Biology

Biology fields

Okayama University *Year* 2006

A novel non-coding DNA family in Caenorhabditis elegans

Yasuo Takashima* Tetsuya Bando†

Hiroaki Kagawa‡

*Okayama University

†Okayama University

‡Okayama University, hkagawa@cc.okayama-u.ac.jp

A novel non-coding DNA family in *Caenorhabditis elegans*

Keywords: repetitive element; palindrome; gene/element association; enhancer; genome organization.

Yasuo Takashima, Tetsuya Bando, Hiroaki Kagawa*

Division of Biomolecular Science, Graduate School of Natural Science and Technology, Okayama University, Okayama 700-8530, Japan

*Corresponding author:
Tel: +81-86-251-7865; Fax: +81-86-251-7876; E-mail: hkagawa@cc.okayama-u.ac.jp

Abbreviations: SINE (LINE), short (long) interspersed element; LTR, long terminal repeat; SSR, simple sequence repeat; *gfp*, green fluorescent protein.

Abstract

Many repetitive elements, for example, SINEs, LINEs, LTR-retrotransposons and other SSRs are dispersed throughout eukaryotic genomes. To understand the biological function of these repetitive elements is of great current research interest. In this study, we report on the identification of a novel non-coding DNA family, designated CE1 family, in the nematode *C. elegans* genome. Some CE1 elements constituted a large palindrome sequence. The CE1 elements were interspersed at 95 sites in the *C. elegans* genome. Most of the CE1 elements were associated with, or are within, protein-coding genes. The sequence of the CE1 elements indicated that some could form a hairpin structure. One of the CE1 family, CE1(*bs258*), is located in the first intron of a novel gene, C46H11.6 which encodes a PDZ/DHR/GLGF domain protein. In *gfp* and *lacZ* reporter gene assays the CE1(*bs258*) element appeared to behave as an enhancer element for the expression of C46H11.6 but no effect on the expression of the opposite direction gene, *pat-10* which encodes the body-wall muscle troponin C. The CE1(*bs258*) RNA transcript was detected by RT-PCR even when CE1(*bs258*) was located in an intron. We conclude that CE1 elements are involved in the expression of adjacent genes and are therefore selectively retained in the *C. elegans* genome. We discussed a biological function of the CE1(*bs258*) having many transcription factor-binding sites.

1. Introduction

Recent studies of the genomes of human, mouse, *Drosophila* and *C. elegans* indicates that these encode 24,194, 26,996, 14,399 and 20,603 genes, respectively (*C. elegans* Sequencing Consortium, 1998; Adams et al., 2000; International Human Genome Sequencing Consortium, 2001; Waterston et al., 2002). Dividing the total genome size by the number of genes leads to an estimate of average gene as 135.2 kb in human, 84.0 kb in mouse, 9.2 kb in *Drosophila* and 4.9 kb in *C. elegans*. This suggests that more complex organisms have more complex gene structure, although there are some exceptions among plants and Amphibia. Repetitive elements are abundant genetic components in eukaryotic genomes, corresponding to 43% of the human genome (International Human Genome Sequencing Consortium, 2001), 37% of the mouse genome (Waterston et al., 2002), ~20% of the *Drosophila* genome (Adams et al., 2000) and 16.5% of the *C. elegans* genome (*C. elegans* Sequencing Consortium, 1998). These data indicate that higher organisms have a higher proportion of repetitive elements. Therefore, it is important to know the structure and function of repetitive elements including, for example, non-coding DNA, *cis*-regulatory elements, and transposons. Eukaryotic genomes contain many types of repetitive elements that consist of short and long interspersed elements (SINEs and LINEs, respectively), LTR-retrotransposons, DNA transposons and other simple sequence repeats (SSRs) (International Human Genome Sequencing Consortium, 2001). As some of the repetitive elements do not encode protein or RNA genes, these elements are undoubtedly involved in regulation of gene expressions or maintenance of chromosomes (*C. elegans* Sequencing Consortium, 1998). The function of repetitive elements of eukaryotic genome is still largely unknown.

In *C. elegans*, several repetitive sequences are located throughout the genome. Tandem and inverted repeats account for 2.7% and 3.6% of the entire genome size and are found, on average, once per 3.6 kb and 4.9 kb, respectively (*C. elegans* Sequencing Consortium, 1998). The *Cer* LTR-retrotransposon family is the most abundant class of retrotransposons (Ganko et al., 2001). Many repetitive families are distributed nonuniformly within the genome and are more likely to be found within an intron than an intergenic region ("intergenic region" means the space between gene *A* and gene *B*) (*C. elegans* Sequencing Consortium, 1998). CeRep26 is a tandem hexamer repeat, TTAGGC, in *C. elegans* which is present at multiple sites along chromosomes in addition to the telomeres (Wicky et al., 1996). A non-random distribution has also been shown for the 711 copies of the CeRep11 family. The CeRep11 elements are distributed all over the autosomes but only a single copy is located on the X chromosome (*C. elegans* Sequencing Consortium, 1998). Most repetitive families are generated by transposition (Smit, 1996; Duret et al., 2000) and these are currently classified into 53 families for the *C. elegans* genome (WormBase, http://www.wormbase.org/). However, these repetitive elements may not always encode an active transposon (Ketting et al., 1997). The *Tc1/mariner* type of transposons are well characterized and many members of this family do not encode an active transposase. Four subfamilies of the *Tc1/mariner* are highly divergent from each other and the other subfamily members; these are probably no longer active in the genome (*C. elegans* Sequencing Consortium, 1998).

The repetitive elements comprise a rich paleontological record and may hold crucial clues about evolutionary events (*C. elegans* Sequencing Consortium, 1998). The difference in size between the *C. elegans* (100.3 Mb) and the *C. briggsae* (104 Mb) genomes is almost entirely due to repetitive elements that account for 16.5% of the *C. elegans* genome in contrast to 22.4% of the *C. briggsae* genome (Stein et al., 2003). The nematodes *C. elegans* and *C. briggsae* diverged from a common ancestor roughly 80-110 million years ago (Stein et al., 2003). A recent study indicates that the *C. briggsae* genome differs dramatically from the *C. elegans* genome in clustering of reproductive genes (Miller et al., 2004). How genome organization occurs in different species is an interesting question. Some may function to affect gene distribution in genomes. Analysis of repetitive elements in *C. elegans* may help us understand their contributions to genome organization.

In this study, we identified a palindromic non-coding DNA element, CE1(*bs258*) and found it associated with a nearby gene. From the study of this element we identified the CE1 family, a novel repetitive family consisting of 95 elements in the nematode *C. elegans* genome. These elements are distributed throughout the nematode chromosomes. The CE1 elements are closely associated with, or are within, genes. We also demonstrated that CE1(*bs258*) acts as an enhancer element for the adjacent gene C46H11.6. Additionally, the CE1(*bs258*) RNA transcript remained stable in RT-PCR experiments. These results suggest that the CE1 family members function as enhancer elements for gene expression regulation. We discussed additional functions of the CE1 family for regulating the concentration of transcription factors and the conformational stabilization of chromatin.

2. Materials and methods

2.1. Worm handling

All the worms derived from the wild-type Bristol N2 were grown under standard conditions, as described by Sulston and Hodgkin (1988).

2.2. Reverse transcription-polymerase chain reaction (RT-PCR)

Mixed staged N2 worms were collected and washed three times with M9 buffer to eliminate the bacteria. Total worm RNA was extracted using the TRISOL® LS Reagent (GIBCO BRL) according to the provided protocol. DNA contamination was checked with normal PCR by using the extracted RNA as a template, and it was confirmed by the absence of a signal.

Total RNA (50 ng/μl) was used for subsequent RT-PCR. RT-PCR was performed for detecting the CE1(*bs258*) RNA transcript. For cDNA synthesis from the CE1(*bs258*) transcript, the oligo DNA primers were designed specifically for CE1(*bs258*) as follows: forward primer, 5'-AAG CGG GCC AAC TTC ATA AC-3' and reverse primer, 5'-CGG CAA AAC ATC ACA ACT TC-3'(primer set 3). Additionally, two control primer sets were designed as follows. Forward primer, 5'-GCG GGA ATT CGA TGG TTT CTT TCC-3' and reverse primer, 5'-GTG TGA GCT CTG TCT GAT ACT TGG-3' (primer set 1) for the partial fragment of the C46H11.6 mRNA. Forward primer, 5'-GCG AAT TCA ACG GCC GGA TAA CCC GAA A-3' and reverse primer, 5'-GCG GAT CCG TGG ACA GAA GAG TTT TGA A-3' (primer set 2) for no detection of the intron of the C46H11.6 pre-mRNA. cDNA was synthesized by reverse transcription by using 5'-RACE System for Rapid Amplification of cDNA Ends Reagent Assembly (GIBCO BRL) with the oligo DNA primers. This cDNA (30 ng/μl of total cDNA per reaction) was then used for normal PCR.

The standard conditions for PCR were as follows: 98°C for 10 sec, 30 cycles of 98°C for 5 sec, 50°C for 10 sec, and 72°C for 1 min, followed by 72°C for 10 min. Finally, these RT-PCR products were confirmed by DNA sequencing using 373S DNA sequencer (Applied Biosystems).

2.3. gfp expression study and lacZ enhancer assay

C46H11.6 genomic DNA clone was derived from pTNC2, which contained two genes *pat-10* and C46H11.6 (Terami et al., 1999). Reporter plasmids; pC46H11BS1888, CE1(*bs258*)::*gfp* and pC46H11BE1085, Δ[CE1(*bs258*)]::*gfp* were constructed by inserting the restriction fragments of pTNC2, *Bam*H I-*Sac* I and *Bam*H I-*Eco*R I, respectively, into *Bam*H I-*Sma* I sites of pPD95.75, the promoterless *gfp* reporter plasmid (a generous gift from A. Fire). These constructs were generated as translational fusions with *gfp*. Reporter plasmids; pC46H11BB1439, CE1(*bs258*)::*gfp* and pC46H11BB1164, Δ[CE1(*bs258*)]::*gfp* were constructed by inserting the PCR fragment into *Bam*H I site of pPD95.77, the promoterless *gfp* reporter plasmid (a generous

gift from A. Fire). The oligo DNA primers were designed as follows: forward primer, 5'-CGG GAT CCG TCG ATT TCA GCA AGA-3' and reverse primers, 5'-GCG GAT CCC GGG GAT AAA TAT ACA-3' and 5'-CAG GAT CCA TCC GGC CGT TGG AAA-3' for pC46H11BB1439 and pC46H11BB1164, respectively. These two constructs were generated as transcriptional fusions with *gfp*. These *gfp* reporter plasmids (100 ng/μl) along with pRF4, a plasmid containing the dominant morphological marker *rol-6(su1006)* (100 ng/μl) were microinjected into the gonadal cavity of young-adult hermaphrodites, as described (Mello and Fire, 1995; Terami et al., 1999).

pPD95HHCE1bs258, CE1(*bs258*)::*lacZ* reporter plasmid was constructed by inserting the PCR fragment into *Hin*d III site of pPD95.18, the Δ*pes-10* minimal promoter::NLS::*lacZ* (a generous gift from A. Fire). The oligo DNA primers were designed specifically for CE1(*bs258*) as follows: forward primer, 5'-CCC AAG CTT CGC ATA ACC CGA AAC-3' and reverse primer, 5'-GGA AGC TTC GGG GAT AAA TAT ACA-3'. This minimal promoter does not drive *lacZ*-expression in any of these tissues on its own (Seydoux and Fire, 1994). The *lacZ* reporter plasmid (100 ng/μl) along with pRF4 (100 ng/μl) were microinjected in young-adult animals. The transgenic animals were fixed and *X-Gal* stained, as previously described (Mello et al., 1991; Terami et al., 1999).

The fluorescent or Nomalski images were captured with Nikon C1 confocal microscope system by using Zeiss Axioplan 2.

2.4. Data mining, multiple sequence alignment and sequence annotation

By using the BLASTN algorithm, preliminary sequences were aligned against public data sets of DNA sequences available online at The Welcome Trust Sanger Institute (http://www.sanger.ac.uk/cgi-bin/blast/submitblast/C.elegans). Sequence alignments were analyzed by using ClustalX 1.8.1 downloaded from ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/. Editing of the final alignment for display was performed using ClustalW 1.8.2 available online at http://clustalw.genome.jp/ (Thompson et al., 1994). Each CE1 element was annotated using WormBase (WS155 version of the database, http://www.wormbase.org/). The targeting sites of transcription factors were searched using MOTIF (Cut off score: 85), which is available online at http://motif.genome.jp/MOTIF.html, supported by TRANSFAC, a database of eukaryotic *cis*- and *trans*-regulatory elements (http://transfac.gbf.de/TRANSFAC; Heinemeyer et al., 1999).

2.5. Phylogenetic analysis

Phylogenetic analysis was performed on the multiple sequence alignments by using distance methods used by ClustalW. The phylogenetic tree was generated by the neighbor-joining methods (Saitou and Nei, 1987) and evaluated by 1,000 bootstrap replications. The phylogram was visualized by TreeView 1.6.6 downloaded from http://taxonomy.zoology.gla.ac.uk/rod/treeview.html (Page, 1996).

2.6. RNA folding prediction

The RNA secondary structure was calculated using the programs available in the GENETYX-MAC version 11.2.6 (Software Development), the MFOLD 3.1.2 (http://bioweb.pasteur.fr/seqanal/interfaces/mfold-simple.html; Zuker et al., 1999), and Vienna RNA Package 1.5 via the web interface to the RNAfold program (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi; Hofacker, 2003).

3. Results

3.1. CE1(bs258) is a member of the CE1 family

During a study of the *cis*-regulatory elements of the body-wall (skeletal) muscle troponin C gene, *pat-10* (Terami et al., 1999), we found a novel repetitive element and designated the 258 bp sequence as CE1(*bs258*). CE1(*bs258*) is located approximately 1 kb upstream of *pat-10* and in the first intron of the gene C46H11.6, a novel gene, encoding a PDZ/DHR/GLGF protein interaction domain protein. *pat-10* and C46H11.6 are transcribed bidirectionally. In a BLAST search using the 258 bp of CE1(*bs258*) as query, 95 non-coding DNA elements were identified in the genome (P value<0.01). We named this 95 element family CE1. After identifying this family we confirmed that it is a novel repeat DNA family by doing a homology search against the 53 repetitive sequence families previously submitted to WormBase. Our results showed no hits to any of the previously identified DNA element families confirming the CE1 family as a novel repetitive family in the *C. elegans* genome. The total sequence of the 95 CE1 elements is 18,832 bp which corresponds to 0.02% of the entire genome. The CE1 elements come in varying lengths, over 231 bp length element accounting for approximately 52.6% of the CE1 elements (Fig. 1).

The 95 elements of CE1 family were classified into five different subfamilies based on 10% nucleotide substitution on a phylogenetic tree (Fig. 2). CE1.1, CE1.2, CE1.3, CE1.4 and CE1.5 comprised 25.3%, 13.7%, 18.9%, 12.6% and 29.5% respectively, of the total CE1 elements. Although CE1(*bs258*) was classified into the CE1.1 subfamily that consisted of 24 members, each member of this subfamily was slightly divergent from each other (Fig. 2 and 3a). The sequence alignments of 24 members of the CE1.1 subfamily are shown and CE1(*bs258*) is labeled as C46H11a+/CE1bs258 in Fig. 3a. This alignment shows that CE1(*bs258*) is slightly divergent (P value<0.00005) from the CE1.1 subfamily.

3.2. CE1(bs258) is constructed by a palindromic DNA

A homology search among sense and anti-sense strands of CE1(*bs258*) showed a highly conserved alignment (P value<0.00001) (Fig. 3b). Moreover, both 5' and 3' ends show extremely high homology (90.0%/40 bp), suggesting that CE(*bs258*) can form a hairpin (or cruciform) structure (Fig. 3b). Additionally, their internal SSRs (simple sequence repeats) consisting of direct repeats of short *k*-mers such as $A_{2\text{-}5}$ and $T_{2\text{-}4}$ appear to contribute palindromic regions. Although each subfamily was slightly divergent, the internal palindromic structure was highly conserved for the CE1.1 subfamily (Fig. 2 and 3a).

3.3. The CE1 elements are nearby or within genes

The chromosomal distribution of the 95 CE1 elements is shown in Fig. 4. The CE1 elements are evenly distributed across all the chromosomes. There are some areas of clustering,

note particularly the left arms of chromosomes II and III, and the center right arm of chromosome IV and V (Fig. 4). Analysis of the DNA sequence for a 10 kb window on each side of a CE1 element identifies 225 gene/element associations (Fig. 5a,b). We found that the number of CE1 elements located within 2 kb of a gene was significantly greater than the expected association (120 gene/element associations; the Wilcoxon test: $p<0.005$; McGhee and Krause, 1997). Further examination showed 50 CE1 elements within 2 kb upstream region of a gene ($p<0.01$) and 47 CE1 elements within 2 kb downstream of a gene ($p<0.005$) (Fig. 5b,c). We designated these gene/element association as "intergenic associations". This means that almost 0.39% of *C. elegans* genes have CE1 elements within 2 kb of genes on either side. The 100.3 Mb *C. elegans* genome contains 20,603 predicted genes, including RNA genes and pseudogenes, with an average density of 1 gene per 4.9 kb (*C. elegans* Sequencing Consortium, 1998). As 53.3% (120/225) of the gene/element associations are within 2 kb of genes (Fig. 5b,c) we explored the possibility that CE1 elements might function as regulatory elements for gene expression (see Section 3.5; Fig. 7).

Some CE1 elements are within an intron of genes (Fig. 4 and 5). We have designated these intragenic gene/element associations as "intronic associations". We found 20 introns of genes containing a CE1 repeat. These elements are located within introns, not exons (i.e., protein coding regions or untranslated regions). The CE1 elements showing an intronic association comprised 31.6% (30/95) of the CE1 family. We designated these as "intronic elements".

We identified 120 gene/element associations using a 2 kb window to determine associations (Fig. 5 and Table 1). We used as a 2 kb window as almost all regulatory sequences are found within this distance relative to a gene in the worm (McGhee and Krause, 1997). Of the 120 genes/element associations 30 are intronic associations and 90 are intergenic associations. We then took these genes and classified redundantly them into 37 biological processes, 11 cellular components and 30 molecular functions, based on the gene ontology (GO) terms described in WormBase (Table 1). Although no GO term was shown in many genes, an interesting tendency was presented. As biological process many genes were related to development (20 genes), morphogenesis (16) and growth (13). As molecular function many genes were related to transcription factor activity (19) and protein kinase activity (11). These tendency lead us to suggest that CE1 elements are associated strategically and closely with genes related to transcriptional regulation for development.

3.4. CE1(bs258) forms a hairpin structure and binds to many regulatory proteins

Our detailed analysis focused on CE1(*bs258*). As DNA CE1(*bs258*) has a thermodynamic minimum free energy (mfe) of -65.3 kcal/mol and as RNA it has a mfe of -109 kcal/mol at 20 °C, a standard growth temperature for worms (Fig. 6a,b). Additionally, structural analysis performed on CE1(*bs258*) predicts a large hairpin structure of approximately 240 bp with four major loops along the structure (Fig. 6c). We have shown that CE1(*bs258*) is transcribed. Using RT-PCR to specific primer sets within the gene C46H11.6 (see Section 2.2; Fig. 6d), we detect transcripts of CE1(*bs258*) (see lanes 3 and 6 on gel in Fig. 6d). It also appears that

CE1(*bs258*) RNA is relatively stable, as other intron RNA from the same gene turns over rapidly (see lanes 2 and 5 on gel in Fig. 6d).

Potential transcription factor-binding within CE1(*bs258*) totaled 61 sites (Table 2 and Fig. 6c). Homologous sequences of transcription factor-binding sites embedded within CE1(*bs258*) are 14, 11, 9, 9 and 7 for heat shock factors, zinc finger transcription factors, high mobility group (HMG), fork head transcription factors and helix-turn-helix transcription factors, respectively (Fig. 6c and Table 2). These results suggest that CE1(*bs258*) could be involved in gene transcription. However, the GATA factors-binding site (A/T)GATA(A/T) such as it has been described in *C. elegans* (Shim et al,1995; Hawkins and McGhee, 1995; Gilleard and McGhee, 2001) was not completely conserved within CE1(*bs258*), suggests that the GATA factor could not recognize these sites. Heat shock factors-binding sites are also contained into CeRep16, a member of the *C. elegans* repetitive DNA family (Jones et al, 1986), suggests that some repetitive DNA elements function to gene expression and supports the present hypothesis that the CE1 family can be a "genomic reservoir" for functional DNA sites (see Section 4.2).

3.5. CE1(bs258) regulates the expression of the associated gene C46H11.6

We next wanted to confirm that CE1(*bs258*) could act as an enhancer/transcription regulator *in vivo*. To do this we performed a *gfp* (green fluorescent protein) expression study to monitor expression of the associated gene C46H11.6 in the presence or absence of CE1(*bs258*). Animals bearing the plasmids with CE1(*bs258*), pC46H11BS1888 and pC46H11BB1439 showed *gfp* expression in body-wall muscles (Fig. 7b,c and d), the vulval muscles (Fig. 7e) and anal depressor muscle (Fig. 7f), whereas animals with the plasmids pC46H11BB1164 and pC46H11BE1281, which does not contain CE1(*bs258*), showed no to little expression in these tissues (Fig. 7g). In a separate study we tested CE1(*bs258*) for its ability to drive *lacZ* expression from the heterologous *pes-10* basal promoter (Fig. 7h). In this assay, *pes-10* basal expression with *lacZ* marker was enhanced by CE1(*bs258*) and observed in intestinal cells (Fig. 7i). In contrast there was no expression of the control vector (Fig. 7j).

As shown in Fig. 7k, the expression of *pat-10* is independent of the presence or absence of CE1(*bs258*) (Terami et al., 1999). Three *lacZ* reporter constructs; pTNCZ7600 with CE1(*bs258*), pTNCZ1248 with the partial fragment of CE1(*bs258*) and pTNCZ647 without CE1(*bs258*), were similarly expressed in body-wall muscles. From the expression profile of pTNCZ292, it is clear that the expression of *pat-10* is only controlled by the 212 bp upstream region. These results indicate that CE1(*bs258*) has only an enhancer activity for the intronic associated gene, C46H11.6, and does not function to regulate the adjacent gene, *pat-10*.

4. Discussion

4.1. Structure and function of CE1(bs258)

CE1(*bs258*) only shows low homology to small RNAs, for example, tRNA, 5S RNA and 7SL RNA (data not shown), and is not flanked by direct or inverted repeats as are most transposons (Fig. 3). These observations suggest that CE1(*bs258*) was not a retrotransposon derived from these small RNAs (Ullu and Tschudi, 1984; Okada, 1991; Kapitonov and Jurka, 2003) but more probably it was amplified by non-target site duplications at the DNA level. This assumption suggests that CE1(*bs258*) could be generated by a number of different paths as DNA, much as other repetitive elements have been generated (Rogers, 1985).

As DNA, the CE1(*bs258*) secondary structure can form stem regions, each of which contains many transcription factor target sequences (Table 2 and Fig. 6), suggesting that the target sequences are structurally protected for binding to the transcriptional regulator by the formation of a double stranded DNA. The A+T content of CE1(*bs258*) is compatible with forming a cruciform structure as described (Mizuuchi et al., 1982; Courey and Wang, 1983; Panyutin et al., 1984; Kurahashi et al, 2004). The cruciform structure could be one of the most important features of CE1(*bs258*). CE1(*bs258*) also act at the RNA level (Fig. 6). We found that CE1(*bs258*) was transcribed *in vivo* and a hairpin structure, which was confirmed by RT-PCR and computer prediction, respectively.

In *C. elegans*, most *cis*-regulatory regions have been shown to extend approximately 1 kb upstream of transcriptional start sites (McGhee and Krause, 1997). Although gene/element association have previously been shown for the *Cer* LTR-retrotransposon family (Ganko et al., 2001; Ganko et al., 2003), the CE1 element/gene associations are extremely high (Fig. 5). CE1(*bs258*) is located within the 1 kb upstream region of the body-wall troponin C gene, *pat-10* (Terami et al., 1999) and within the first intron of the C46H11.6 gene, which have a head-to-head gene orientation and are expressed in body-wall muscles. We showed that the C46H11.6 expression in body-wall muscles depended on CE1(*bs258*) whereas the *pat-10* expression was independent of CE1(*bs258*) (Fig. 7). Furthermore CE1(*bs258*) acts as an enhancer was confirmed by the Δ*pes-10*::*lacZ* enhancer assay (Fig. 7). The CE1 elements are associated with a number of different types of genes (Table 1). Although the expression pattern for all these genes is not yet known, a few of these genes have been analyzed in detail. In the AIY thermosensory interneuron, CE1(*C40H5b*-) and CE1(*ZK652*+), for example, are located within the seventh intron of *ttx-3* (Hobert et al., 1997) and the third intron of *ceh-23* (Altun-Gultekin et al., 2001), respectively. 31.6% (30/95) of the CE1 elements are associated within an intron (Fig. 5), suggesting that these elements may control expression of the genes they are embedded within. Studying the genes associated with CE1 elements may shed light on the biological functions of repetitive elements that are distributed throughout the genome.

4.2. Evolution and Biological significance of the CE1 elements

A BLAST search using CE1(*bs258*) as query retrieved many hits only in *C. elegans*, but not in the genome of *C. briggsae*. The only one hit sequence in *C. briggsae* showed a homology to CE1(*bs258*) of 62.0%/211 bp (P value<0.01). Although the CE1(*bs258*) ortholog was conserved within the upstream region of the *pat-10* ortholog (CBG10771) and had a CE1(*bs258*) homology of 62.0% in the *C. briggsae* genome, the C46H11.6 ortholog (CBG09116) was not located in the upstream region of the *pat-10* ortholog. As suggested elsewhere gene pairs are rarely conserved between the two genomes (Felix, 2004; Bando et al., 2005). Moreover, only a few repeat families are shared, suggesting that most elements were acquired after the two sister species diverged or that the two species are undergoing rapid genome evolution (Stein et al., 2003). The *C. elegans* CE1 elements existed only a single copy with a low homology in the *C. briggsae* genome, suggesting that the diffusion to all chromosomes of the CE1 elements occurred, probably, after the two species diverged.

Below we present a model for the function of CE1 elements in *C. elegans* that encompasses the data presented in this paper. On the bases of genomic distribution (Fig. 4 and 5) and the internal target sequence (Table 2) of the CE1 elements, we propose CE1 elements can store transcriptional regulators and act as "genomic reservoir" for sequestering such factors. When the CE1 elements are folded into a cruciform, or hairpin, structure the store potential is expanded and the distance to the adjacent genes is probably contracted. When genes are expressed, it is necessary that there is a structure to store/concentrate these transcriptional regulators in the local area to be gathered later for the formation of the transcriptional complex. Additionally, the distance to the target gene should be reduced. The CE1 elements would thus serve as a localized genomic reservoir for transcription factors, increasing the local concentration at transcribed genes where CE1 elements have a higher than expected association with genes. In *C. elegans*, palindromic repetitive element, CeRep16 containing heat shock factor-binding sites has been known (Jones et al., 1986). A higher-ordered structure of the elements associated with genes is extremely significant for biological functions and evolutionary reasons. This assumption will be confirmed by experiments in future.

Recently, the role of SINEs as powerful reagents for molecular systematics has been recognized (Shedlock and Okada, 2000), even though their molecular function within biological process is largely unknown. The most important features of the CE1 family are the chromosomal distribution and gene/element associations (Fig. 4 and 5). The CE1 elements are a novel repetitive family dispersed along all the chromosomes (Fig. 4) and are closely associated with or within genes (Fig. 5). CE1(*bs258*) has a potential to bind transcriptional regulators *in silico* (Table 2) and actually functions as an enhancer element for the expression of the nearby gene C46H11.6 *in vivo* (Fig. 7). These results suggest that transcription factors shown in table 2 could bind CE1(*bs258*) and function gene expression regulation. A spread distribution of the CE1 elements in the genome also suggests that the CE1(*bs258*)-binding proteins could function on the chromatin structure. We assume that CE1 elements harboring many *cis*-regulatory components of eukaryotic genes were strategically selected for once distributed into the vicinity of a gene; this enabled the CE1 family to

enhance the expression of the adjacent genes and thus contribute to genome organization and genome evolution.

Although our knowledge of the detailed distribution of all of the identified repetitive elements is limited, some repetitive elements are primarily localized to the vicinity of a gene. This should act as a hint as to function. Furthermore, CE1(*bs258*) clearly functions as a gene regulatory element for the expression of the associated gene C46H11.6. We conclude that repetitive elements can significantly contribute to enhance gene expression and to control genome structure.

Acknowledgements

We would express our sincere gratitude to Dr. Naruya Saitou (National Institute of Genetics) and Dr. Norihiro Okada (Tokyo Institute of Technology) for their invaluable discussion, Dr. Donald G. Moerman (University of British Columbia) for proofreading and critical comments. We thank Dr. Andrew Fire (Stanford University) for providing the pPD vector series for *gfp* expression analysis and *lacZ* enhancer assay. *C. elegans* strains were provided by the Caenorhabditis Genetic Center (CGC) which is funded by the NIH National Center for Research Resources (NCRR). This study was supported by a grant from the Ministry of Education, Culture, Sports and Technology of Japan to H.K.

References

Adams, M. D., Celniler, S. E., Holt, R. A. and others (195 co-authors), 2000. The genome sequence of *Drosophila melenogaster*. Science 287, 2185-2195.

Altun-Gultekin, Z., Andachi, Y., Tsalik, E. L., Pilgrim, D., Kohara, Y., Hobert, O., 2001. A regulatory cascade of three homeobox genes, *ceh-10*, *ttx-3* and *ceh-23*, controls cell fate specification of a defined interneuron class in *C. elegans*. Development 128, 1951-1969.

Bando, T., Ikeda, T., Kagawa, H., 2005. The homeoproteins MAB-18 and CEH-14 insulate the dauer collagen gene *col-43* from activation by the adjacent promoter of the spermatheca gene *sth-1* in *C. elegans*. J. Mol. Biol. 348, 101-112.

C. elegans Sequencing Consortium, 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282, 2012-2018.

Courey, A. J., Wang, J. C., 1983. Cruciform formation in a negatively supercoiled DNA may be kinetically forbidden under physiological conditions. Cell 33, 817-829.

Duret, L., Marais, G., Biemont, C., 2000. Transposons but not retrotrasnsposons are located preferentially in regions of high recombination rate in *Caenorhabditis elegans*. Genetics 156, 1661-1669.

Felix, M-A., 2004. Genomes: a helpful cousin for our favourite worm.. Curr. Biol. 14, R75-R77.

Ganko, E. W., Fielman, K. T., McDonald, J. F., 2001. Evolutionary history of *Cer* elements and their impact on the *C. elegans* genome. Genome Res. 11, 2066-2074.

Ganko, E. W., Bhattacharjee, V., Schliekelman, P., McDonald, J. F. 2003. Evidence for the contribution of LTR retrotransposons to *C. elegans* gene evolution. Mol. Biol. Evol. 20, 1925-1931.

Gilleard, J. S., McGhee, J. D. 2001. Activation of hypodermal differentiation in the *Caenorhabditis elegans* embryo by GATA transcription factors ELT-1 and ELT-3. Mol. Cell. Biol. 21, 2533-2544.

Hawkins, M. G., McGhee, J. D. 1995. *elt-2*, a second GATA factor from the nematode *Caenorhabditis elegans*. J. Biol. Chem. 270, 14666-14671.

Heinemeyer, T., Chen, X., Karas, H., Kel, A. E., Kel, O. V., Liebich, I., Meinhardt, T., Reuter, I., Schacherer, F., Wingender, E., 1999. Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. Nucleic Acids Res. 27, 318-322.

Hobert, O., Mori, I., Yamashita, Y., Honda, H., Ohshima, Y., Liu, Y., Ruvkun, G., 1997. Regulation of inerneuron function in the *C. elegans* thermoregulatory pathway by the *ttx-3* LIM homeobox gene. Neuron 19, 345-357.

Hofacker, I. L., 2003. Vienna RNA secondary structure server. Nucleic Acids Res. 31, 3429-3431.

International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of human genome. Nature 409, 860-921.

Jones, D., Russnak, R. H., Kay, Candido, E. P. 1986. Structure, expression, and evolution of a heat shock gene locus in *Caenorhabditis elegans* that is flanked by repetitive elements. J Biol Chem. 261, 12006-12015.

Kapitonov, V. V., Jurka, J., 2003. A novel class of SINE elements derived from 5S rRNA. Mol. Biol. Evol. 20, 694-702.

Ketting, R. F., Fischer, S. E., Plasterk, R. H., 1997. Target choice determinants of the Tc1 transposon of *Caenorhabditis elegans*. Nucleic Acids Res. 25, 4041-4047.

Kurahashi, H., Inagaki, H., Yamada, K., Ohye, T., Taniguchi, M., Emanuel, B. S., Toda, T., 2004. Cruciform DNA structure underlies the etiology for palindrome-mediated human chromosomal translocations. J. Biol. Chem. 279, 35377-35383.

McGhee, J. D., Krause, M. W., 1997. Transcription factors and transcriptional regulation. In: Riddle, D.L., Blumenthal, T., Meyer, B. J., Priess, J. R. (Eds.), *C. elegans* II. Cold Spring Harbor Laboratory Press, Plainview, New York, pp. 147-184.

Mello, C., Fire, A., 1995. DNA transformation. Methods Cell Biol. 48, 451-482.

Mello, C. C., Kramer, J. M., Stinchcomb, D., Ambros, V., 1991. Efficient gene transfer in *C. elegans* extra chromosomal maintenance and integration of transforming sequence. EMBO J. 10, 3959-3970.

Miller, M, A., Cutter, A. D., Yamamoto, I., Ward, S., Greenstein, D., 2004. Clustered organization of reproductive genes in the *C. elegans* genome. Curr. Biol. 14, 1284-1290.

Mizuuchi, K., Mizuuchi, M., Gellert, M., 1982. Cruciform structures in palindromic DNA are favored by DNA supercoiling. J. Mol. Biol. 156, 229-43.

Okada, N., 1991. SINEs. Curr. Opin. Genet. Dev. 1, 498-504.

Page, R. D., 1996. TreeView: an application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. 12, 357-358.

Panyutin, I., Klishko, V., Lyamichev, V., 1984. Kinetics of cruciform formation and stability of cruciform structure in superhelical DNA. J. Biomol. Struct. Dyn. 1, 1311-1324.

Rogers, J. H., 1985. The origin and evolution of retroposons. Int. Rev. Cytol. 93, 187-279.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406-425.

Seydoux, G., Fire, A., 1994. Soma-germline asymmetry in the distributions of embryonic RNAs in *Caenorhabditis elegans*. Development 120, 2823-2834.

Shedlock, A. M., Okada, N., 2000. SINE insertions: powerful tools for molecular systematics. BioEssays 22, 148-160.

Shim, Y. H., Bonner, J. J., Blumenthal, T. 1995. Activity of a *C. elegans* GATA transcription factor, ELT-1, expressed in yeast. J. Mol. Biol. 253, 665-676.

Smit, A. F., 1996. The origin of interspersed repeats in the human genome. Curr. Opin. Genet. Dev. 6, 743-748.

Stein, L. D., Bao, Z., Blasiar, D. and others (36 co-authors), 2003. The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. PLoS Biol. 1, 166-192.

Sulston, J., Hodgkin, J., 1988. Methods. In: Wood, B. P. (Eds.), *The nematode Caenorhabditis elegans*. Cold Spring Harbor Laboratory Press, Plainview, New York. pp. 587-606.

Terami, H, Williams, B. D., Kitamura, S., Sakube, Y., Matsumoto, S., Doi, S., Obinata, T., Kagawa, H., 1999. Genomic organization, expression, and analysis of the troponin C gene *pat-10* of *Caenorhabditis elegans*. J. Cell Biol. 146, 193-202.

Thompson, J. D., Higgens, D. G., Gibson, T. J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment though sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673-4680.

Ullu, E., Tschudi, C., 1984. Alu sequences are processed 7SL RNA genes. Nature 312, 171-172.

Waterston, R. H., Lindbald-Toh, K., Birney, E. and others (222 co-authors), 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520-562.

Wicky, C., Villeneuve, A. M., Lauper, N., Codourey, L., Tobler, H., Muller, F., 1996. Telomeric repeats (TTAGGC)n are sufficient for chromosome capping function in *Caenorhabditis elegans*. Proc. Natl. Acad. Sci. USA. 93, 8983-8988.

Zuker, M., Mathews, D. H., Turner, D. H. 1999. Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide. In: Barciszewski, J., Clark, B. F. C. (Eds.), RNA Biochemistry and Biotechnology. Kluwer Academic Publishers, Dordrecht, pp11-43.

Figure legends

Fig. 1. Distribution of the sequence length of the CE1 elements. The black bars denote copy numbers of the CE1 elements.

Fig. 2. Phylogenetic relationships of 95 members of the CE1 family inferred by the neighbor-joining (N-J) method. Bootstrap values (1,000 replications) are shown above or in the branches. Branch lengths represent nucleotide substitutions per site. The subfamily names are shown on the right. The numbers of the members are indicated by the numbers provided in parentheses. CE1(*bs258*) is labeled as C46H11a+/CE1bs258 and shaded.

Fig. 3. Multiple sequence alignment of the CE1.1 subfamily. (a) Multiple sequence alignment of 24 members of the CE1.1 subfamily. (b) A potential palindrome structure of CE1(*bs258*). The sense and anti-sense strand of CE1(*bs258*) were aligned. Asterisk (*) and dot (.) indicate an identical nucleotide and highly conserved nucleotide. Dash (-) indicates the gap inserted to optimize the alignment. Nucleotides contributing to palindrome sequences within CE1(*bs258*) are shaded.

Fig. 4. Chromosomal distributions of the CE1 element/gene association in the *C. elegans* genome. Chromosomes were divided into three regions (left, centric, and right) marked by vertical hash marks. The centric region is known as high gene density region. Closed and open circles represent the CE1 elements with intronic association and within 2 kb intergenic association, respectively. Open triangles represent the CE1 elements within 5 kb intergenic association. The CE1 elements lacking a gene within 5 kb periphery are marked by closed triangles.

Fig. 5. Distance distributions of the CE1 element/gene association. (a) 225 CE1 element/gene within a 10 kb window were sorted into 1 kb bins. The white bar denotes intronic element contributions (top rectangle of the 0-1000 bp column). Distribution of the CE1 associations upstream and downstream of a gene within a window of 1 bp to (b) 10 kb and (c) 2 kb. A model gene is represented by a small gray bar in the center and is not scaled to size.

Fig. 6. A common DNA/RNA secondary structure prediction for CE1(*bs258*). A mountain plot represents a secondary structure in a plot of height versus position computed with the (a) DNA and (b) RNA parameter set. The curve represents the mfe (minimum free energy) structure. (c) A large hairpin structure with four major loops along the structure is the main feature of this putative common DNA/RNA folds. Eight circles represent potential transcriptional regulators that bind to CE1(*bs258*). (d) CE1(*bs258*) exists stably as both DNA and RNA. Detection of the CE1(*bs258*) transcript by RT-PCR. CE1(*bs258*) exists as both DNA (*) and RNA (**). Each PCR product was analyzed by electrophoresis on a 2% (wt/vol) agarose gel and visualized with 0.5 μg of ethidium bromide/ml. Lane M, 100 bp DNA Ladder; 1-3, genomic PCR products; 4-6, RT-PCR products; and lane C, no DNA contamination into the extracted RNA was checked by RT-PCR handling

without reverse transcriptase, and it was confirmed by no signal.

Fig. 7. The expression patterns of C46H11.6::*gfp* and *pat-10*::*lacZ* reporter genes associated with or without CE1(*bs258*), and *lacZ* enhancer assay of CE1(*bs258*). (a) Schematic structure and summary of expression patterns of C46H11.6::*gfp* are shown. (b-g) C46H11.6::*gfp* reporter plasmids associated with CE1(*bs258*) (pC46H11BS1888 and pC46H11BB1439) were expressed in body-wall (d) and vulval muscles (e) and anal depressor muscle (f), and weakly expressed in intestinal cells, whereas C46H11.6::*gfp* reporter expressions without CE1(*bs258*) (pC46H11BB1164 and pC46H11BE1085) were hardly detectable in these somatic tissues (g). (h) Schematic structure and summary of *lacZ* enhancer assay of CE1(*bs258*) are shown. (i and j) The assay vector pPD95HHCE1bs258 was expressed and *lacZ*-stained in intestinal cells, whereas the control vector pPD95.18 was not expressed in any of tissues on its own. (k) Schematic structure and summary of expression patterns of *pat-10*::*lacZ* are shown (Terami et al., 1999). All reporter constructs of *pat-10*::*lacZ* associated with CE1(*bs258*) (pTNCZ7600), with the partial fragment of CE1(*bs258*) (pTNCZ1248), and without CE1(*bs258*) (pTNCZ647) were expressed in body-wall and vulval muscles and anal depressor muscle. Plus (+) and minus (-) indicate intensity of reporter gene expressions, respectively. The scale bars represent 50 μm.

Figure 1 Takashima et al.

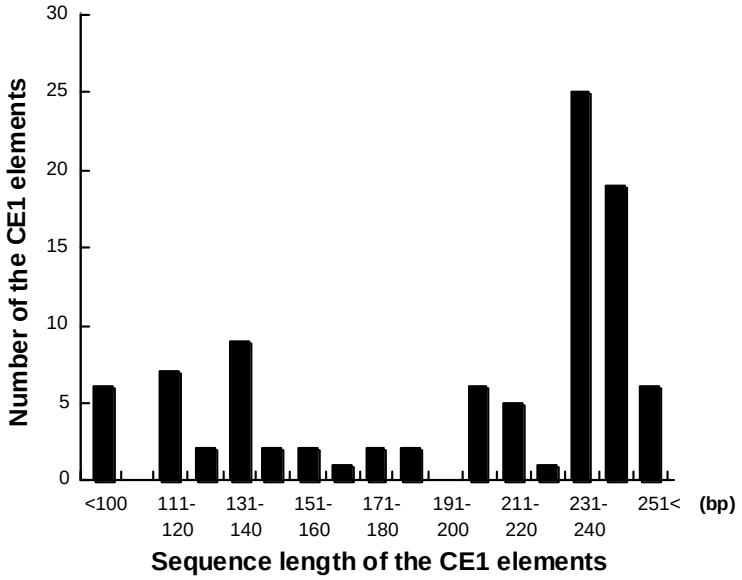

Figure 2 Takashima et al.

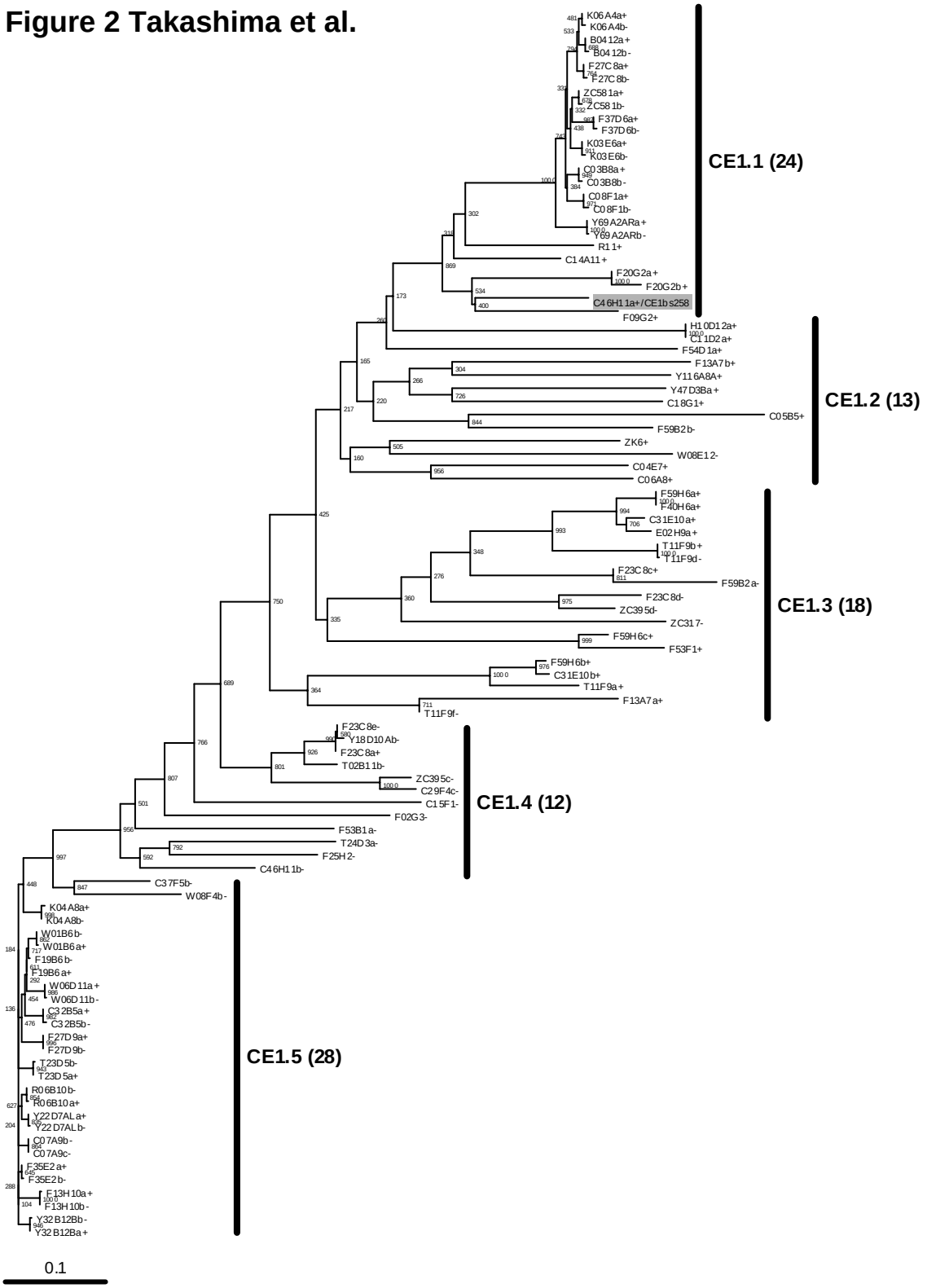

Figure 3 Takashima et al.

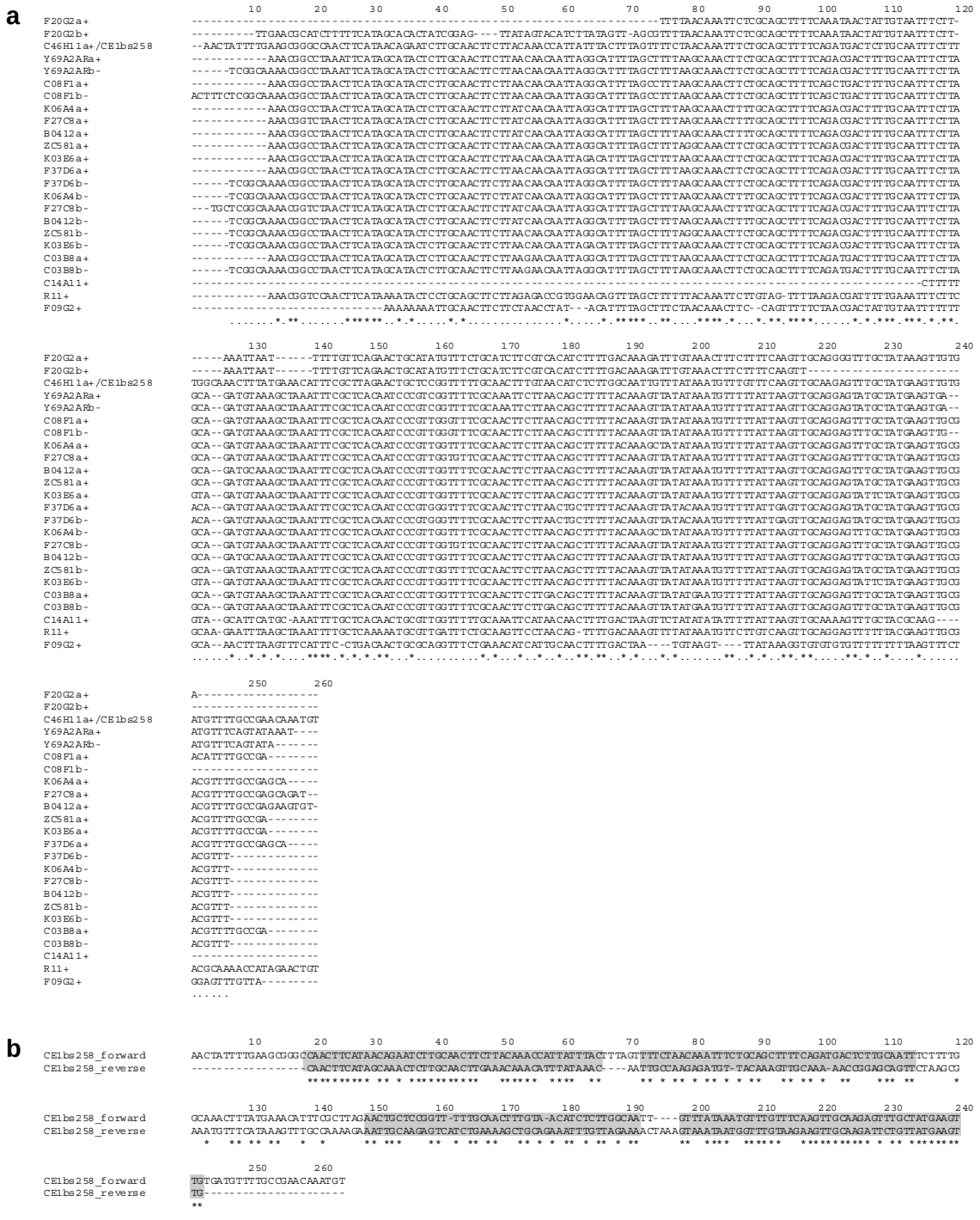

Figure 4 Takashima et al.

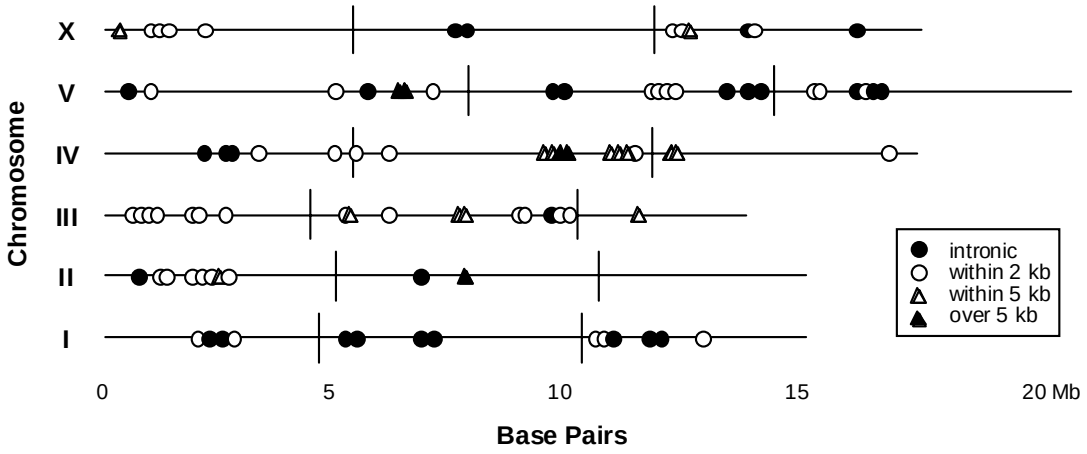

Figure 5 Takashima et al.

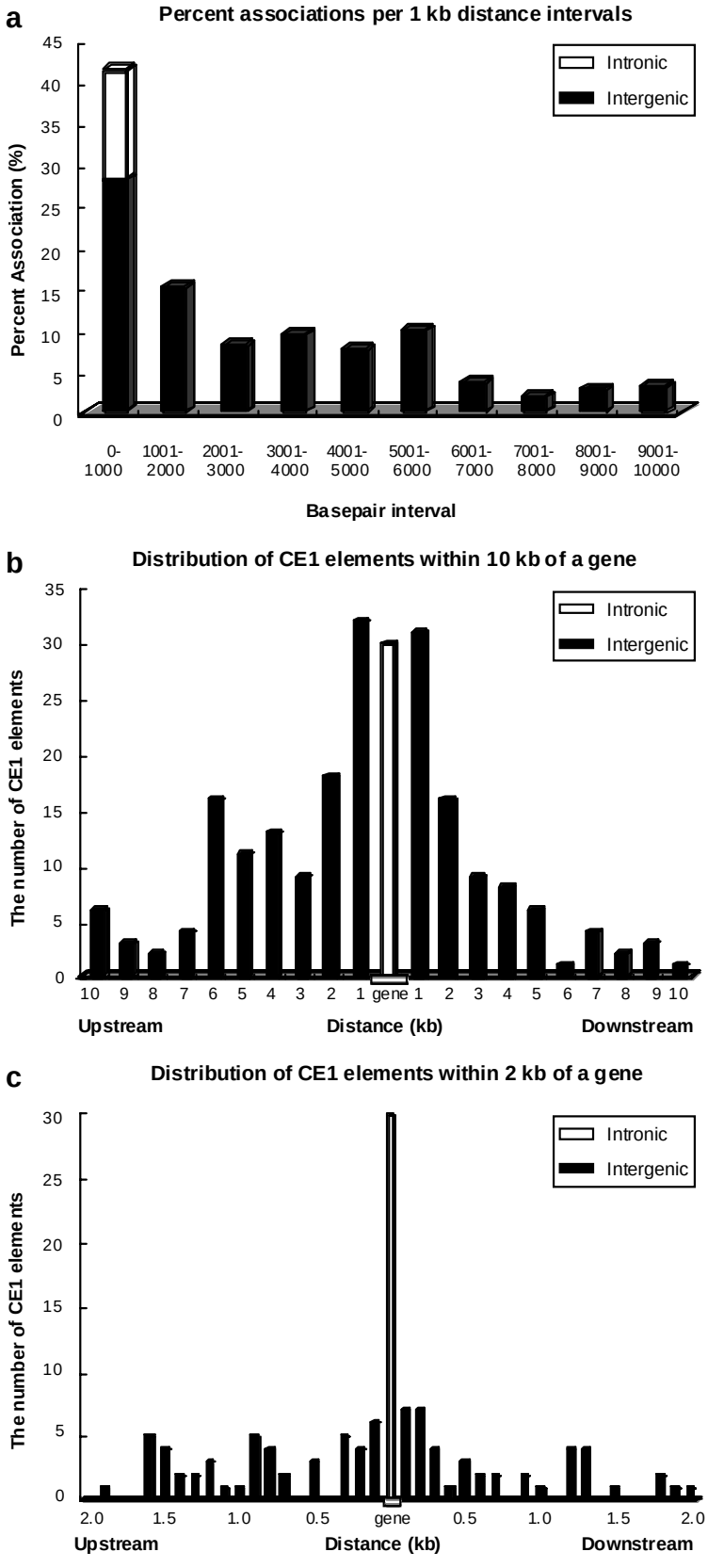

Figure 6 Takashima et al.

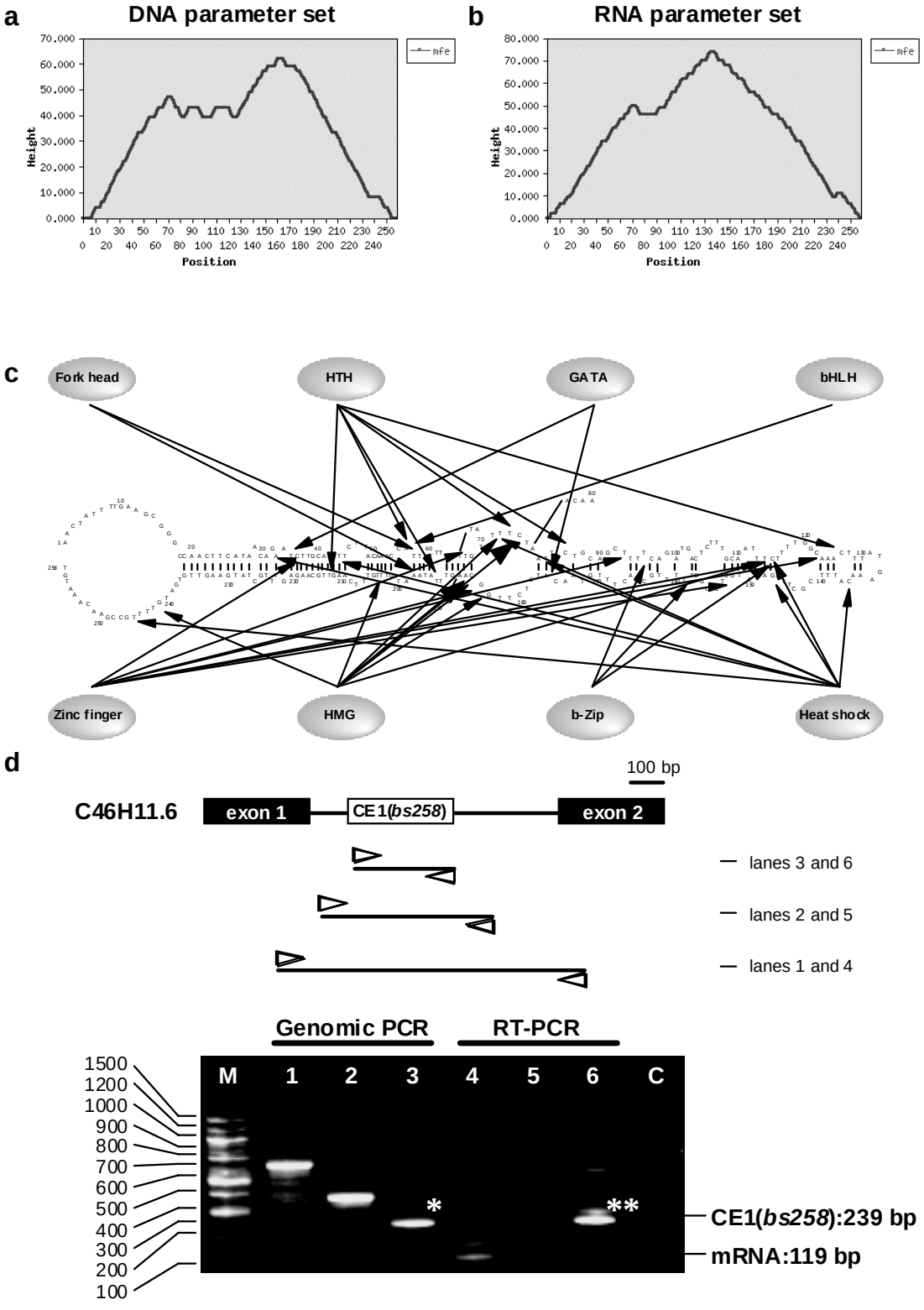

Figure 7 Takashima et al.

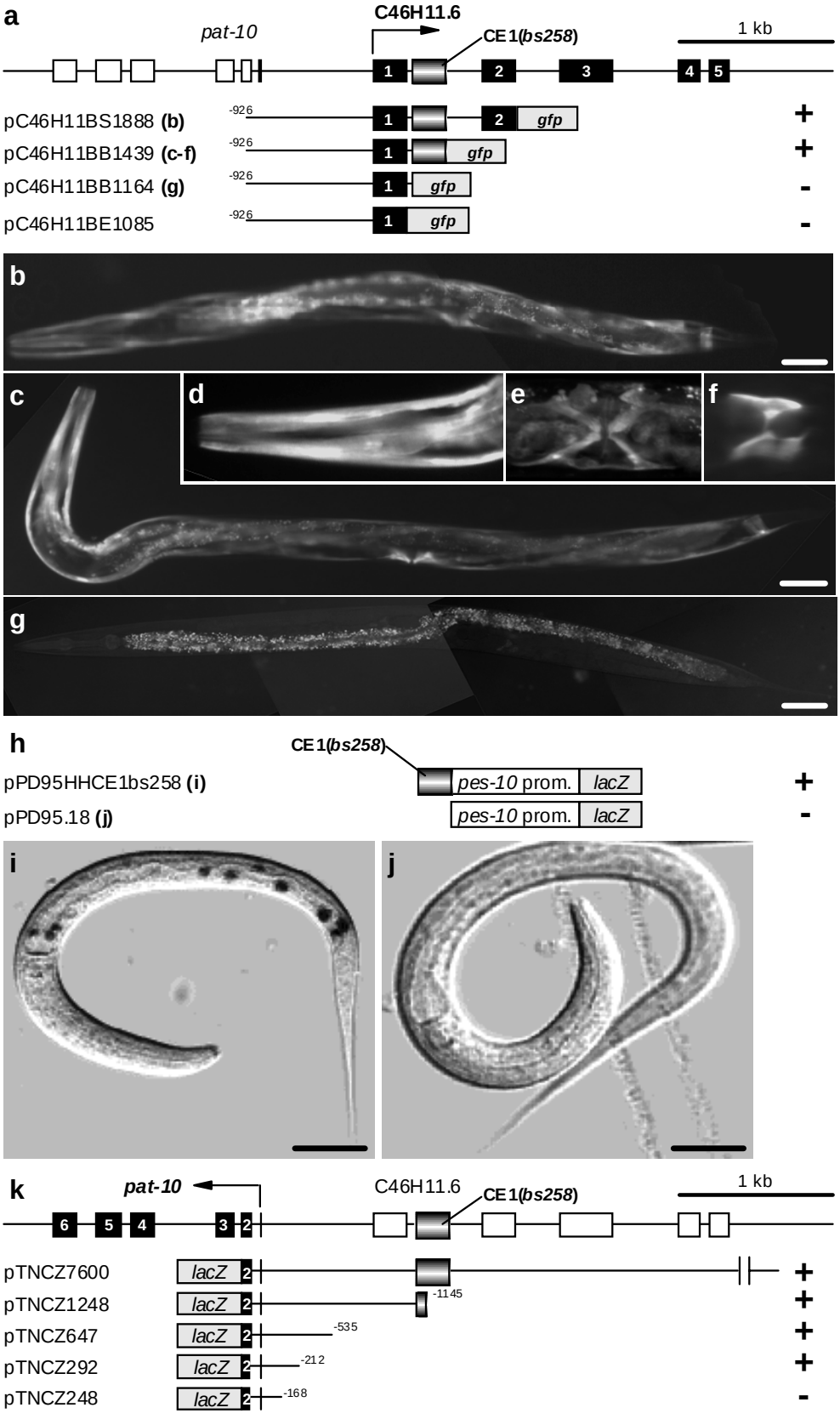

Table 1
Summary of the GO annotation of the CE1-associated genes within a 2 kb window

Biological process		Cellular component		Molecular function	
embryonic development	11	membrane	7	zinc ion binding	5
larval development	7	nucleus	3	transcription factor activity	4
growth	6	integral to membrane	2	sequence-specific DNA binding	4
gametogenesis	6	cytoplasm	2	transporter activity	3
reproduction	5	mitochondrial inner membrane	1	protein serine/threonine kinase activity	3
positive regulation of growth rate	5	intracellular	1	protein kinase activity	3
locomotory behavior	5	gap junction	1	G-protein coupled receptor activity	3
regulation of transcription, DNA-dependent	4	flagellum	1	calcium ion binding	3
physiological process	4	endoplasmic reticulum	1	ATP binding	3
transport	3	contractile fiber	1	transmembrane receptor protein tyrosine kinase activity	2
protein amino acid phosphorylation	3			transferase activity, transferring hexosyl groups	2
post-embryonic body morphogenesis	3			structural constituent of cuticle	2
hermaphrodite genital morphogenesis	3			steroid hormone receptor activity	2
proteolysis	2			protein-tyrosine kinase activity	2
protein binding	2			protein binding	2
positive regulation of body size	2			epidermal growth factor receptor activity	2
phosphate transport	2			DNA binding	2
morphogenesis of an epithelium	2			astacin activity	2
metabolism	2			transferase activity, transferring groups other than amino-acyl groups	1
regulation of cell fate specification	1			structural molecule activity	1
oviposition	1			structural constituent of muscle	1
oocyte construction	1			phosphoinositide binding	1
negative regulation of vulval development	1			metallopeptidase activity	1
muscle thin filament assembly	1			MAP kinase activity	1
muscle contraction	1			ligand-dependent nuclear receptor activity	1
locomotion	1			integrase activity	1
intracellular signaling cascade	1			growth factor activity	1
ergosterol biosynthesis	1			C-8 sterol isomerase activity	1
embryonic cleavage	1			binding	1
DNA integration	1				
determination of adult life span	1				
dauer larval development	1				
cytokinesis	1				
cuticle biosynthesis	1				
ciliary or flagellar motility	1				
cell adhesion	1				
No gene ontology terms	51	No gene ontology terms	61	No gene ontology terms	53

Table 2
Transcription factor-binding sites within CE1(*bs258*)

Potential binding factor	Consensus	Position on CE1(*bs258*)
bHLH		
Dfd; Deformed	NNNNNNATTAMYNNNN	50..65
HTH		
CdxA	WWTWMTR	200..194, 61..55, 81..87, 191..197, 126..132, 70..76
Nkx-2.5/Csx, tinman homolog	CWTAATTG	208..214
HMG		
SRY; sex-determining region Y gene product	AAACWAM	208..202, 72..66, 118..112, 194..188, 244..238, 75..81, 196..185
Sox-5	NNAACAATNN	195..186
Mat1-Mc; M-box interacting with Mat1-Mc	YCNATTGTYW	185..194
GATA		
GATA-1; GATA-binding factor 1	SNNGATNNNN	39..30
GATA-3; GATA-binding factor 3	NNGATARNG	182..173
Zinc finger		
ADR1; alcohol dehydrogenase gene regulator 1	NGGRGK	157..152
BR-C; Broad-Complex Z3	NNNTAAACWARNNNN	76..62
BR-C; Broad-Complex Z4	WWWRKAAASAWAW	199..187
PBF; PBF (MPBF)	ANNWAAAGNNN	71..61
Dof1 / MNB1a - single zinc finger transcription factor	NNNWAAAGCNN	71..61, 98..88, 133..123
Dof2 - single zinc finger transcription factor	NMNNAAAGNNN	71..61, 122..112
Dof3 - single zinc finger transcription factor	SMANAAAAAA	34..43
PBF (MPBF)	NWNWAAAGNGN	71..61
Fork head		
Freac-4; Fork head RElated ACtivator-4	CYWAWGTAAACANWGN	69..54
HNF-3beta; Hepatocyte Nuclear Factor 3beta	KGNANTRTTTRYTTW	195..209, 53..67
HFH-1; HNF-3/Fkh Homolog 1	NAWTGTTTATWT	187..198
HFH-2; HNF-3/Fkh Homolog 2	NAWTGTTTRTTT	197..208
HFH-3; HNF-3/Fkh Homolog 3 (= Freac-6)	KNNTRTTTRTTTA	197..209
HFH-8; HNF-3/Fkh Homolog-8	NNNTGTTTATNYR	187..199
XFD-1; Xenopus fork head domain factor 1	YAWGTAAAYAWWRY	67..54
XFD-2; Xenopus fork head domain factor 2	WNWATAAACAWNNR	199..186
Leusine zipper		
TCF11; TCF11/KCR-F1/Nrf1 homodimers	GTCATNNWNNNNN	103..91
C/EBPalpha; CCAAT/enhancer binding protein alpha	NNATTRCNNAANNN	170..157, 127..114
Heat shock		
HSF; heat shock factor (Drosophila)	AGAAN	75..71, 86..82, 116..112, 47..43, 31..35, 142..138, 146..150
HSF; heat shock factor (yeast)	AGAAN	146..150, 75..71, 86..82, 116..112, 31..35, 47..43, 248..252
Pol II promoter element		
Cap signal for transcription initiation	NCANNNNN	153..146, 87..94, 239..232
AP-1; activator protein 1	NNTGACTCANN	98..108
GCN4	NARTGACTCW	97..106