Genes Genet. Syst. (2015) **90**, p. 43–53

CG-containing oligonucleotides and transcription factor-binding motifs are enriched in human pericentric regions

Yoshiko Wada[1,2], Yuki Iwasaki[1†], Takashi Abe[3], Kennosuke Wada[1], Ikuo Tooyama[2] and Toshimichi Ikemura[1*]

[1]*Department of Bioscience, Nagahama Institute of Bio-Science and Technology, Nagahama, Shiga 526-0829, Japan*
[2]*Shiga University of Medical Science, Molecular Neuroscience Research Center, Seta Tsukinowa-cho, Otsu City, Shiga 520-2192, Japan*
[3]*Institute of Science and Technology, Department of Information Engineering, Faculty of Engineering, Niigata University, Niigata 950-2181, Japan*

(Received 16 October 2014, accepted 20 November 2014)

Unsupervised data mining capable of extracting a wide range of information from big sequence data without prior knowledge or particular models is highly desirable in an era of big data accumulation for research on genes, genomes and genetic systems. By handling oligonucleotide compositions in genomic sequences as high-dimensional data, we have previously modified the conventional SOM (self-organizing map) for genome informatics and established BLSOM for oligonucleotide composition, which can analyze more than ten million sequences simultaneously and is thus suitable for big data analyses. Oligonucleotides often represent motif sequences responsible for sequence-specific binding of proteins such as transcription factors. The distribution of such functionally important oligonucleotides is probably biased in genomic sequences, and may differ among genomic regions. When constructing BLSOMs to analyze pentanucleotide composition in 50-kb sequences derived from the human genome in this study, we found that BLSOMs did not classify human sequences according to chromosome but revealed several specific zones, which are enriched for a class of CG-containing pentanucleotides; these zones are composed primarily of sequences derived from pericentric regions. The biological significance of enrichment of these pentanucleotides in pericentric regions is discussed in connection with cell type- and stage-dependent formation of the condensed heterochromatin in the chromocenter, which is formed through association of pericentric regions of multiple chromosomes.

Key words: CG dinucleotide, chromocenter, oligonucleotide composition, pericentric heterochromatin, self-organizing map

INTRODUCTION

Oligonucleotide composition varies significantly even among genomes with the same G+C% and represents a "genome signature" (Karlin et al., 1998). Oligonucleotide composition can also specify both inter- and intragenomic differences (Abe et al., 2006a; Iwasaki et al., 2013). The present study analyzes intragenomic, global characteristics of oligonucleotide composition in the human genome at the level of tens of kb or much longer. Oligonucleotides such as tetra-, penta- and hexanucleotides often represent motif sequences responsible for sequence-specific protein binding (e.g., transcription factor binding). Occurrences of such functionally important oligonucleotides should differ between regions within a single genome, and regions enriched with these specific oligonucleotides are thought to be distinguishable from the majority of genomic sequences.

A self-organizing map (SOM) is an unsupervised clustering method that can visualize high-dimensional complex data on one plane (Kohonen et al., 1996). For oligonucleotide compositions handled as high-dimensional data (e.g., 1024 dimensions for the pentanucleotide

Edited by Toshihiko Shiroishi
* Corresponding author. E-mail: t_ikemura@nagahama-i-bio.ac.jp
† Present Address: National Research Institute of Fisheries Science, 2-12-4 Fukuura, Kanazawa, Yokohama, Kanagawa 236-8648, Japan

composition), we have modified the conventional SOM to produce the batch-learning self-organizing map (BLSOM) (Kanaya et al., 2001; Abe et al., 2003), which, unlike the conventional SOM, is suitable for high-performance parallel computing and thus for big data analyses. BLSOM can analyze more than ten million sequences simultaneously (Abe et al., 2006b), and the BLSOM for oligonucleotide composition can cluster genomic sequence fragments (e.g., 1-kb sequences) from a wide range of microorganisms according to phylotype without information regarding species (Abe et al., 2005; Uehara et al., 2011; Nakao et al., 2013). BLSOM for oligonucleotide compositions in much longer genomic sequences (e.g., 50 and 100 kb) derived from a wide range of eukaryotes can also cluster the sequences according to species, by unveiling global genome characteristics with no prior knowledge other than oligonucleotide composition (Abe et al., 2006a; Iwasaki et al., 2014), and the present study has focused on global characteristics of oligonucleotide composition in human chromosomal DNAs.

In an era of big data accumulation of genomic sequences, unsupervised data mining, which can extract a wide range of information without prior knowledge or particular models, has become increasingly important, because it allows us to acquire the least expected knowledge from a massive number of genomic sequences. BLSOM can reveal and visualize oligonucleotides that contribute to the clustering of genomic sequences according to a category of interest, and thus provides an unsupervised data mining strategy. For example, when analyzing tetra- and pentanucleotide compositions in human genomic sequences of 20, 50 and 100 kb, we unexpectedly found that a wide variety of transcription factor-binding (TFB) motif oligonucleotides are enriched in pericentric heterochromatin regions although we know that there are few protein-coding genes in these regions (Iwasaki et al., 2013). Importantly, the enrichment of particular sets of TFB motif oligonucleotides in the pericentric region is chromosome-specific, and the TFB motif-enriched sequences differ from well-characterized repetitive sequences such as alphoid, beta- and gamma-satellite, Alu, and L1. TFs and TFB sequences (TFBSs) have various functions other than transcriptional regulation (Probst and Almouzni, 2011), and various TFBSs located in genomic regions most likely unrelated to transcription have been widely observed (MacQuarrie et al., 2011). In addition, TF-mediated looping interactions between genome portions that are remote from each other in the one-dimensional sequence, such as chromosome conformation capture (Sanyal et al., 2012), have attracted wide attention.

Since human pericentric heterochromatin regions are poor in protein-coding genes, the enrichment of TFBSs in pericentric regions may relate to novel functions of TFs other than conventional transcriptional regulation. A well-known function of mammalian pericentric regions is formation of condensed heterochromatin in chromocenters, which support the association of pericentric DNAs of homologous and nonhomologous chromosomes and function as headquarters for chromosomal DNA organization in interphase nuclei (Maison and Almouzni, 2004; Probst and Almouzni, 2011). Importantly, groups of chromosomes forming individual chromocenters depend on cell type and stage; i.e., the size and number of chromocenters differ among tissues of the same organism. When considering molecular mechanisms to achieve this cell type- and stage-dependent chromocenter formation, the enrichment of chromosome-dependent combination of TFBSs in pericentric heterochromatin regions (Iwasaki et al., 2013) is of particular interest, because the cellular complement of individual TFs is regulated in ways specific to cell type and stage. We have therefore proposed that the clustering of TFBSs in pericentric regions is involved in cell type- and stage-dependent chromocenter formation and supports condensed heterochromatin formation in chromocenters through TF-mediated chromatin interactions (Iwasaki et al., 2013).

During this previous study, we noticed that a class of CG-containing oligonucleotides is also enriched in human pericentric regions. The present study analyzes these CG-containing oligonucleotides, because methylation at C in CG dinucleotides is a well-known epigenetic modification; moreover, the binding of methylated-CpG-binding domain proteins (MBDs), as well as several structurally unrelated methyl-CpG-binding zinc-finger proteins, to methylated CG induces histone deacetylation, subsequent chromatin condensation, and heterochromatinization (Bogdanović and Veenstra, 2009). Since CG methylation in pericentric regions is thought to be an essential process for heterochromatinization in pericentric regions, we propose in this study that the CG-containing oligonucleotides enriched in pericentric regions may play roles in the formation of condensed heterochromatins in chromocenters and thus in the nuclear organization of chromosomal DNAs.

MATERIALS AND METHODS

BLSOM A SOM is an unsupervised clustering method that can visualize high-dimensional complex data on one plane (Kohonen et al., 1996). BLSOM for oligonucleotide composition was constructed as described by Abe et al. (2003), and oligonucleotides diagnostic for category-dependent separation were visualized as described by Abe et al. (2005). The BLSOM program can be obtained from y_wada@nagahama-i-bio.ac.jp. We previously constructed BLSOMs for tetra- and pentanucleotide compositions in 20-, 50-, and 100-kb sequences derived from the human genome and found that these BLSOMs for sequences with different lengths yielded almost identical conclusions

(Iwasaki et al., 2013).

U-matrix Distances of weight vectors between neighboring lattice points on a BLSOM can be visualized as grayness levels with a U-matrix method (Ultsch, 1993), and this provides information regarding dissimilarity of oligonucleotide composition in local areas on the BLSOM as a grayness level. The U-matrix was constructed as described previously (Iwasaki et al., 2013).

Human genome sequence The genome sequence of *Homo sapiens* (GRCh37) was obtained from the NCBI ftp site (http://www.ncbi.nlm.nih.gov/genomes/).

RESULTS

BLSOM for human genomic sequences To study intragenomic differences of oligonucleotide compositions in the human genome, we constructed pentanucleotide BLSOM for 50-kb sequences (abbreviated as Penta in Fig. 1A), based on our previous finding explained in MATERIALS AND METHODS. Lattice points containing sequences from a single chromosome are indicated in a color specifying the chromosome and those containing sequences from multiple chromosomes are indicated in black. Most lattice points are black, showing that most human sequences are not separated according to chromosome. However, several characteristic zones with colored or white lattice points are observed and designated as specific zones (SZs); a similar pattern was obtained for 100-kb Penta and tetranucleotide BLSOM (data not shown). Lattice points with no genomic sequence after BLSOM calculation are left white, and our previous studies (Abe et al., 2003, 2005) showed that lattice points containing sequences with oligonucleotide compositions very distinct from other sequences are surrounded by white lattice points. Because colored lattice points in an SZ are often surrounded by white lattice points, sequences located in an SZ (SZ sequences) are thought to have specific pentanucleotide compositions very distinct from most other

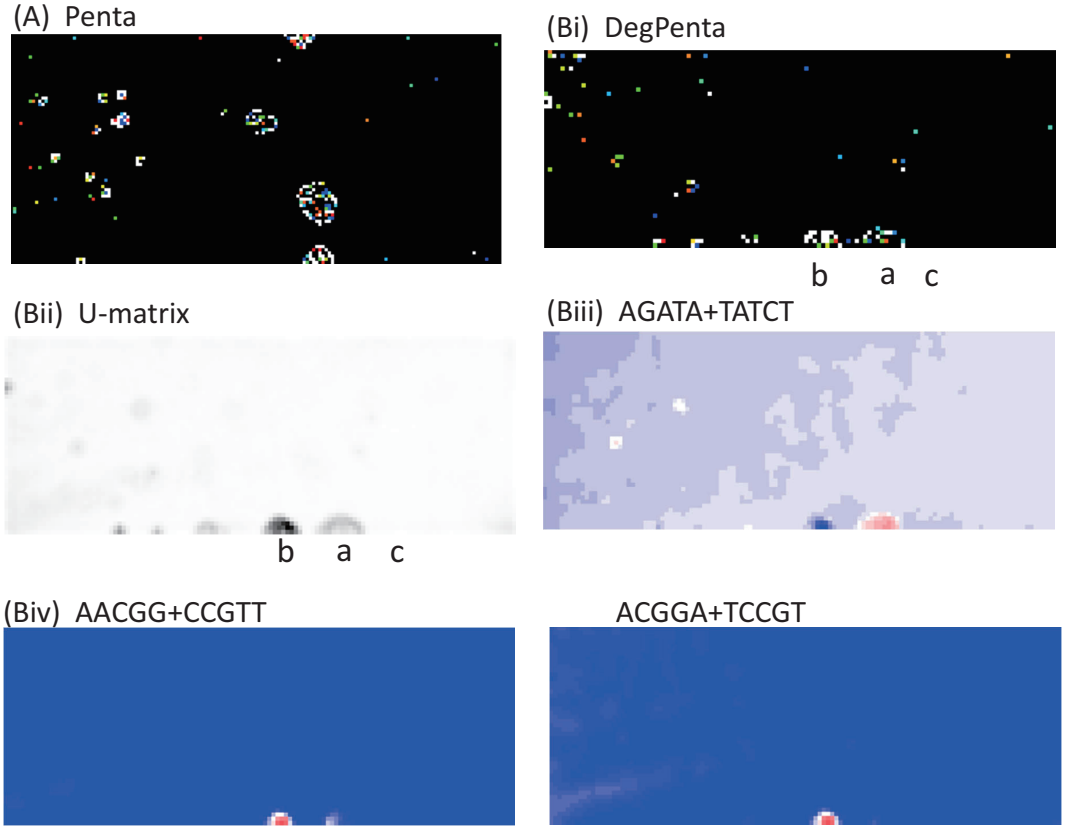

Fig. 1. Pentanucleotide BLSOM for 50-kb human sequences. (A and Bi) Penta and DegPenta. Map size was chosen as an average number of sequences of ten at one lattice point. Lattice points containing sequences from multiple chromosomes are indicated in black and those containing sequences from a single chromosome are indicated in color; for colors, see Supplementary Fig. S4A. Different colors specify the 24 chromosomes, but this is not important here because a colored lattice point indicates only that the lattice point contains sequences derived from a single chromosome. (Bii) U-matrix for DegPenta. (Biii, Biv) TFBS and CG-containing pentanucleotides, respectively, enriched in SZs. The expected frequency of each pentanucleotide from the mononucleotide composition was calculated at each lattice point, and the observed/expected ratio for the pentanucleotide is indicated in color, as described previously (Iwasaki et al., 2013).

human sequences, and this peculiar composition of SZs is confirmed below with an alternate method.

Figure 1A shows mirror-symmetric-type locations of SZs. Our previous studies, which analyzed a wide variety of eukaryotic, bacterial and archaeal genomes (Abe et al., 2003, 2006a), showed that sequences from a single genome are often split vertically into two territories according to the transcriptional and/or replicational direction of genes and genomic segments present in the sequence. In DNA databanks (DDBJ/EMBL/NCBI), only one strand of each pair of complementary sequences is registered. To study the general characteristics of genomic sequences, differences in oligonucleotide composition between two complementary strands are not important; furthermore, the obtained map should not be affected by the strand registered in DNA databanks. Therefore, we previously developed another type of BLSOM, in which the frequencies of a pair of complementary pentanucleotides (e.g., AAAAC and GTTTT) in each fragment are summed (Abe et al., 2005). This BLSOM (DegPenta, Fig. 1Bi) yielded a simpler pattern than Penta, because the mirror-symmetric split of SZs in the vertical direction disappears.

Dissimilarity of oligonucleotide composition between neighboring lattice points (and thus between sequences belonging to neighboring lattice points) can be visualized, using a U-matrix (Ultsch, 1993), with a grayness level. Dark gray lines on the U-matrix for DegPenta (Fig. 1Bii) correspond primarily to borders of SZs, showing that the oligonucleotide composition in SZ sequences is dissimilar to that of most other genomic sequences and supporting the above-mentioned peculiar oligonucleotide composition in SZs. Three major SZs are marked a, b and c in Fig. 1, Bi and Bii.

Oligonucleotides enriched in SZs Oligonucleotides, such as penta- and hexanucleotides, often represent motif sequences responsible for the sequence-specific binding of proteins such as transcription factors (Wingender, 1988), and their frequency should differ from that expected from the mononucleotide composition in the genome, and should differ among genomic regions. BLSOM can visualize diagnostic oligonucleotides responsible for clustering (self-organization) of sequences according to categories of interest and thus provides information about characteristics of sequences belonging to the category of interest (e.g., SZ sequences). In Fig. 1, Biii and Biv, and Fig. 2, after calculating the expected pentanucleotide frequency from mononucleotide composition at each lattice point, the observed/expected ratio for one pentanucleotide is indicated, with red, blue and white showing over-, under- and moderate representation, respectively, compared to the expected level (Abe et al., 2006a). This observed/expected ratio is useful for analyzing compositional fea-

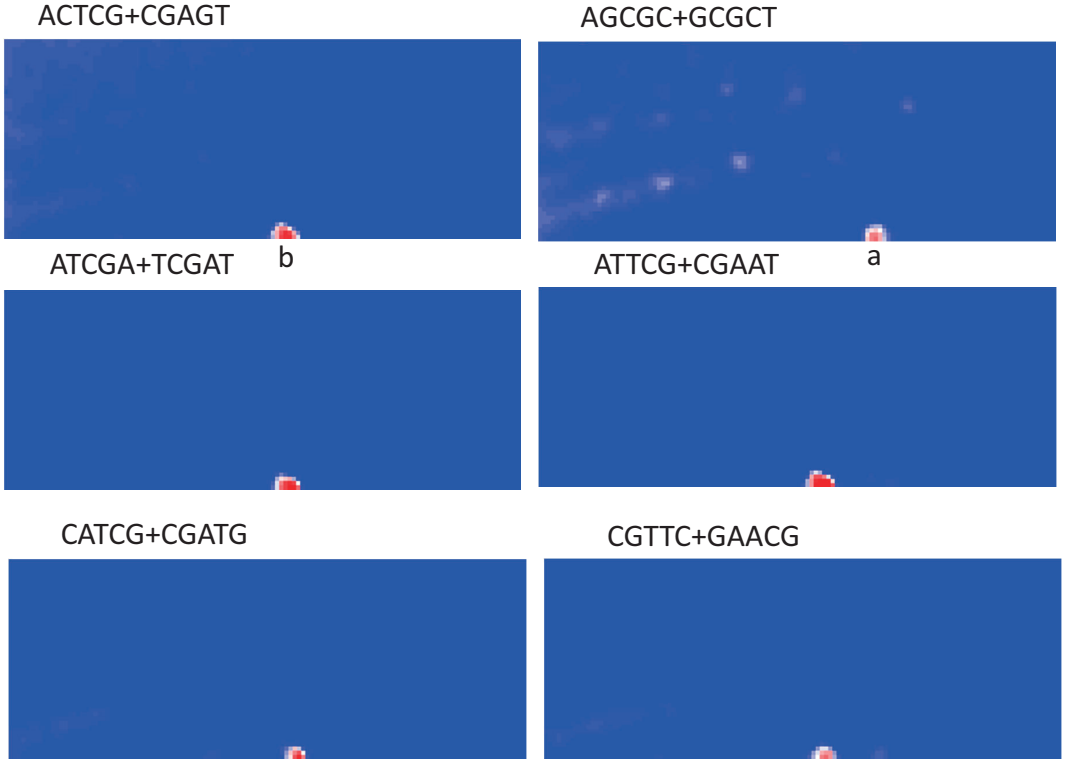

Fig. 2. CG-containing pentanucleotides enriched in SZs on DegPenta listed in Fig. 1Bi. Lattice points are marked as described in Fig. 1, Biii and Biv.

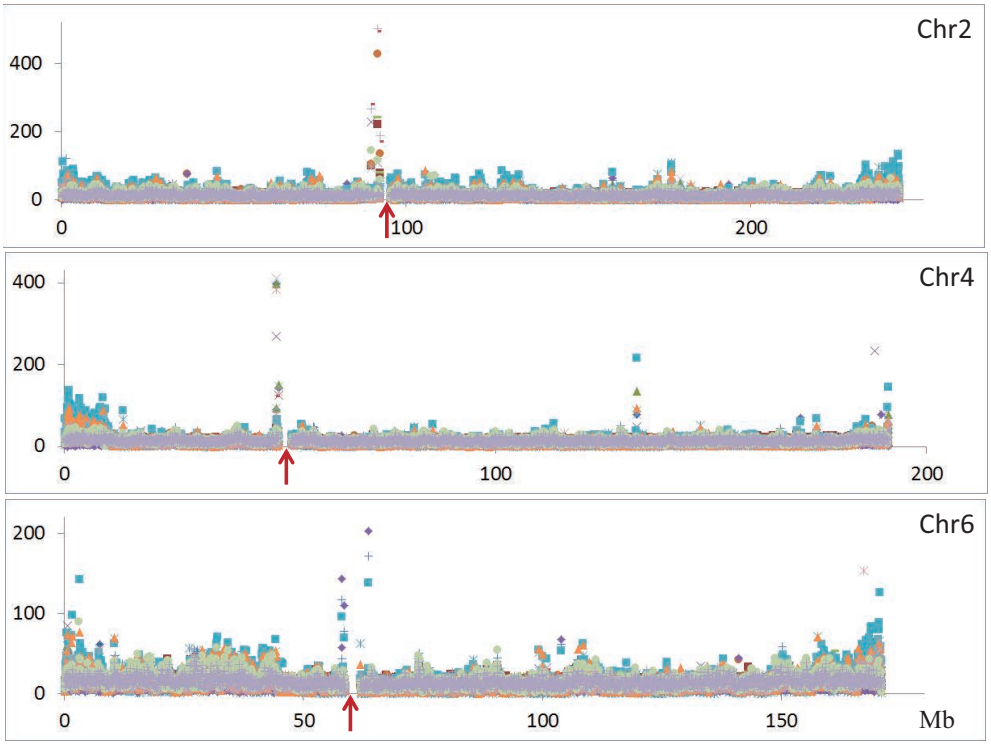

Fig. 3. Distribution of CG-containing pentanucleotides on individual chromosomes. Numbers of pentanucleotides per 50 kb are plotted with colored symbols distinguishing pentanucleotides enriched in SZs: ◆, AACGG/CCGTT; ■, AATCG/CGATT; ▲, ACGGA/TCCGT; X, ACTCG/CGAGT; *, AGCGC/GCGCT; ●, ATCGA/TCGAT; +, ATCGC/GCGAT; –, ATTCG/CGAAT; –, CATCG/CGATG; ◆, CGATA/TATCG; ■, CGCAG/CTGCG; ▲, CGCTC/GAGCG; X, CGGAA/TTCCG; *, CGTTC/GAACG; ●, CTCGA/TCGAG; and +, TCGAA/TTCGA. The centromeric region is marked with a red arrow. The colored symbols are listed in Supplementary Fig. S4B.

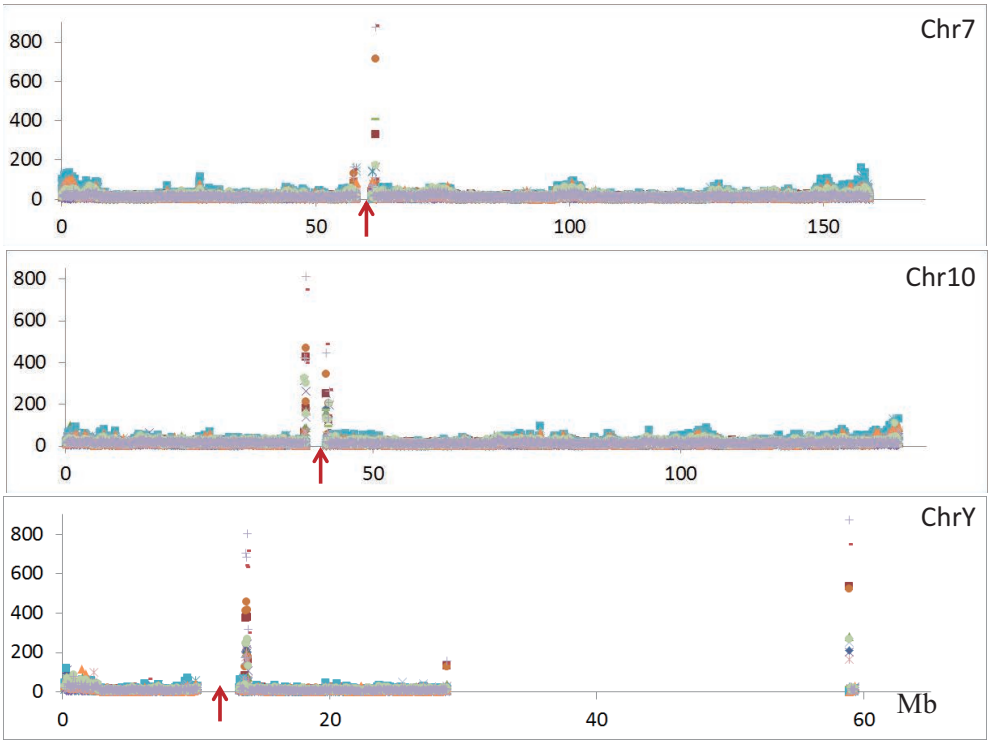

Fig. 4. Distribution of CG-containing pentanucleotides on individual chromosomes. Numbers of pentanucleotides per 50 kb are plotted as described in Fig. 3, and the centromeric region is marked with a red arrow.

tures of genomic sequences located in a certain zone such as an SZ, independent of a simplex effect reflecting the mononucleotide composition of genomic sequences.

We focus here on pentanucleotides that are enriched (red) in SZs but underrepresented (blue) in most other areas. Overrepresentation of a certain oligonucleotide only in an SZ should yield useful information for understanding the biological significance of sequences belonging to SZs, especially when a function of the oligonucleotide is known. We previously found that nine TFB motif pentanucleotides are enriched in SZ sequences (Iwasaki et al., 2013). AGATA/TATCT, listed in Fig. 1Biii, corresponds to one such pentanucleotide, a TFBS motif recognized by GATA-family TFs. While the result for this TFBS pentanucleotide was not presented in the previous paper, this pentanucleotide is clearly enriched in SZa (Fig. 1Biii); examples of other TFBS pentanucleotides enriched in SZa, b and c are presented in Supplementary Fig. S1.

Figures 1Biv and 2 show eight examples of CG-containing pentanucleotides that are enriched in SZs but underrepresented in most other regions. Of the 122 possible CG-containing pentanucleotide pairs, 16 show these enriched patterns, although the actual overrepresentation level between SZs, and even within one SZ, differs among pentanucleotide pairs, as previously found for TFBS motifs (see Supplementary Fig. S1B). It should be mentioned here that most of the residual CG-containing pentanucleotides are underrepresented in almost all areas including SZs (Iwasaki et al., 2013, 2014) because of a clear suppression of CG in mammalian genomes, which has been connected with methylation at C in a CG dinucleotide followed by spontaneous deamination and mutation to TG (Walsh and Bestor, 1999).

CG-containing pentanucleotides enriched in SZs on each chromosome DegPenta for 50-kb sequences revealed that 16 CG-containing pentanucleotides, as well as nine TFBS pentanucleotides found previously, are

Table 1. High occurrence of CG-containing pentanucleotides in pericentric regions

Chr	AACGG/ CCGTT	AATCG/ CGATT	ACGGA/ TCCGT	ACTCG/ CGAGT	AGCGC/ GCGCT	ATCGA/ TCGAT	ATCGC/ GCGAT	ATTCG/ CGAAT	CATCG/ CGATG	CGATA/ TATCG	CGCAG/ CTGCG	CGGAA/ TTCCG	CGTTC/ GAACG	CTCGA/ TCGAG	TCGAA/ TTCGA
1	h														h
2		H		H		H		H2	H					H	H1
3					h										
4			H2	H				h				H1	H		
5					h										
6							H2			H1					
7		H				H		H1	H					h	H2
8											h				
9					h										
10	H	H	H	H		H		H2	H					H	H1
11					h										h
12											h			h	
13															
14					h										
15					h										
16		H		H		H		H2	H					H	H1
17															
18					h										
19															
20	h	H	h	H	h			H2				h	h	H	H1
21	h	H	H	H		H		H2						H	H1
22		H		h		H		H2	H					h	H1
X	H1				h										h
Y	H	H		H		H		H2						H	H1

H: higher occurrence of the respective pentanucleotide in the pericentric region than that of any pentanucleotide in other regions; h: higher occurrence of the respective pentanucleotide in the pericentric region than that of this pentanucleotide in other regions. CGCTC/GAGCG, for which H and h are observed in no chromosome, is not listed, to reduce the table size.

enriched in a restricted portion of sequences (i.e., SZ sequences) but underrepresented in most human genomic sequences. To clarify the biological significance of SZ sequences, we previously investigated their chromosomal locations and found that SZ sequences enriched for TFBSs are mainly localized in pericentric heterochromatin regions (Iwasaki et al., 2013). This finding was unexpected because we know that there are few protein-coding genes in these regions, which are thus thought to be poor in TFBSs; unsupervised data mining such as BLSOM can therefore provide such unexpected knowledge.

Having obtained this unexpected knowledge with the unsupervised mining method, we can use a much simpler and more standard method to investigate the novel finding in more detail. As previously done for TFBS pentanucleotides, Figs. 3 and 4 simply plot the occurrences (per 50 kb) of each of the 16 CG-containing pentanucleotide pairs of interest on individual chromosomes, with symbols distinguishing each pentanucleotide; for their pericentric clustering on other chromosomes, refer to Table 1. Human centromeric regions are composed of highly repetitive sequences such as alphoid repeats, and the centromeric regions are not included in "*Homo sapiens* high coverage assembly GRCh37" reported by the Genome Reference Consortium (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/) because they lack unique sequences required for sequence assembly. Unsequenced centromeric regions are detected as open spaces with no data for pentanucleotide occurrence; the centromeric region is marked with a red arrow. Importantly, the highest occurrence of pentanucleotides of interest (per 50 kb) is observed in pericentric regions, as previously found for TFBS pentanucleotides; occurrence per 100 kb yields almost the same conclusion. The highest occurrence of several CG-containing pentanucleotides in pericentric regions is found not only on the six chromosomes displayed in Figs. 3 and 4, but also on most other chromosomes (see Supplementary Figs. S2 and S3, and Table 1). Importantly, the set of pentanucleotides that produces the highest peak differs among chromosomes; i.e.,

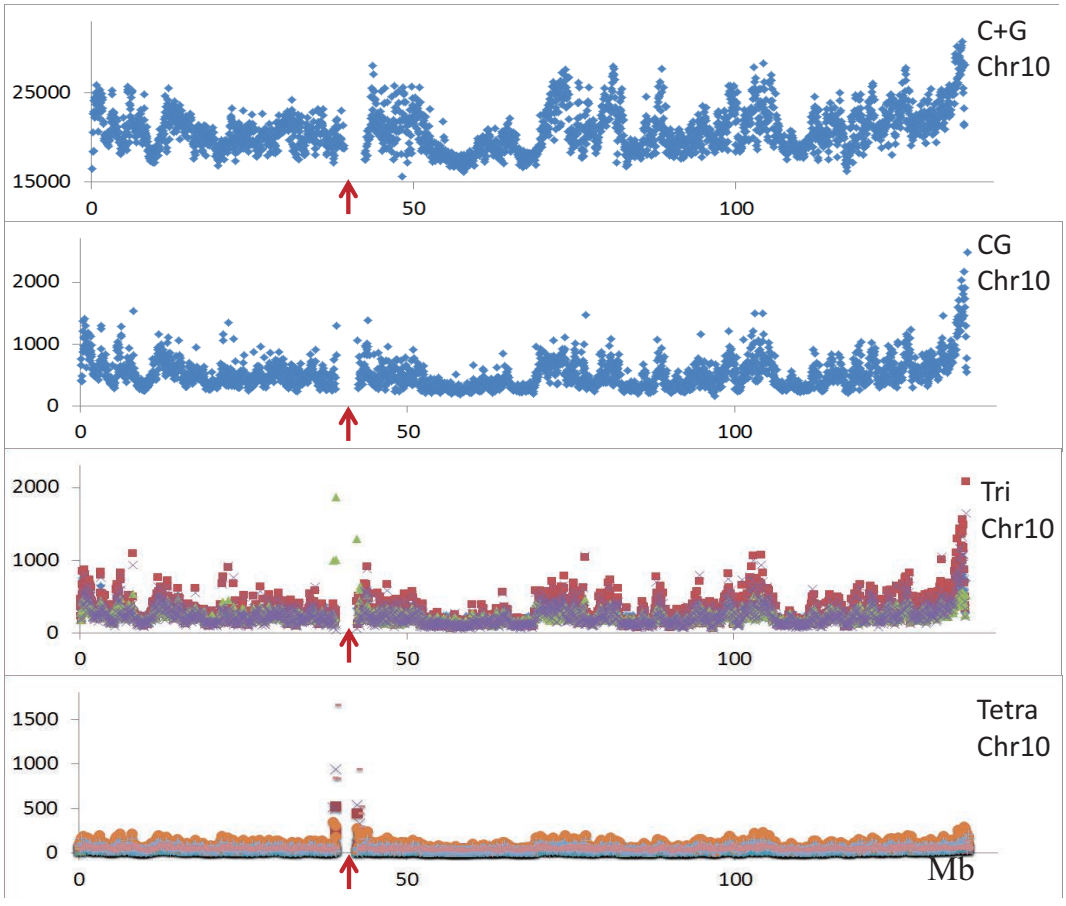

Fig. 5. Distribution of C+G, CG, and CG-containing tri- and tetranucleotides per 50 kb on chr10. Different trinucleotide pairs (Tri) are specified with the following symbols: ◆, ACG/CGT; ■, CCG/CGG; ▲, CGA/TCG; and X, CGC/GCG. Different tetranucleotide pairs (Tetra) are specified with the following symbols: ◆, ACGA/TCGT; ■, ATCG/CGAT; ▲, CCGA/TCGG; X, CGAA/TTCG; *, CGAC/GTCG; ● CGAG/CTCG; +, GCGA/TCGC; and –, TCGA/TCGA. The centromeric region is marked with a red arrow. The colored symbols are listed in Supplementary Fig. S4C.

chromosome-dependent enrichment occurs, as found for TFBS pentanucleotides.

In more detail, more than 200 occurrences of the pentanucleotide pairs (per 50 kb) are observed in pericentric regions of nine chromosomes. Only for chrY, a similar enrichment is also observed in the subtelomeric region of its long arm (Fig. 4). Table 1 summarizes high occurrence of each of the 16 CG-containing pentanucleotides on each chromosome. H shows higher occurrence of the respective pentanucleotide in pericentric regions than that of any pentanucleotides in other regions of the same chromosome, except for the subtelomeric region of chrY; H1 and H2 show the highest and the second highest case, respectively; and h shows higher occurrence of the respective pentanucleotide in pericentric regions than that of this pentanucleotide in other regions of the same chromosome (see chr1, 3, and X in Supplementary Figs. S2 and S3). Focusing on chromosome-dependent enrichment in more detail, a similar set of seven CG-containing pentanucleotides is enriched in the group of chromosomes comprising chr2, 7, 16, 22 and Y, but completely different sets are enriched in, for example, chr6 and chrX (Table 1, Figs. 3 and 4, Supplementary Figs. S2 and S3). To conclude, pericentric regions display the highest occurrence of CG-containing pentanucleotides (H and h in Table 1) for 21 chromosomes, but not for chr13, 17 or 19. It is possible that the unsequenced centromeric region of the latter three chromosomes is enriched for CG-containing pentanucleotides.

Oligonucleotides other than pentanucleotides

Because only 16 CG-containing pentanucleotide pairs among the 122 possible pairs are enriched in pericentric regions, their enrichment should not reflect a simple chromosomal distribution of CG dinucleotide occurrence, but be affected by sequence context surrounding CG. This prediction can be tested by directly analyzing CG dinucleotide occurrence on individual chromosomes. Figure 5 shows the occurrence of G+C and CG (per 50 kb) on chr10. High occurrence of CG is mainly observed in

Table 2. High occurrence of TFBS motif pentanucleotides in pericentric regions

Chr	AATCA/ TGATT	AATCT/ AGATT	AGATA/ TATCT	ATTGG/ CCAAT	CTATC/ GATAG	CTTCC/ GGAAG	GCCAA/ TTGGC	TATCA/ TGATA
1		h						
2	h	h					H1	
3		h	h					
4	H1		H2					
5		h	h					
6								H1
7	H1	h	h					
8		h	h					
9								
10	H1	h						
11		h	h					
12		h	h			h		
13								
14		h	h					
15		h	h				h	
16	H1	h	h					
17		h	h					
18								
19		h	h					
20			h					
21		h	h	h	h			
22	h						h	
X		h	h		h			h
Y	h	h	h	h				

H and h are defined in Table 1. ACCAC/GTGGT, for which H and h are observed in no chromosome, is not listed, to reduce the table size.

G+C-rich regions, which are known to be gene-rich (Bernardi, 2004), and no enrichment is observed in pericentric regions; other chromosomes show similar results (data not shown).

Next, analyzing all four pairs of CG-containing trinucleotides, enrichment of CGA/TCG is observed in pericentric regions in chr2, 7, 10, 16, 20, 21 and Y (see green triangles for Tri Chr10 in Fig. 5), on which six CGA/TCG-containing pentanucleotides (AATCG/CGATT, ACTCG/CGAGT, ATCGA/TCGA, ATTCG/CGAAT, CTCGA/TCGAG and TCGAA/TTCGA) are enriched (Table 1). Table 1 also shows that the other two CGA/TCG-containing pentanucleotides (ATCGC/GCGAT and CGATA/TATCG) are not enriched in these seven chromosomes, but are enriched in chr6; thus, sequence surrounding CGA/TCG produces the chromosome-dependent difference. Successively analyzing all eight CGA/TCG-containing tetranucleotide sets, two pairs (ATCG/CGAT and CGAA/TTCG) and one palindromic tetranucleotide (TCGA) are enriched in pericentric regions on these seven chromosomes, and one example is shown in Fig. 5 (Tetra Chr10). The finding that less than half of the CGA/TCG-containing tetranucleotides are enriched in pericentric regions in these seven chromosomes again shows the importance of sequence context surrounding this trinucleotide. In addition, among eight ATCG/CGAT- or CGAA/TTCG-containing and four TCGA-containing pentanucleotides enriched in the seven chromosomes, only five, two and three pentanucleotides are included in the 16 pairs of interest listed in Table 1, showing that enrichment of the CG-containing pentanucleotides is not a mere reflection of the above-mentioned CG-containing tetranucleotides (ATCG/CGAT, CGAA/TTCG and TCGA).

The finding that nine of the 16 pentanucleotide pairs listed in Table 1 contain CGA/TCG in their sequences shows that CGA/TCG is an important element of the pentanucleotides of interest. However, these nine correspond to a minor set of the total of 44 CGA/TCG-containing pentanucleotides; and importantly, CG-containing pentanucleotides that lack CGA/TCG are also enriched in pericentric regions. It should also be mentioned that no CG-containing pentanucleotides among the 16 pairs of interest are composed only of C and G. For example, pentanucleotides constituting the core element of the

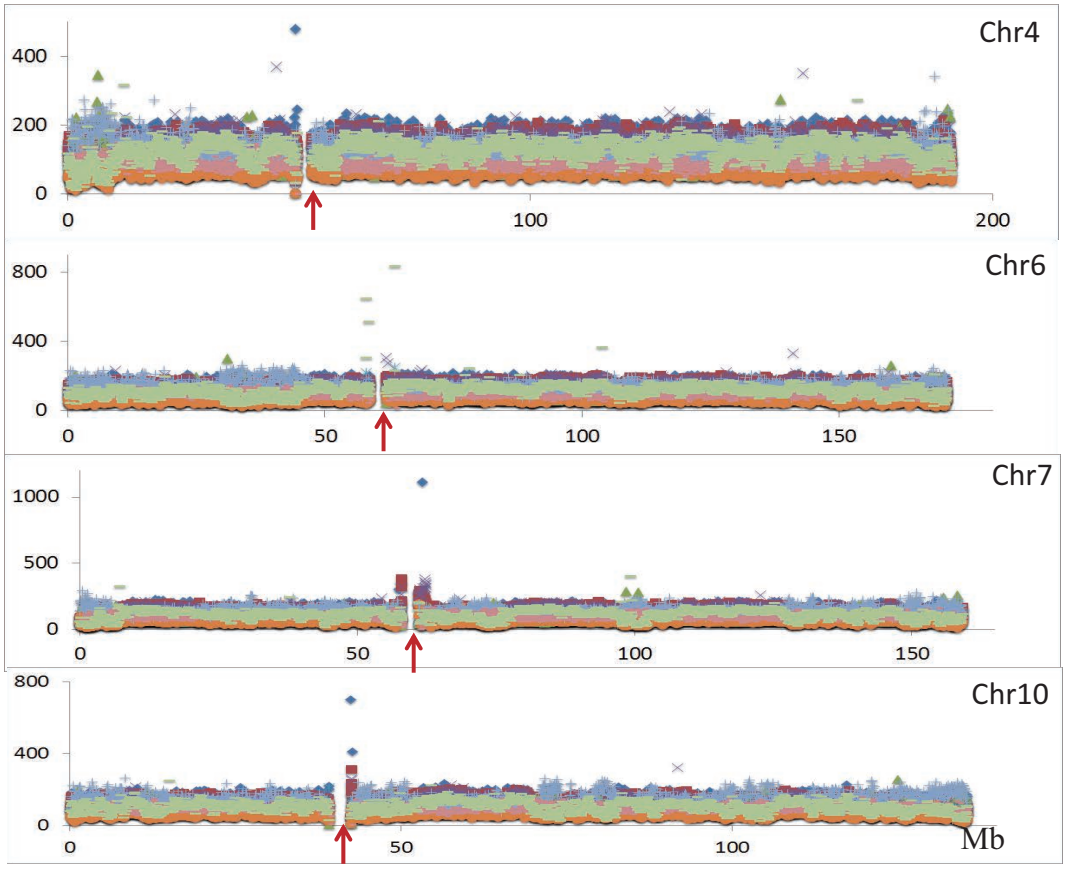

Fig. 6. Distribution of TFBS pentanucleotides on individual chromosomes. Numbers of TFBS pentanucleotides per 50 kb are plotted with symbols distinguishing pentanucleotides as follows: ◆, AATCA/TGATT; ■, AATCT/AGATT; ▲, ACCAC/GTGGT; X, AGATA/TATCT; *, ATTGG/CCAAT; ●, CTATC/GATAG; +, CTTCC/GGAAG; –, GCCAA/TTGGC; and –, TATCA/TGATA. The centromeric region is marked with a red arrow. The colored symbols are listed in Supplementary Fig. S4D.

TFBS for SP1, GGGCGG, are primarily enriched in G+C-rich regions, which are known to be gene-rich (Bernardi, 2004). The intrachromosomal location of various CG-containing oligonucleotides probably reflects their functional differences.

TFBS motif pentanucleotides Because the enrichment of CG-containing pentanucleotides in pericentric regions is chromosome-dependent, we have compared this chromosome dependency with that of TFBS pentanucleotides. As done for CG-containing pentanucleotides in Table 1, Table 2 lists high occurrence of TFBS pentanucleotides in pericentric regions on each chromosome and reveals a similar chromosome dependency for these two types of pentanucleotides. For example, the group of chromosomes comprising chr2, 7, 16 and Y is enriched for a similar set of CG-containing pentanucleotides and also for a similar set of TFBSs, and the pattern of enrichment clearly differs from that of chr8 for these two types of pentanucleotide; examples of chromosome-dependent distribution of TFBS pentanucleotides are listed in Fig. 6. This similar chromosome dependency for CG-containing and TFBS pentanucleotides suggests that these two types of pentanucleotides enriched in pericentric regions have a related function.

DISCUSSION

Biological function of CG-containing pentanucleotides enriched in pericentric regions CpG islands (Bird, 1987), which play important roles in transcriptional regulation, are typically a few hundred bp in length, and exist preferentially in gene-rich regions. Therefore, CG-containing pentanucleotides enriched in 50-kb (and 100-kb) SZ sequences in pericentric regions, which are poor in protein-coding genes, probably have roles that are different from canonical CpG islands. Oligonucleotides such as penta- and hexanucleotides often represent motif sequences responsible for protein binding, and we previously proposed that clustering of TFBS pentanucleotides in pericentric regions plays roles in cell type- and cell stage-dependent formation of chromocenter heterochromatin, through TF-mediated chromatin interactions (Iwasaki et al., 2013). As mentioned above, the similarity of chromosome dependency between CG-containing and TFBS pentanucleotides indicates that these two types of pentanucleotides may have a related function. The strict sequence requirement surrounding a CG (i.e., dependency on the context sequence) found in this study is consistent with a view that these CG-containing pentanucleotides act as motif sequences for protein factors because a one-base difference in oligonucleotide sequence is usually crucial for protein binding.

Here, we propose that CG-containing pentanucleotides of interest act as binding-motif sequences for protein factors that are involved in heterochromatin formation of pericentric regions, such as factors that bind to methylated and/or unmethylated CG-containing oligonucleotide motifs. The human methylated-CG-binding protein MeCP2 requires an A/T-rich sequence that follows the methylated C for its binding (i.e., the sequence context surrounding the methylated C is important) and is involved in the formation of chromatin loops and nuclear organization (Klose et al., 2005; Bogdanović and Veenstra, 2009). In addition, the production of methylated-CG-binding proteins such as MeCP2 is regulated in ways that are specific to cell type and stage, and these proteins associate with various chromatin modifiers to establish a repressive chromatin environment, providing a connection between DNA methylation and chromatin modification (Bogdanović and Veenstra, 2009). These observations are consistent with our proposal that CG-containing oligonucleotides are involved, along with TFBS motifs, in condensed heterochromatin formation in chromocenters, which function as headquarters in nuclear organization.

Oligonucleotides longer than pentanucleotides Oligonucleotide composition is an example of high-dimensional data. To obtain novel knowledge from high-dimensional data in an efficient way, especially from big data, an informatics tool with strong visualization power becomes essential, and our group has introduced BLSOM for oligonucleotide composition. When considering actual motif sequences for the binding of various proteins, information about oligonucleotides longer than pentanucleotides becomes important. In the case of PC computers, because the pentanucleotide is the longest oligonucleotide routinely analyzable with BLSOM, longer oligonucleotides have not been analyzed in the present study. While information about pentanucleotides can provide information about longer oligonucleotides in an indirect way, we have started to analyze longer oligonucleotides directly, using high-performance supercomputers. SZ regions similar to those found here (Fig. 1) have been observed for hexanucleotide BLSOM (to be published elsewhere).

REFERENCES

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

Abe, T., Sugawara, H., Kinouchi, M., Kanaya, S., and 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., Kanaya, S., Kinouchi, M., and Ikemura, T. (2006a) Self-Organizing Map (SOM) unveils and visualizes hidden sequence characteristics of a wide range of eukaryote genomes. Gene **365**, 27–34.

Abe, T., Sugawara, H., Kanaya, S., and Ikemura, T. (2006b)

Sequences from almost all prokaryotic, eukaryotic, and viral genomes available could be classified according to genomes on a large-scale Self-Organizing Map constructed with the Earth Simulator. J. Earth Simulator **6**, 17–23.

Bernardi, G. (2004) Structural and Evolutionary Genomics: Natural Selection in Genome Evolution. Elsevier, Amsterdam; New York.

Bird, A. (1987) CpG islands as gene markers in the vertebrate nucleus. Trends Genet. **3**, 342–347.

Bogdanović, O., and Veenstra, G. J. C. (2009) DNA methylation and methyl-CpG binding proteins: developmental requirements and function. Chromosoma **118**, 549–565.

Iwasaki, Y., Wada, K., Wada, Y., Abe, T., and Ikemura, T. (2013) Notable clustering of transcription-factor-binding motifs in human pericentric regions and its biological significance. Chromosome Res. **21**, 461–474.

Iwasaki, Y., Abe, T., Okada, N., Wada, K., Wada, Y., and Ikemura, T. (2014) Evolutionary changes in vertebrate genome signatures with special focus on coelacanth. DNA Res. **21**, 459–467. doi:10.1093/dnares/dsu012.

Kanaya, S., Kinouchi, M., Abe, T., Kudo, Y., Yamada, Y., Nishi, T., Mori, H., and Ikemura, T. (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, A. M., and Mrazek, J. (1998) Comparative DNA analysis across diverse genomes. Annu. Rev. Genet. **32**, 185–225.

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

Klose, R. J., Sarraf, S. A., Schmiedeberg, L., McDermott, S. M., Stancheva, I., and Bird, A. P. (2005) DNA binding selectivity of MeCP2 due to a requirement for A/T sequences adjacent to methyl-CpG. Mol. Cell **19**, 667–678.

MacQuarrie, K. L., Fong, A. P., Morse, R. H., and Tapscott, S. J. (2011) Genome-wide transcription factor binding: beyond direct target regulation. Trends Genet. **27**, 141–148.

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

Nakao, R., Abe, T., Nijhof, A. M., Yamamoto, S., Jongejan, F., Ikemura, T., and Sugimoto, C. (2013) A novel approach, based on BLSOMs (Batch Learning Self-Organizing Maps), to the microbiome analysis of ticks. ISME J. **7**, 1003–1015.

Probst, A. V., and Almouzni, G. (2011) Heterochromatin establishment in the context of genome-wide epigenetic reprogramming. Trends Genet. **27**, 177–185.

Sanyal, A., Lajoie, B. R., Jain, G., and Dekker, J. (2012) The long-range interaction landscape of gene promoters. Nature **489**, 109–113.

Uehara, H., Iwasaki, Y., Wada, C., Ikemura, T., and Abe, T. (2011) A novel bioinformatics strategy for searching industrially useful genome resources from metagenomic sequence libraries. Genes Genet. Syst. **86**, 53–66.

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 (eds.: S. Gielen and B. Kappen), pp. 864–867. Springer, London, UK.

Wingender, E. (1988) Compilation of transcription regulating proteins. Nucl. Acids Res. **16**, 1879–1902.

Walsh, C. P., and Bestor, T. H. (1999) Cytosine methylation and mammalian development. Genes Dev. **13**, 26–34.