Sz1 and Sz3 sequences had TfChIP and/or AcChIP but had neither protein-coding nor noncoding gene except for one noncoding gene in one Sz1 sequence; for details, see the lend for Fig. S9. As observed for Sz2 sequences, the centromeric Sz1- and -3 sequences that had no gene were primarily composed of one or a few types of TFBSs while the one type often repeated several times within one Sz sequence. This contrasted to the TfChIP data in euchromatic regions harboring genes and indicated possible functions other than the conventional transcriptional regulation. Detailed classification of TfChIP data according to the number and type of TFBSs appears to be important and is planned as a separate study.

Graur et al. (2013) recently reported critical remarks about the ENCODE data. Because centromeric regions have highly repetitive sequences, TfChIP data might be artifactually enriched during multiple alignment and mapping of the sequences. To test this possibility, we examined the mapability (i.e., uniqueness) of TfChIP sequences, which was available from the UCSC Genome Browser. The sizes of individual TfChIP in centromeric regions were primarily a few or several hundred base pairs, and a significant portion of each sequence was covered with the highest mapability (1.0). This indicated that the TfChIP that were difficult to be mapped appeared not to be included in the published data. While we cannot exclude the possibility mentioned by Graur et al. (2013) that TFBs may occur even without any functions, we will discuss possible biological significances of TFBs in pericentric regions.

## Discussion

### Biological significance of notable TFB-motif clustering

We first discuss a possible biological significance of clustering of TFB-motif oligonucleotides, as well as of TFBs, in pericentric regions. A well-known function of pericentric regions is the formation of condensed heterochromatin in chromocenters, which supports the association of pericentric DNAs of homologous and nonhomologous chromosomes (Maison et al. 2002; Maison and Almouzni 2004; Probst et al. 2009; Probst and Almouzni 2011). A group of chromosomes forming individual chromocenters depends on cell types and stages, i.e., the size and number of chromocenters differ among organisms and among tissues of the same organism. Furthermore, large chromocenters are formed by grouped regions of pericentric, nucleolar, and telomeric heterochromatin as well as sex chromosomes. All TFB-rich sequences revealed by the 1-kb analysis in Figs. 6 and S5 (pericentric, subtelomeric,

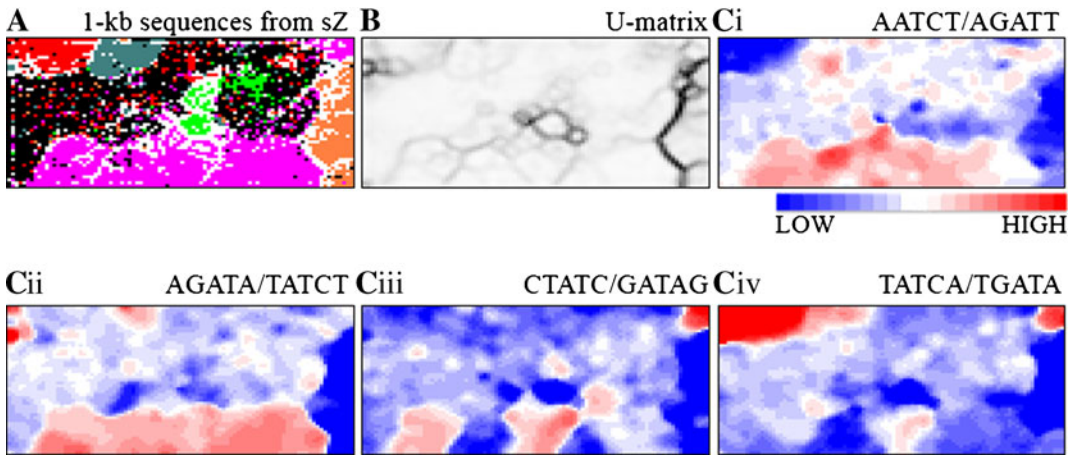


**Fig. 5** DegePanta-BLSOM for 1-kb sequences derived from Sz. (*A*) Lattice points containing sequences from multiple Sz are indicated in *black*, and those containing sequences from a single sZ are indicated in *color*: Sz1 (■), Sz2 (■), Sz3 (■), Sz4 (■), and Sz5 (■). The Sz2-specific sequences (*pink*) formed a major extended territory at the *bottom* and a minor territory at the *top left*. This split in the Sz2 territory may relate to the internal segmentation observed within Sz2 in Fig. 3a. (*B*) U-matrix. (*Ci–iv*) TFB pentanucleotides specifically enriched in Sz-specific core territory. The observed/expected ratio for each TFB pentanucleotide was calculated as described in Fig. 3b and indicated in *color* presented under the Civ panel

and PARs) coincided with the sequences involved in the chromocenter formation.

Pericentric heterochromatin was once thought to be stable in composition and transcriptionally inert, but has been shown to be surprisingly dynamic (Maison et al. 2002; Maison and Almouzni 2004; Probst et al. 2009; Probst et al. 2010; Probst and Almouzni 2011). Mouse centromere-derived double-stranded transcripts appear to be involved in establishing the heterochromatin structure (Maison et al. 2002), and Dicer-related RNA interference machinery is involved in the formation of the centromeric heterochromatin in higher vertebrate cells (Fukagawa et al. 2004). A strand-specific burst in transcription of mouse pericentric satellites is required for chromocenter formation during early mouse development (Probst et al. 2010), and mouse TLF (TBP-like factor) is required for chromocenter formation in haploid round spermatids (Martianov et al. 2002). Long nuclear noncoding RNA is transcribed from the periphery of pericentric heterochromatin (Maison et al. 2011) and centromere RNA is a key component for the assembly of nucleoproteins at the nucleolus and centromere (Du et al. 2010; Wong et al. 2011). Notable clustering of TFB motifs in pericentric regions may relate, at least in part, to these transcriptions in the regions. For the extremely complex and dynamic organization of mammalian nuclei, a wide variety of interactions of RNAs with DNAs and/or proteins should be dynamically intertwined. Transcripts in pericentric regions may engage in chromocenter formation through RNA-protein, RNA–RNA and RNA–DNA interaction. TFB-motif enrichment in pericentric regions may play roles in the RNA-mediated nuclear organization, and actually the centromeric Sz4 sequences often had noncoding RNA genes.

We also focus attention on the current understanding that TFs and TFBSs have various functions other than transcriptional regulations (Probst et al. 2009; Probst and Almouzni 2011); various TFBSs in genomic regions most likely unrelated to transcription have been experimentally proven (MacQuarrie et al. 2011). The TFB-motif-rich sequences located in the protein

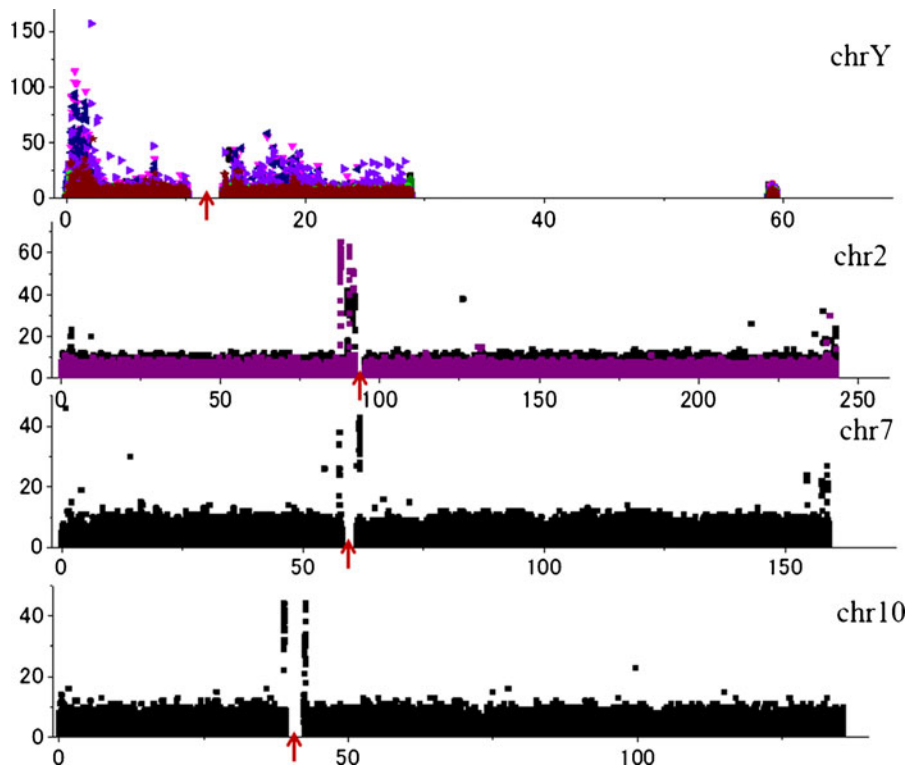


**Fig. 6** Distribution of TFB pentanucleotides on individual chromosomes. Numbers of TFB pentanucleotides per 1 kb were plotted with colored symbols distinguishing TFB pentanucleotides: for chr Y, as described in Fig. 4: for other chromosomes, AATCA/TGATT (■) and GCCAA/TTGGC (■)

gene-poor centromeric regions may have novel functions other than the conventional transcriptional regulation. While neither protein-coding nor noncoding transcript was observed in most of centromeric Sz1-3 sequences, a significant portion of the sequences had TfChIP and/or AcChIP. In formation of highly complex nuclear organizations, various types of protein–DNA and protein–protein interactions should play crucial roles. TFs can bind a wide range of proteins including other TFs. If a certain combination of TFs binds to the TFB-motif-rich sequences in pericentric regions, sequences on different chromosomes can gather together through protein–protein interactions. When considering molecular mechanisms to achieve cell-type and -stage-dependent chromocenter formation, the chromosome-dependent combinatorial clustering of TFB motifs is of particular interest because the content of individual TFs is regulated in a way specific to the cell types and stages.

The interaction of centromeric/pericentric, telomeric/subtelomeric and pseudoautosomal regions with each other and with various internal genomic regions should play crucial roles in the global organization of chromosomal DNAs within interphase nuclei; the pericentric regions may provide the primary largest-scale hubs, and the telomeric/subtelomeric regions and PARs may provide the secondary large-scale hubs, possibly depending on the size of TFB motif clusters. Much smaller-scale TFB-motif-rich regions in internal regions (e.g., orphan-type enrichment in Figs. 6 and S5) may interact with the primary and/or secondary large-scale hubs, through protein–protein

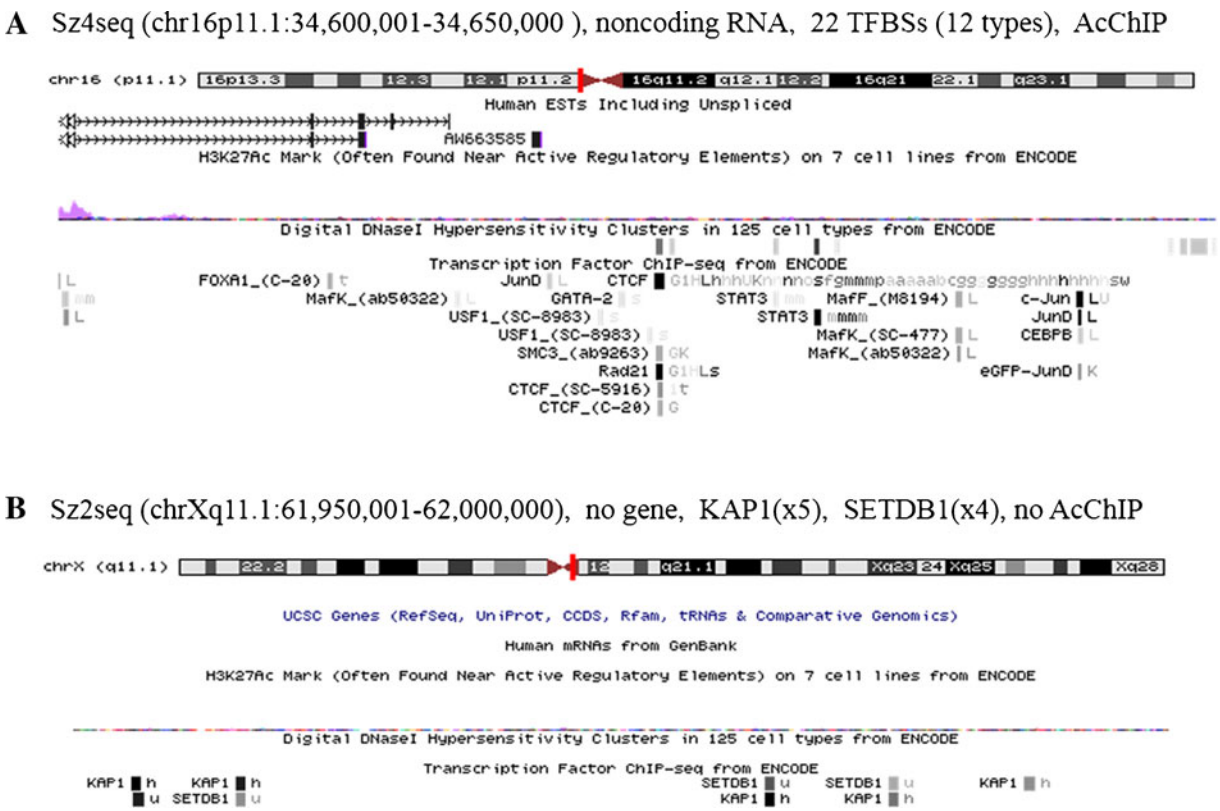


**Fig. 7** TfChIP and AcChIP data in 50-kb Sz sequences located within the centromeric band region. A *narrow brown bar* in the centromeric region (marked with brown in the chromosome ideogram) showed the position of the Sz sequence. **a** An example of Sz4 sequences. Noncoding RNA gene (*arrowed dash line*), AcChIP (*pale violet chevron mark*) and TfChIP (*small vertical bar*) were listed according to the UCSC Genome Browser. **b** An example of Sz2 sequences. Only TfChIP data were observed; KAP1 and SETDB1 were repeated 5 and 4 times, respectively. The figures were downloaded from the UCSC Genome Browser, but some data were not clear in the presented figures, and therefore, the nucleotide position and the name and number of TFs observed in the Sz sequence were additionally listed in *bigger letters* in each figure

interaction that is supported by various TFs and TF-associated proteins. Pairing of two PARs may depend on similar mechanisms because TFB motifs are evidently enriched there.

Our examination of primary sequences derived from the same Sz core but belonging to different chromosomes showed that the 1-kb sequences often had distinct primary sequences when analyzed with a dot-matrix method (data not shown). For formation of a certain supramolecular complex containing a combinatorial set of TFs, combinatorial enrichment of a certain set of TFB motifs, rather than the primary sequence, may be important.

## Oligonucleotides other than pentanucleotides analyzed here

This study focused mainly on TFB-consensus pentanucleotides. In the simple counting of occurrences of TFB-motif pentanucleotides per 1 kb, however, we occasionally found almost the same occurrence for two pentanucleotides, which turned out to be the components of one TFB hexa- or heptanucleotide. This shows that analysis of pentanucleotides can provide, at least in part, information concerning the occurrence of TFB oligonucleotides longer than pentanucleotides. Therefore, we next discuss the cases in which the pentanucleotide was not registered as a TFB-consensus motif but enriched specifically in Sz (Fig. S4B). Such pentanucleotides often corresponded to a component of longer TFB-consensus motifs of vertebrates; for details, see the legend for Fig. S4.

To examine the generality of the present findings, BLSOM for oligonucleotides longer than pentanucleotides will become important because the number of TFB-motif hexa- and heptanucleotides exceeds the number of TFB pentanucleotides. In this study, we used a high-performance PC, not a supercomputer, for BLSOM calculation. The analysis of much higher dimensional data such as the 16,384-dimensional data for heptanucleotides could not be conducted using a PC. The BLSOM algorism, however, is suitable for large-scale parallel computing with supercomputers, and we recently had an opportunity to use a high-performance supercomputer for a test trial to construct DegeHexa- and DegeHepta-BLSOMs for all 50-kb sequences derived from the human genome. Again, at least six Sz were observed on both BLSOMs, and Sz1-5 sequences observed on the DegePenta-BLSOM, most of which were derived from pericentric regions, were located primarily in Sz on the DegeHexa- and DegeHepta-BLSOMs (our unpublished data). This indicated that hexa- and heptanucleotide compositions in pericentric regions also differed from a major portion of the human genome. Since the study of hexa- and heptanucleotide compositions requires much more analyses than that for pentanucleotide composition, a separate study is planned.

**Acknowledgments** This work was supported by Grant-in-Aid for Scientific Research (C) and for Young Scientists (B) from the Ministry of Education, Culture, Sports, Science and Technology of Japan and by the Grant-in-Aid for JSPS Fellows (JSPS KAKENHI Number 24•9979). We thank Hewlett-Packard Japan, Ltd. and SCSK Corp. Japan for giving us an opportunity to use HP ProLiant DL980 G7 as a test trial.



## References

Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T (2003) Informatics for unveiling hidden genome signatures. Genome Res 13:693–702

Abe T, Sugawara H, Kinouchi M, Kanaya S, Ikemura T (2005) Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. DNA Res 12:281–290

Abe T, Sugawara H, Kinouchi M, Kanaya S, Ikemura T (2006) A large-scale Self-organizing map (SOM) unveils sequence characteristics of a wide range of eukaryote genomes. Gene 365:27–34

Abe T, Wada K, Iwasaki Y, Ikemura T (2009) Novel bioinformatics for inter- and intraspecies comparison of genome signatures in plant genomes. Plant Biotech 26:469–477

Bernardi G (2004) Structural and evolutionary genomics: natural selection in genome evolution. Elsevier, New York

Bernardi G, Olofsson B, Filipski J et al (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958

Du Y, Topp CN, Dawe RK (2010) DNA binding of centromere protein C (CENPC) is stabilized by single-stranded RNA. PLoS Genet 6:e1000835

Fukagawa T, Nogami M, Yoshikawa M et al (2004) Dicer is essential for formation of the heterochromatin structure in vertebrate cells. Nat Cell Biol 6:784–791

Gentles AJ, Karlin S (2001) Genome-scale compositional comparisons in eukaryotes. Genome Res 11:540–546

Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E (2013) On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol 5:578–590

Ikemura T (1985) Codon usage and transfer RNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34

Ikemura T, Aota S (1998) Global variation in G+C content along vertebrate genome DNA: possible correlation with chromosome band structures. J Mol Biol 203:1–13

Kanaya S, Kinouchi M, Abe T et al (2001) Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the E. coli O157 genome. Gene 276:89–99

Karlin S, Campbell AM, Mrazek J (1998) Comparative DNA analysis across diverse genomes. Annu Rev Genet 32:185–225

Kohonen T, Oja E, Simula O, Visa A, Kangas J (1996) Engineering applications of the self-organizing map. Proc IEEE 84:1358–1384

MacQuarrie KL, Fong AP, Morse RH, Tapscott SJ (2011) Genome-wide transcription factor binding: beyond direct target regulation. Trends Genet 27:141–148

Maison C, Almouzni G (2004) HP1 and the dynamics of heterochromatin maintenance. Nat Rev Mol Cell Biol 5:296–304

Maison C, Bailly D, Peters AH et al (2002) Higher-orderstructure in pericentric heterochromatin involves a distinct pattern of histone modification and an RNA component. Nat Genet 30:329–334

Maison C, Bailly D, Roche D et al (2011) SUMOylation promotes de novo targeting of HP1α to pericentric heterochromatin. Nat Genet 43:220–227

Martianov I, Brancorsini S, Gansmuller A, Parvinen M, Davidson I, Sassone-Corsi P (2002) Distinct functions of TBP and TLF/TRF2 during spermatogenesis: requirement of TLF for heterochromatic chromocenter formation in haploid round spermatids. Development 129:945–955

Probst AV, Almouzni G (2011) Heterochromatin establishment in the context of genome-wide epigenetic reprogramming. Trends Genet 27:192–206

Probst AV, Dunleavy E, Almouzni G (2009) Epigenetic inheritance during the cell cycle. Nat Rev Mol Cell Biol 10:192–206

Probst AV, Okamoto I, Casanova M, Marjou FE, Baccon PL, Almouzni G (2010) A strand-specific burst in transcription of pericentric satellites is required for chromocenter formation and early mouse development. Dev Cell 19:625–638

Ultsch A (1993) Self organized feature maps for monitoring and knowledge acquisition of a chemical process. In Proc. ICANN'93, Int. Conf. on Artificial Neural Networks, edited by S Gielen, B Kappen. London: Springer: 864–867

Wong LH, Brettingham-Moore KH, Chan L et al (2011) Centromere RNA is a key component for the assembly of nucleoproteins at the nucleolus and centromere. Genome Res 17:1146–1160