GBE

Genome Biology and Evolution

SMBE

GC Content Heterogeneity Transition of Conserved Noncoding Sequences Occurred at the Emergence of Vertebrates

Nilmini Hettiarachchi[1,2] and Naruya Saitou[1,2,*]

[1]Department of Genetics, School of Life Science, Graduate University for Advanced Studies (SOKENDAI), Mishima, Japan

[2]Division of Population Genetics, National institute of Genetics, Mishima, Japan

*Corresponding author: E-mail: saitounr@nig.ac.jp.

Accepted: September 9, 2016

Abstract

Conserved non-coding sequences (CNSs) of Eukaryotes are known to be significantly enriched in regulatory sequences. CNSs of diverse lineages follow different patterns in abundance, sequence composition, and location. Here, we report a thorough analysis of CNSs in diverse groups of Eukaryotes with respect to GC content heterogeneity. We examined 24 fungi, 19 invertebrates, and 12 non-mammalian vertebrates so as to find lineage specific features of CNSs. We found that fungi and invertebrate CNSs are predominantly GC rich as in plants we previously observed, whereas vertebrate CNSs are GC poor. This result suggests that the CNS GC content transition occurred from the ancestral GC rich state of Eukaryotes to GC poor in the vertebrate lineage due to the enrollment of GC poor transcription factor binding sites that are lineage specific. CNS GC content is closely linked with the nucleosome occupancy that determines the location and structural architecture of DNAs.

Key words: conserved non-coding sequence, CNS, fungi, invertebrates, vertebrates.

Introduction

Conserved non-coding sequences have been studied for over a decade with findings highlighting their functional importance in organisms. Various studies on vertebrate CNSs (Bejerano et al. 2004; Lee et al. 2010; Takahashi and Saitou 2012; Babarinde and Saitou 2013; Matsunami and Saitou 2013; Saber et al. 2016) and plant CNSs (Kaplinsky et al. 2002; Guo and Moose 2003; Inada et al. 2003; Kritsas et al. 2012; Baxter et al. 2012; Hettiarachchi et al. 2014) reported CNSs to have a regulatory function related to transcription and development. It has also been found that these conserved regions are under purifying selection (Drake et al. 2006; Casillas et al. 2007; Takahashi and Saitou 2012; Babarinde and Saitou 2013). Lee et al. (2010) experimentally verified the function of the "ancient" vertebrate CNSs they identified in their study. Clarke et al. (2012) experimentally verified the functions of two CNSs conserved between vertebrates and invertebrates which have repression function on central nervous system and hindbrain.

Other than their function, some previous studies have also highlighted surprising nucleotide frequency patterns in the flanking regions of animal CNSs. Vavouri et al. (2007) reported a drop of AT content in the flanking regions of CNSs in *Takifugu rubripes, Homo sapiens, Caenorhabditis elegans*, and *Drosophila melanogaster* genomes. Babarinde and Saitou (2013) reported a sharp decrease in GC content of flanking regions towards CNSs.

The AT drop pattern near the boundaries of CNSs has been observed in plant CNSs. Kritsas et al. (2012) reported the AT drop near *Arabidopsis thaliana* and *Brachypodium distachyon* CNSs. Hettiarachchi et al. (2014) reported the AT drop near the boundaries of grass, monocot, and eudicot lineage specific CNSs. Along with the AT drop, an increase in the nucleosome occupancy probability for these CNSs was also reported (Baxter et al. 2012; Hettiarachchi et al. 2014). Seridi et al. (2014) showed a drop of nucleosome occupancy toward the center of the *D. melanogaster* CNSs. So far, there have not been many studies on CNSs and its relations with nucleosome occupancy. Further it has to be noted that because it is documented and known that the nucleosome, which is the repetitive unit of chromatin is inhibitory to transcription factor

binding, experimental evidence is required to verify the functionality and the molecular mechanism by which CNSs located in folded chromatin regions can act as regulatory elements. This aspect of structural architecture of CNSs and their functionality has yet to be fully explained. Apart from the above, Polychronopoulos et al. (2014) showed that distance between CNSs follow a power-law like distribution pattern. They tested this feature for amniotic, mammalian, fly, and worm CNSs and found that this pattern for CNSs remained even after they removed the closest genes to the CNSs from the analysis. Babarinde and Saitou (2016) discovered that the physical distance between one CNS and its nearest gene is often well conserved between mouse and human genomes.

So far, various structural features and distribution patterns have been identified for CNSs. In this study, we focused on four aspects of CNSs; GC content, nucleotide frequency patterns, nucleosome occupancy probability, and substitution pattern. In our previous analysis on lineage specific plant CNSs, the determined CNSs were GC rich (Hettiarachchi et al. 2014). However, Babarinde and Saitou (2013) reported that mammalian CNSs are GC poor. There seem to be a GC content heterogeneity in CNSs of different groups of organisms. This feature might be related to lineage specific nucleotide preferences in regulatory elements. In order to determine where in the line of evolution this GC content heterogeneity for CNSs first appeared, we conducted the analysis on fungal genomes as well as genomes of invertebrates and non-mammalian vertebrates to obtain eukaryote wide perspective of the features of the CNSs.

Materials and Methods

Genome Sequences Compared in the Analysis

Repeat masked genomes of 24 fungi, 19 invertebrates, and 12 non-mammalian vertebrates were downloaded from Ensembl release 78 (see supplementary table S1, Supplementary Material online, for their list). The analyses were focused on the nuclear genomes.

Identification of Lineage Common CNSs

BLAST 2.2.25+ (Altschul et al. 1997) was used for performing homology searches in this study.

Common to invertebrates

The BLASTn search was done for individual orders in group invertebrates. The genomes considered for this analysis include orders Diptera, Lepidoptera, Hymenoptera, and Nematoda. BLASTn search was done with *D. melanogaster* as the query and *D. ananassae* as the subject database. The cut off e-value for the search was 0.001. The alignments without any overlap with a coding region for both query and the subject were considered for subsequent analyses. The best hits selected based on the e-value were searched in *D. persimilis*. Similarly the four mosquito genomes were searched against each other (*A. gambiae* vs. *A. darlingi* and *A. aegypti* vs. *C. quinquefasciatus*) and best hits obtained from the mosquito genomes were searched in best hits obtained for fly genomes to obtain the Diptera common CNSs. Similarly Lepidoptera, Hymenoptera, and Nematode common CNSs were obtained by pairwise chain search.

Common to Non-Mammalian Vertebrates

The BLASTn searches for non-mammalian vertebrates were performed in a similar manner with a cutoff e-value of 0.001. The initial search for birds was done with *G. gallus* as the query and *Meleagris gallapavo* as the subject database. The best hit results were searched in *Anas platyrhynchos*. The best hits from this step were searched in *Taeniopygia guttata* finally to obtain bird common CNSs. The best hits from previous step were searched in the following new species, *Pelodiscus sinensis*, *Anolis carolinensis*, and *Xenopus tropicalis* with the expectation of finding reptilian, reptilian and amphibian shared CNSs. The CNSs that are found in all teleost fishes were found with the same strategy within the group non-mammalian vertebrates.

Common to Fungal Genomes

Fungal common CNSs were determined for nine different orders. Determining lineage common CNSs for fungi follows the same method used for invertebrate and non-mammalian vertebrates.

A depiction of the pipeline used in identifying the lineage common CNSs is provided in supplementary figure S10, Supplementary Material online.

Setting Percentage Identity Cutoff for the CNSs

We used the gene based approach as Babarinde and Saitou (2013) to set the percentage identity cutoff for CNSs (supplementary table S2, Supplementary Material online). This step is needed to identify the conserved regions that might actually be under selective constraint against regions that are not under functional constraint but appear conserved because they did not have enough time to accumulate mutations. For this analysis, we considered only one-to-one orthologous cDNA sequences for the reference genome of a particular group and the most basal species within the same group. For the invertebrates cDNA searches, we considered *D. melanogaster* and *A. darlingi* with respect to Diptera, *Danius plexippus* and *Bombyx mori* for lepidotera, *Atta cephalotes* and *Nasonia vitripennis* for hymenoptera, and *C. briggsae* and *C. japonica* for nematodes. Similarly for non-mammalian vertebrate cDNA searches, we used *G. gallus* and *T. guttata* for birds, *G. gallus* and *P. sinensis* for protein conservation between birds and group chelonia, *G. gallus* and *A. carolinensis*

for all reptiles, and *G. gallus* and *X. tropicalis* to find level of protein conservation between reptiles and amphibians. *Tetraodon nigroviridis* and *Danio rerio* were used to determine the protein conservation level for teleost fish. Same strategy was followed for fungal genomes (supplementary tables S4A–C) after performing BLASTn searches on query and the subject, reciprocal best hits were selected for each of the above mentioned pairs of species. The average percentage identities for the reciprocal best hits were considered as the cutoff threshold for the CNSs in the respective groups.

GC Content and Related Analyses

The GC content of the CNSs were determined and compared with the non-coding GC content of the reference genomes for determining GC content of CNSs and the reference genome. Multiple sequence alignments for the CNSs and the background non-coding regions were constructed with clustalw. Based on these multiple sequence alignments, ancestral sequences were constructed using FASTML (Ashkenazy et al. 2012), and ancestral GC content of CNSs and background non-coding regions were determined. The Multiple sequence alignments of CNSs constructed with clustalw were used to determine their substitution patterns using MEGA6 (Tamura et al. 2013). To obtain the GC content distribution in the flanking regions and inside of CNSs, we analyzed the GC content distribution in 1000 bp flanking regions and the center (20 bp) of the CNSs by a moving window analysis (10 bp window with 1 base step size). The statistical significance of the CNS GC content of lineages were assessed with regards to the non-coding GC content of the reference genome considered for each lineage by *t*-test. The flanking regions of birds, birds-Chelonian shared, reptilian, reptilian and amphibian shared CNSs were used for identification of isochore-like regions. The flanking regions were extended up to 12 kb regions and the classification of isochore like regions were based on Costantini et al. (2006).

Determination of Nucleosome Occupancy Probability

Nucleosome occupancy probability was determined by using the computational model produced by Kaplan et al. (2010) by considering nucleotide preferences in nucleosome regions. The link to the program is http://genie.weizmann.ac.il/software/nucleo_prediction.html. The nucleosome occupancy probabilities for all groups in the study were computed. Initially we extracted a total of 8000 bases from the center of the CNSs. Then the average nucleosome occupancy probability was calculated for each site along the complete length of 8000 bases. The same analysis was done for random samples with the same length and the same number of sequences as the CNSs. The statistical significance of the nucleosome occupancy of CNSs was determined between the CNSs and the random samples of sequences by *t*-test.

Association of Histone Modifications with CNSs

Certain histone modification signals are known to be signatures for some genomic regulatory regions such as promoters and distal enhancers. H3K4Me3 has been found to be highly associated with gene promoter regions (Tserel et al. 2010) whereas H3K4Me1 and H3K27ac are known to be related with nucleosome regions that flank enhancer elements (Heintzman et al. 2009; Creyghton et al. 2010). The regions with these histone modification signals are considered to be active enhancer positions in numerous studies as stated above.

We determined histone modifications associated with zebrafish and nematode CNSs. The coordinates for histone modifications data for *C. elegans* were downloaded from modencode (http://www.modencode.org/) project and zebrafish chromatin signature marks were retrieved from Bogdanović et al. (2012). This analysis was done only for the two above mentioned species as the histone modification data is not available for the rest of the species used in this study.

Predicted Target Genes of CNSs

We considered the closest gene to the CNS as the most plausible likely target gene. For the CNSs that were found inside introns or UTR regions, the gene that they reside in was considered as their target gene. The genes were considered based on the reference genomes used for each group in the analysis. The GO analysis for the target genes were performed using DAVID (The Database for Annotation, Visualization, and Integrated Discovery, version 6.7).

Transcription Factor Binding Site (TF Binding Site) Analysis for the CNSs in Vertebrates

The TF binding site data for human was downloaded from UCSC table browser (GRch37/hg19). A total of 4286829 binding sites were considered for 150 transcription factors. In order to determine the ancestral TF binding sites that are shared across lineages we searched (Blastp) the human TF gene sequences in *A. thaliana*, *O. sativa*, and *C. briggsae* protein coding genes and determined the union of TF genes that are shared across lineages. Also we tried to compare this data with random expectation by searching all the longest transcripts of protein coding genes of human against *A. thaliana*, *O. sativa*, and *C. briggsae* protein coding genes. All Blastp searches were performed with e-value < 0.00001.

Transcription Factor Binding Site Analysis for Plants

Fifty six thousand five hundred and twenty eight transcription factor binding site data for 27 transcription factors were downloaded from supporting data provided by Heyndrickx et al. (2014). The binding site information is based on *A. thaliana* genome. In order to test for TF binding site genes that are shared across lineages we tested the homology of

these sequences in human, chicken, and fugu genomes. And, to compute the random expectation of *A. thaliana* genes that find homology in other lineages, we searched all the protein coding genes of *A. thaliana* in the above-mentioned genomes.

The transcription factor data is provided in supplementary table S4, Supplementary Material online.

Results

Fungi Lineage Common CNSs

We identified 467 Eurotiales, 1,536 Pleosporales, 339 Hypocreales, 201 Schizosaccharomycetales, 2,412 Slerotiniaceae, 288 Magnaporthales, 26 Saccharomycetales, 22,053 Pucciniales, and 669 Ustilaginales CNSs, respectively (table 2). Despite the long divergence times [minimum divergence between two species was 100 million years ago (mya)], the fungal species had considerable number of CNSs. The average lengths of the CNSs were above 50 bp for all orders. The length distributions for the CNSs are provided in supplementary figure S1, Supplementary Material online.

We compared the GC contents for the fungal CNSs along with the reference genome non-coding regions. Eurotiales, Pleosporales, Hypocreales, and Schizosaccharomycetales showed statistically significantly higher GC content than the genomic average (table 1). Sclerotiniacea and Pucciniales had higher GC CNSs compared with the genomic average, but the values were not statistically significant. Ustilaginales CNSs were not significantly different from genomic non-coding GC and Saccharomycetales had too low number of CNSs for any statistical inference. The fungal CNSs showed a pattern of being predominantly GC rich.

Invertebrate Lineage Common CNSs

Invertebrate lineage common CNSs were higher in number compared with fungal CNSs. We identified 1194, 20513, 15573, and 5121 CNSs for orders Diptera, Lepidoptera, Hymenoptera, and Nematoda, respectively. Their length distributions are provided in supplementary figure S2, Supplementary Material online. Lepidoptera, Hymenoptera, and Nematode CNSs were GC rich compared with the genomic average. The Diptera CNSs showed a very low GC content of 20.72% compared with all other orders. Inside the order Diptera, drosophids alone showed a slightly higher but not statistically significant GC value than the genomic average of *D. melanogaster*. Mosquito CNSs were considerably GC poor (29.92%) compared with the reference genome (*Aedes aegypti*). The CNSs that are common to the order Diptera (the CNSs that are shared between drosophids and mosquitoes) showed a pattern of being GC poor. Even when we considered species pairs with roughly the same divergence time between drosophids and mosquitoes, namely, *D. melanogaster–D. ananassae* (44.2 mya; Saisawang and Ketterman 2014) and *Culex quinquefasciatus–Aedes aegypti* (43.3 mya; Marinotti et al. 2013) the pattern of GC remained

Table 1

Statistical Significance of CNS GC Content and Relative GC Content Change Respect to Reference Genome Non-Coding GC Contents for Plants, Fungi, Invertebrates, and Vertebrate CNSs

Lineage	Groups of species used in the analysis	Reference genome content (%)	CNS content (%)	Relative content change	*P*-value
Plants	Eudicots	35.60	36.09	0.013	NS
	Grasses	43.00	45.33	0.054	2.859E−121
	Monocots	43.00	47.16	0.096	4.326E−06
	Angiosperm	35.60	37.34	0.048	NS
Fungi	Eurotiales	46.55	52.21	0.121	0.004
	Pleosporales	48.78	51.70	0.059	1.352E−05
	Hypocreales	45.21	49.28	0.090	0.035
	Schizosaccharomycetales	33.35	35.70	0.070	0.001
	Pucciniales	45.67	47.77	0.045	0.014
	Sclerotiniaceae	40.81	41.72	0.022	0.03
	Magnaporthales	51.52	51.08	−0.008	NS
	Ustilaginales	50.72	50.62	−0.002	NS
Invertebrates	Diptera	40.41	20.72	−0.487	5.910E−193
	Hymenoptera	33.19	37.26	0.122	7.700E−72
	Lepidoptera	31.65	39.17	0.237	0.000
	Nematoda	36.31	41.23	0.135	3.20E−06
Vertebrates	Birds	41.26	41.33	0.002	NS
	Birds and chelonia	41.26	39.50	−0.042	4.100E−32
	Reptiles	41.26	37.81	−0.083	6.120E−184
	Reptiles and amphibian	41.26	37.42	−0.093	1.600E−110
	Teleost fish	45.35	42.99	−0.052	2.960E−94
	Mammals	43.3	36.34	−0.161	4.596E−32

Table 2
Number of Lineage Common CNSs, Mean, and Mode Lengths of CNSs Identified for Groups Fungi, Invertebrate, and Non-Mammalian Vertebrates

Lineage	Groups used in the analysis	Number of species	Number of CNSs	Mean lengths of CNSs (bp)	Mode length (bp)
Fungi	Eurotiales	4	467	89.64	61
	Pleosporales	3	1536	142.89	42
	Hypocreales	4	339	74.34	35
	Schizosaccharomycetales	3	201	73.47	64
	Sclerotiniaceae	2	2412	150.44	50
	Magnaporthales	2	288	103.03	84
	Saccharomycetales	2	26	111.69	55
	Pucciniales	2	22053	85.44	53
	Ustilaginales	2	669	122.46	46
Invertebrates	Diptera	7	1194	25.46	23
	Hymenoptera	4	15573	58.41	34
	Lepidoptera	3	20513	105.26	51
	Nematoda	5	5121	61.90	37
Non-mammalian vertebrates	Birds	4	25011	261.33	37
	Birds and chelonia	5	8007	323.72	48
	Reptiles	6	4477	300.22	46
	Reptiles and amphibian	7	2305	253.54	98
	Teleosts	5	15168	116.38	65

the same where drosophid CNSs were GC rich and mosquito CNSs were GC poor. Therefore the very low GC content observed for Diptera group could not have solely been brought about by the age of the CNSs. Even though the extant CNS GC content is low, our ancestral GC content analysis showed that the ancestral CNSs of Diptera had been considerably GC richer (about 26%) compared with the current CNSs GC content.

Non-Mammalian Vertebrate Lineage Common CNSs

We found 25,011 CNSs that are commonly shared among the four bird species used in the analysis. Between bird and *Pelodiscus sinensis* (Chinese softshell turtle), there are 8,007 shared CNSs. Reptilian, reptile and amphibian shared CNSs were 4477 and 2305, respectively. Fifteen thousand one hundred sixty eight CNSs were identified for the five teleost species used in the analysis (the length distributions are provided in supplementary fig. S3, Supplementary Material online). All non-mammalian vertebrate CNSs had GC content lower than the genomic average of the reference genomes (*Gallus gallus* and *Tetraodon nigroviridis*) that were considered in the analysis.

The Pattern in GC Content Transition

Hettiarachchi et al. (2014) identified the lineage specific plant CNSs to be GC rich. Here we found that fungi and invertebrate CNSs are also predominantly GC rich. In this study, we observed a transition from GC rich state in plants, fungi, and invertebrate CNSs to GC poor state in non-mammalian vertebrate CNSs (table 1). Babarinde and Saitou (2013) reported that mammalian CNSs are GC poor. This shows some change in nucleotide preference that occurred in the vertebrate lineage with regards to CNSs or putative regulatory elements compared with other eukaryotes. Close observation of Diptera group revealed that the relative GC content change for mosquitoes is much higher than for drosophids (fig. 1), and the relatively high GC content change in Diptera group was brought about mainly due to mosquito and drosophid shared CNSs (tables 1 and 3). The distribution of the average GC content for CNSs of different lineages along with the reference genome GC content is shown in supplementary figure S4, Supplementary Material online. Here, for better representation purpose, the drosophid and mosquito GC contents are shown separately.

Nucleosome Occupancy and GC Content Distribution for CNSs

The CNSs are located in numerous structurally diverse regions. One reason for this scenario is the diverse nucleotide composition in these regulatory regions. This diversity in GC content of the CNSs in turn results in positioning them in open chromatin or heterochromatin regions. Nucleosome occupancy is known to be related to regulation of genes (Jiang and Pugh 2009). Further nucleosome occupancy has been reported to be directly associated with nucleotide composition (Kaplan et al. 2010; Tillo and Hughes 2009; Gaffney et al. 2012). It has been reported that GC rich sequences have a high propensity to form nucleosomes whereas lower GC regions will prefer an open chromatin conformation (Warnecke et al. 2008; Washietl et al. 2008). We found that Lepidoptera, Hymenoptera, and Nematode CNSs showed high nucleosome occupancy probabilities. These CNSs tend to be located in a

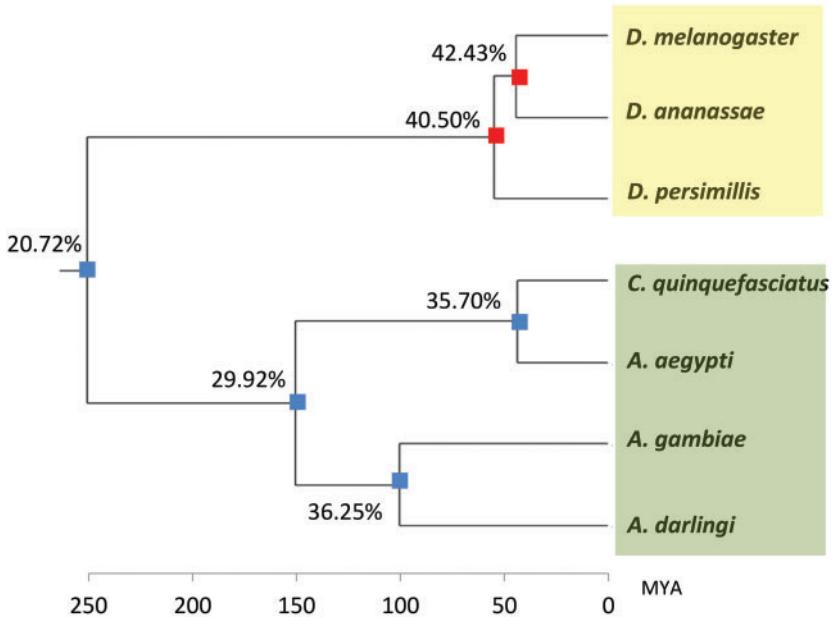

Fig. 1.—GC content change in order Diptera. In these analyses, the order Diptera contain drosophids (yellow background) and mosquitoes (green background). The common CNSs identified at each node was considered for determining the GC content and red squares correspond to high GC CNSs whereas blue squares correspond to low GC CNSs. The actual GC contents are provided in the phylogenetic tree for groups' drosophids, mosquitoes, and order Diptera.

Table 3

Relative GC Content Change of Drosophids, Mosquitoes, and Order Diptera with Respect to the Reference Genome

Lineage	CNS GC content (%)	Reference genome GC (%)	Relative GC content change of CNSs
Drosophid	40.50	40.41	0.0022
Mosquitoes	29.92	38.38	−0.2200
Diptera	20.72	40.41	−0.4872

well-positioned nucleosome region. Diptera CNSs showed lower nucleosome occupancy which goes in line with their very low GC content.

The Diptera CNSs are specially located in very low GC open chromatin regions margined by two well positioned nucleosomes. This structurally constrained conformation may also be related to the functional aspect of CNSs which are yet to be fully elucidated. These GC-rich flanks have been documented as container sites by Valouev et al. (2011). Kundaje et al. (2012) showed that the container sites are a distinct feature of transcription factor binding sites. To perform regulatory functions, the transcription factors should be able to identify the correct binding site or motif from a large arena of similar regions. Therefore, it can be assumed that the accurate finding of the correct binding site may lie in the sequence features along with the unique structural architecture of the binding sites. These high GC flanks might be essential for keeping the proper structural architecture of the actual transcription factor binding sites which are embedded inside these regions.

We found that vertebrate CNSs were GC poor and they showed a lower nucleosome occupancy probability. The teleost CNSs showed a low nucleosome occupancy toward the center of the CNSs where the CNS center was flanked by two nucleosome regions. One interesting feature observed was that the bird, bird and Chelonian shared, reptilian, reptilian and amphibian shared CNSs are flanked by long stretches of high GC regions or in other words stretches of GC rich isochore-like regions making CNSs the only low GC genomic area in that region (supplementary fig. S5, Supplementary Material online and fig. 2). One clear observation was that the CNSs were flanked by highly GC rich isochore-like regions (H2 and H3) compared with the random samples (classification of flanking regions is based on Costantini et al. [2006]). The statistical significance was assessed by *t*-test. H2 and H3 regions in bird, bird and Chelonian shared, reptilian, reptilian and amphibian shared CNSs were significantly higher compared with random sequences (bird—1.00E−05, 1.00E−05; bird and Chelonian shared—1.00E−05 [H2], 1.00E−05 [H3];

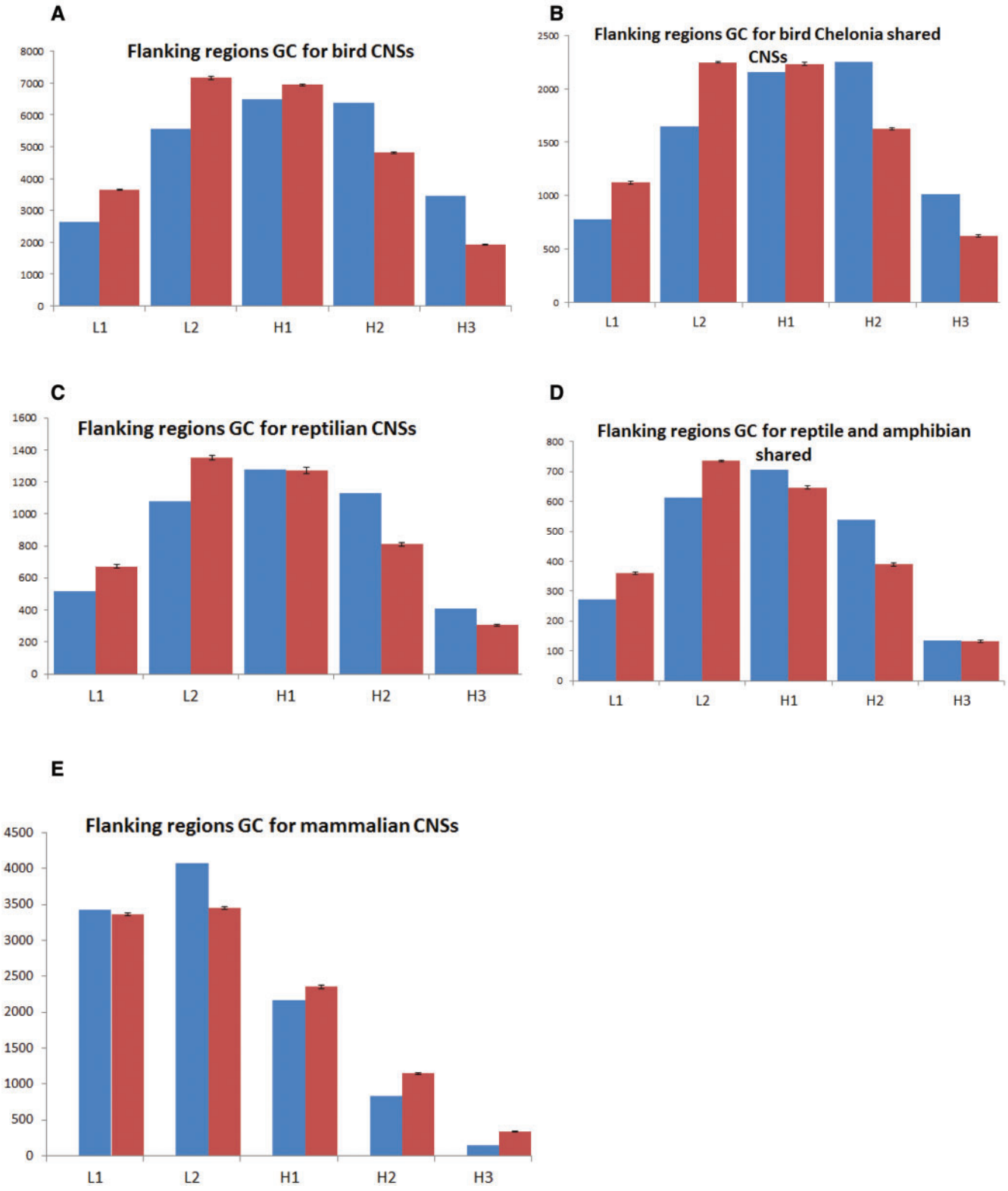

Fig. 2.—Isochore distribution for the flanking regions of CNSs and random samples. Red bars represents the random samples and the blue bars represent the flanking regions of CNSs. The X axis gives the groups into which isochores are classified into (<37-L1, 37 > =<40-L2, 41 > =<45-L3, 46 > =<52-L4, > =53-L3). The Y axis gives the frequency of each isochore segment in flanking regions of CNSs and random samples. H2 and H3 regions

reptilian—2.30E−05 [H2], 5.10E−05 [H3]; reptilian and amphibian shared—1.40E−05 [H2], NS [H3]). The low GC regions were more abundant in random samples than for the flanking regions of CNSs. The isochore analysis for mammalian CNSs showed that mammalian CNS flanking regions are predominantly L2 type (fig. 2*E*) and the mammalian CNSs were also located in high GC flanking regions compared with CNSs (supplementary fig. S5*E*, Supplementary Material online). Figure 3*A–D* gives examples of the GC content distribution for bird, reptilian, nematode, and fungi CNSs, respectively. The GC content distribution for reptile and bird CNSs show a decline in GC inside the CNSs compared with the surrounding flanking regions. Nematode (invertebrates) CNSs and Eurotiales (fungi) show an elevation in the GC

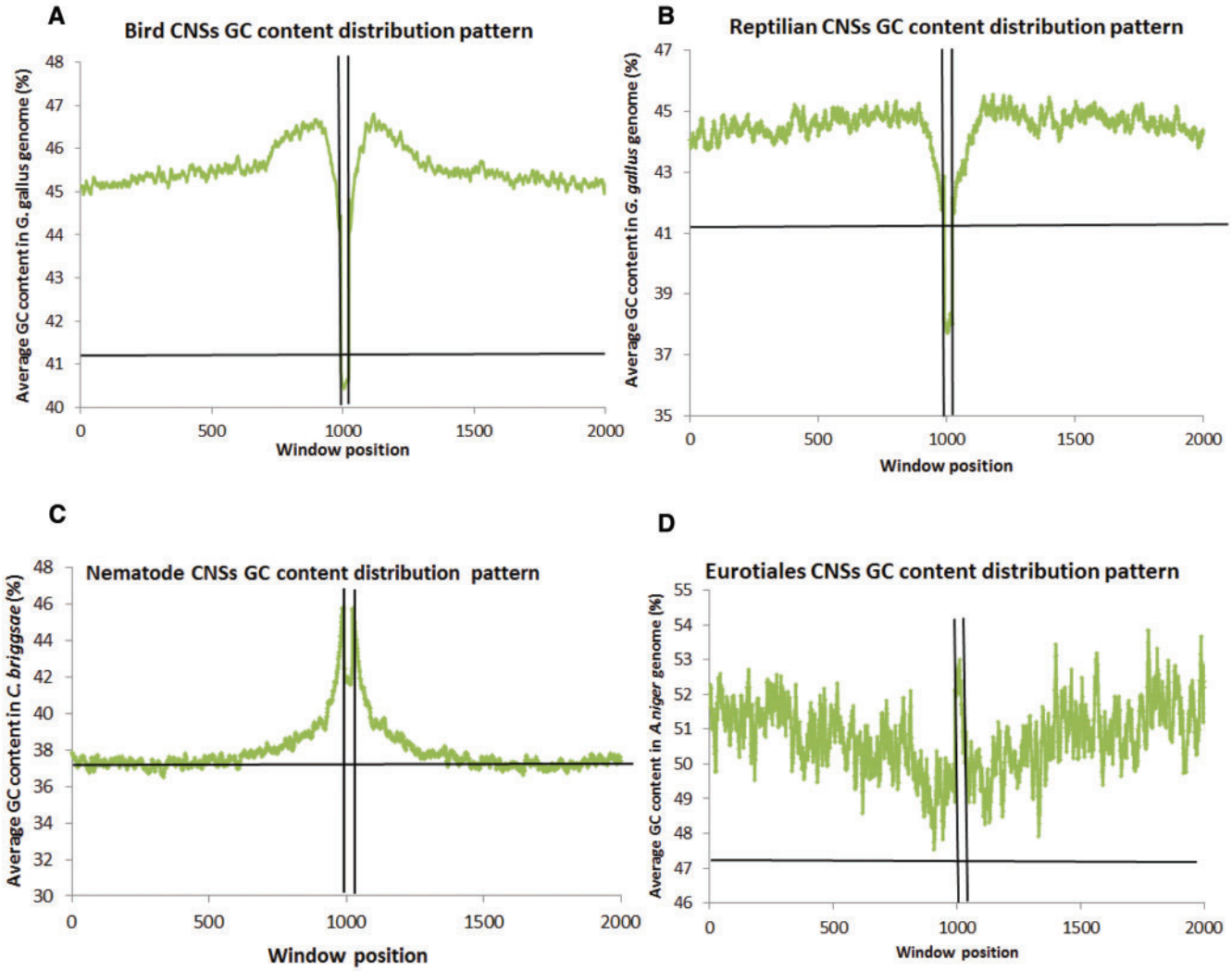

Fig. 3.—GC content distribution of the CNSs across the center of CNSs and the flanking regions. The 1000th nucleotide position corresponds to the center of the CNSs. The horizontal back line represents the level of non-coding GC content of the reference genome. The vertical lines represent the margins of the flanking regions. (*A*) bird CNSs GC distribution. (*B*) Reptilian CNSs GC distribution. (*C*) Nematode CNSs GC content distribution. (*D*) Eurotiales CNSs GC content distribution.

Fig. 2.—Continued

in bird, bird and Chelonian shared, reptilian, reptilian and amphibian shared (no significance in H3 regions in reptilian and amphibian shared CNSs compared with random samples) CNSs were significantly higher compared with random sequences (bird—1.00E-05, 1.00E-05; bird and Chelonian shared—1.00E-05 [H2], 1.00E-05 [H3]; reptilian—2.30E-05 [H2], 5.10E-05 [H3]; reptilian and amphibian shared—1.40E-05 [H2], NS [H3] at 95% confidence P value < 0.05).

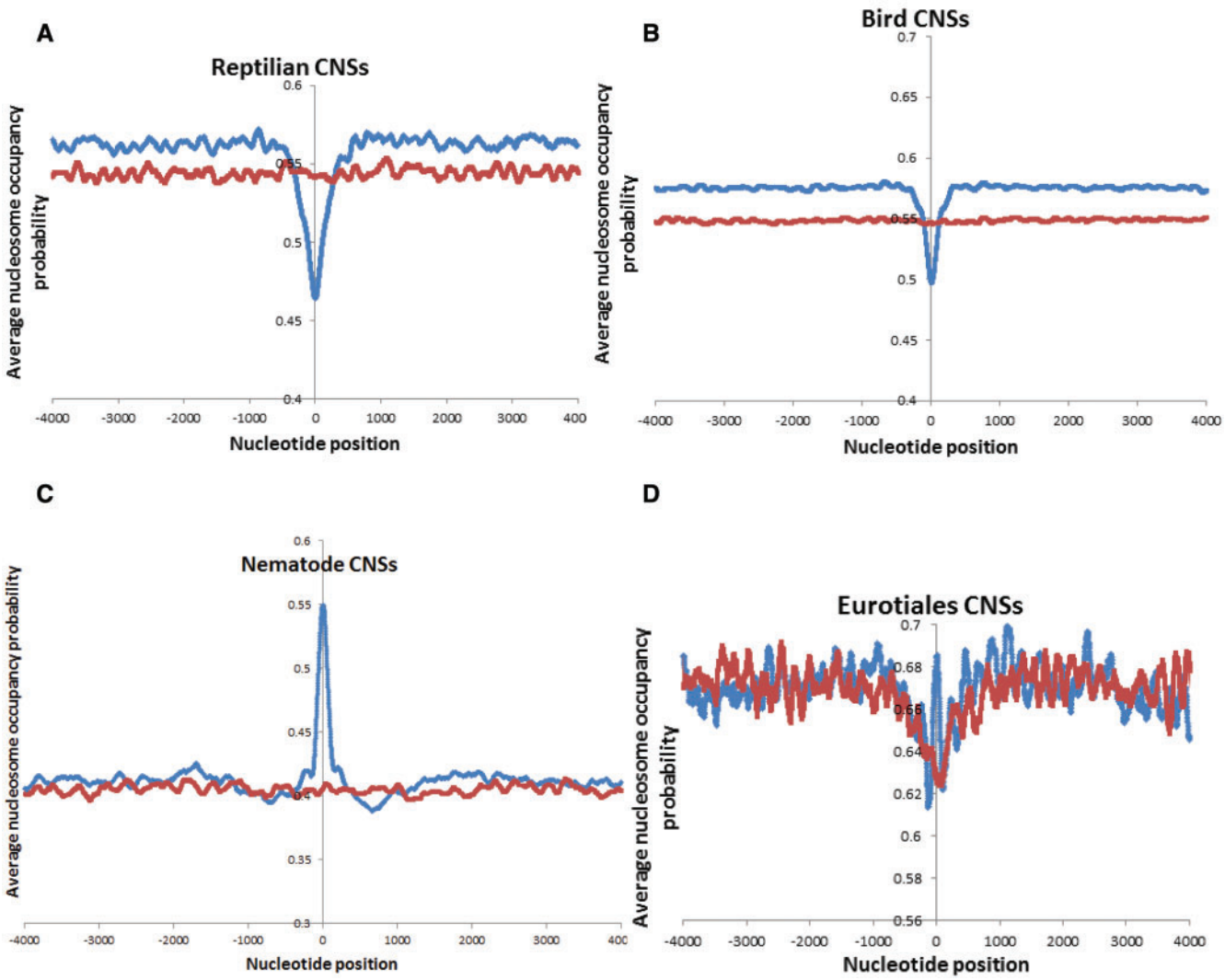

Fig. 4.—Nucleosome occupancy probability for the CNSs of different lineages. The 0th position represents the center of the CNSs. 8000bp flanks were considered for this analysis. The blue and red colors represent the nucleosome occupancy for CNSs and random samples, respectively. (*A*) Reptilian CNSs average nucleosome occupancy probability, (*B*) bird CNSs average nucleosome occupancy probability, (*C*) Nematode CNSs average nucleosome occupancy probability, and (D) Eurotiales CNSs average nucleosome occupancy probability. (The statistical significance between the center 1000 bases of CNSs and center 1000 bases of random sequences for Reptilian, bird, Nematode, and Eurotiales are 2.37E-81, 0.00, 9.78E-73, and 3.87E-22, respectively.).

content of the CNSs in comparison to the surrounding genomic regions. The GC content distribution patterns for fungi, invertebrates and vertebrate CNSs are given in supplementary figure 6A–*C*, Supplementary Material online, respectively.

Figure 4*A–D* provides nucleosome occupancy probability distributions for reptile, bird, nematode, and Eurotiale CNSs, respectively. The nucleosome occupancy follows a similar pattern where the low GC reptile CNSs have low nucleosome occupancy probability whereas nematode and Eurotiales, with high GC CNSs, show a higher nucleosome occupancy probability compared with the flanking regions. The random samples in all instances showed no apparent elevation or decline in the nucleosome occupancy when compared with the CNSs.

Histone Modifications Related with CNSs

The histone modification signals related to the CNSs were determined for nematode and teleost conserved regions found in the analysis. Histone modifications have been studied in many organisms and they are regarded as gene regulatory signals which enable genes to be activated or repressed.

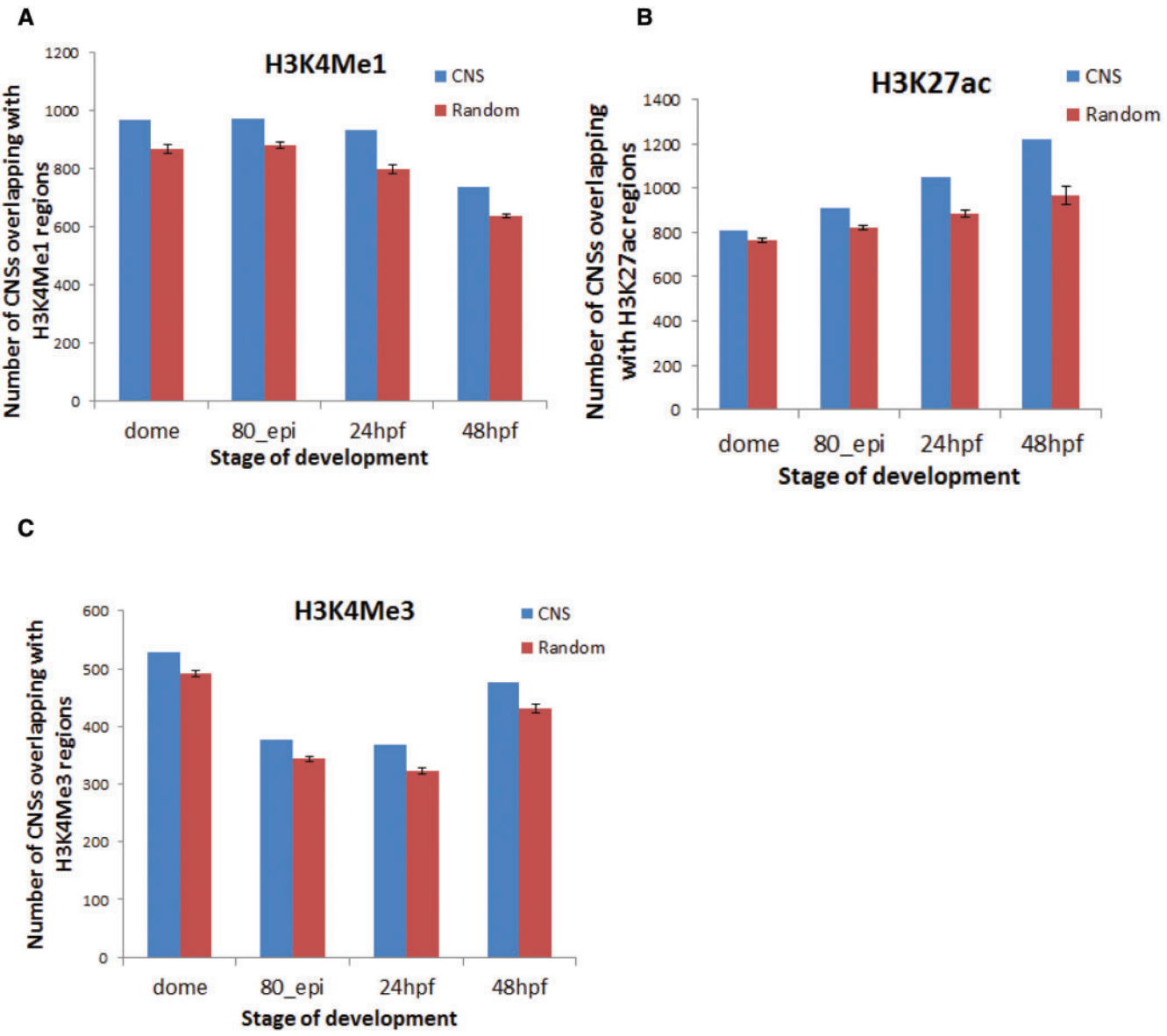

FIG. 5.—Chromatin modification signals overlapping with teleost fish CNSs at different development stages. (*A*) CNSs overlapping with H3K4Me1 regions. (*B*) CNSs overlapping with H3K27ac regions. (*C*) CNSs overlapping with H3K4Me3 regions. The statistical significance of the number CNSs overlapping with chromatin modification signals was compared with random samples overlapping chromatin modification signals with *t*-test at 95% confidence.

Certain histone modification signals are thought to have a direct impact on regulation of genes (Hebbes et al. 1994; Kalmykova et al 2005; Buenrostro et al. 2013).

The teleost CNSs (tested with zebra fish chromatin modification data) showed an overrepresentation for H3K27ac and H3K4Me1 with respect to random expectation (fig. 5). The over-representation of H3K27ac and H3K4Me1 signals in the CNSs compared with random samples of sequences picked from the genome was statistically significant at $P < 0.0001$ (*P*-values for H3K4Me1 regards to dome, 80_epi, 24hpf, 48hpf—9.60E−05, 1.00E−05, 1.00E−05, 1.00E−05; *P*-values for H3K27ac regards to dome, 80_epi, 24hpf, 48hpf—1.88E−03, 1.50E−05, 1.00E−05, 9.90E−05). These modification signals are known to be related to active enhancer regions (Creyghton et al. 2010). H3K4Me3 which is related with promoter regions also showed an overrepresentation in CNSs compared with random samples of sequences (*P*-values for H3K4Me1 regards to dome, 80_epi, 24hpf, 48hpf—1.19E−04, 3.10E−05, 1.00E−05, 1.09E−04).

The number of CNSs that overlapped with H3K27ac regions advances with the development stage whereas for H4K4Me1, many CNSs overlap with this chromatin mark at very early stage of development such as the dome stage. Similarly, nematode CNSs also showed an overrepresentation

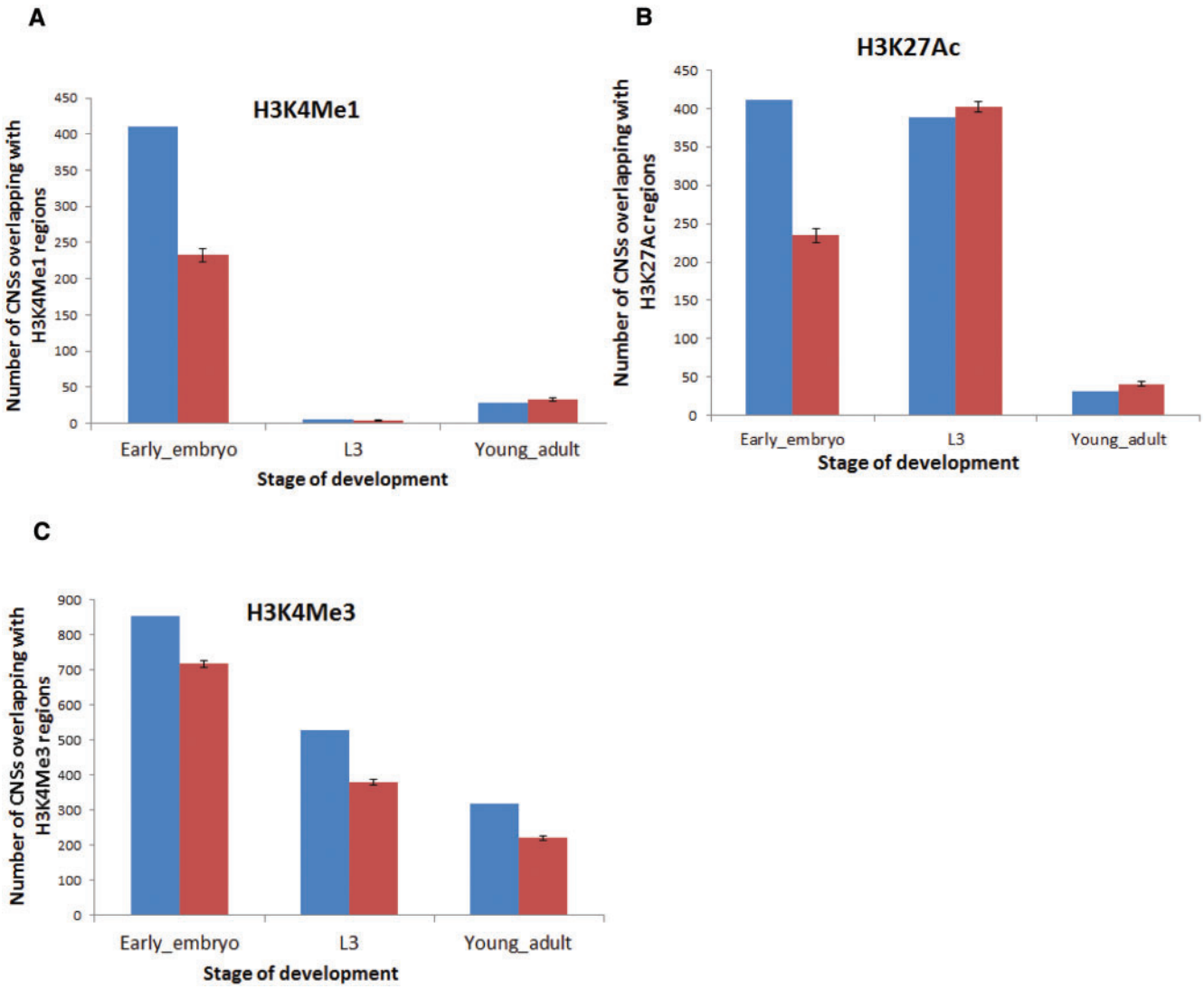

FIG. 6.—Chromatin modification signals overlapping with Nematode CNSs at different development stages. (*A*) CNSs overlapping with H3K4Me1 regions. (*B*) CNSs overlapping with H3K27ac regions. (*C*) CNSs overlapping with H3K4Me3 regions. The statistical significance of the number CNSs overlapping with chromatin modification signals was compared with random samples overlapping chromatin modification signals with *t*-test at 95% confidence.

of H3K4Me1 and H3K27ac signals at early embryo stage compared with random samples of sequences (statistical significances for H3K4Me1, H3K27ac at embryo stage are 4.40E−05, 4.20E−05, respectively). Many CNSs overlapped with H3K4Me3 regions during early development stage when compared with later stages such as L3 stage or the young adult (fig. 6).

The Predicted Target Genes for CNSs and Functional Classification

The closest genes to the CNSs were considered as the likely target gene. The target genes were determined based on the reference genomes used in the analysis. The gene ontology analysis was performed based on the likely target genes. The GO analysis for the likely target genes showed that Diptera and nematode CNS-associated genes were enriched in transcription regulation and DNA binding (supplementary tables S4A and B, Supplementary Material online). This analysis was only done for these two invertebrate groups as the gene ontology data was not available for other groups of interest. The GO analysis for bird CNS-associated genes showed a pattern similar to invertebrates. The highly enriched GO terms were related to regulation of transcription, regulation of RNA metabolic processes, and DNA binding, whereas the most underrepresented was related to certain receptor classes and enzyme activity related proteins. The GO analysis could only

be determined for the Diptera, nematode, and teleosts due to limited availability of data for other genomes (supplementary table S3, Supplementary Material online).

Transcription Factor Binding Site Analysis for Vertebrates

The human transcription factor binding site data from UCSC were used as the vertebrate reference in determining the binding site characteristics for non-mammalian vertebrates. We found that many of the binding sites for ubiquitous transcription factors are GC rich (fig. 7*A*). For example, SMC3, SP1, USF2, and ATF1 are known to be ubiquitous transcription factors, and they have higher than average genomic GC content: 49.87%, 51.57%, 51.95%, and 49.74%, respectively.

Transcription factors such as SP1, ATF specifically bind to sites that are overrepresented in housekeeping gene promoter regions (Farrél et al. 2007). The GC content for CNSs we found for non-mammalian vertebrates were lower than the genomic average for the non-coding region of the reference genome. We found that many of the underrepresented binding sites in CNSs are GC rich and are also related to ubiquitous transcription factor binding sites. The tissue specific binding sites were over-represented compared with the ubiquitous binding sites. For example, the overrepresented binding sites such as SETDB1 are found to be related to repression of genes encoding developmental regulators, and help to maintain the embryonic stem cell state (Bilodeau et al. 2009). Another overrepresented binding site was for transcription factor ZNF263 which is related to regulation of cell growth, cell differentiation, and development (Okubo et al. 1995).

Several other transcription factors such as MAFs, MAFF, and MAFK are considered to be important in gene expression in mammals (Kannan et al. 2012). MAFF and MAFK correspond to two low GC binding sites which are overrepresented in our set of CNSs for non-mammalian vertebrates. MAFK has a function in neuronal differentiation and in general, Maf family transcription factors are considered to be regulators of tissue specific gene expression (Kataoka 2007). In general, one evident pattern we observed is that the GC poor binding sites are related to tissue or stage specific gene expression, regulation of transcription and development, whereas high GC binding sites correspond to ubiquitous activity.

Transcription Factor Binding Sites in Plants

We focused on the *A. thaliana* transcription factor binding sites. Out of the 26 binding sites we analyzed, the low GC binding sites (11/26) mostly seem to be related to ubiquitous activity and the high GC sites appear to facilitate binding of transcription factors related to transcription and development and tissue and state restricted expression (fig. 7*B*). The binding sites such as AP2, SEP3, SOC1, LFY, GL1, and GLT1 are related to ubiquitous expression (Lee and Lee 2010), and their GC were lower than the genomic average GC for the non-coding region. Binding sites such as PRR5, PRR7, PIF4, PIF5, TOC1, and FUS3 have larger than average GC content (42.07%, 42.51%, 38.77%, 39.72%, 46.56%, and 36.84%), and are related to transcription regulation and development.

This finding is opposite to what was observed for the vertebrate CNSs. That is, the vertebrates binding sites related to regulation of transcription and development and tissue specific expression were GC poor. It appears that vertebrates and plants have formulated different sequence preferences when it comes to binding sites related to tissue specific and ubiquitous binding. This in a way explains the heterogeneity we observed for high GC plant CNSs and GC poor vertebrate CNSs. Even though the CNSs seem to be related to same GO function in both lineages, their sequence preferences differ. These analyses were restricted to *A. thaliana* and human, as a comprehensive transcription factor binding site data is not available for other organisms under study. We did not perform the transcription factor binding site overrepresentation analysis for plant CNSs, as the number is too low for any statistical inference.

The GC Content Transition in the Vertebrate Lineage

In order to explain the origin of GC content heterogeneity, we decided to look into the evolutionary dynamics of the transcription factors. Because we found that the CNSs are overrepresented in TF binding sites (supplementary fig. S7, Supplementary Material online), we obtained a cue that the transcription factor binding site evolution may have played an important role in the evolutionary dynamics of CNS GC content. Upon closer examination of the TF binding site data for vertebrates (human as the reference), we found that ubiquitous TF binding sites have higher GC than the tissue or stage specific binding sites when compared with the genomic GC content (supplementary fig. S8, Supplementary Material online). In comparison, we observed that plant TF binding sites follow the opposite pattern, whereby plant tissue specific binding sites were GC rich and the ubiquitous binding sites were found to be GC poor when compared with the genomic average (supplementary fig. S9, Supplementary Material online). This heterogeneity in GC content with regards to the CNSs in lineages might be attributable to the tissue or stage specific TFs. After testing 150 vertebrate TFs shared with *A. thaliana*, *O. sativa*, and *C. briggsae* protein coding genes, we found that vertebrate tissue or stage specific TFs are more lineage specific than the ubiquitous ones. The tissue specific TFs were significantly less shared than the ubiquitous TFs (supplementary fig. S8, Supplementary Material online). The lineage specific features should come from tissue specific TFs that are not shared across lineages. Because vertebrates evolved more tissue specific TFs that are lineage specific which are GC poor, in turn more binding sites that are GC poor, they show the characteristic feature of low GC CNSs among other lineages.

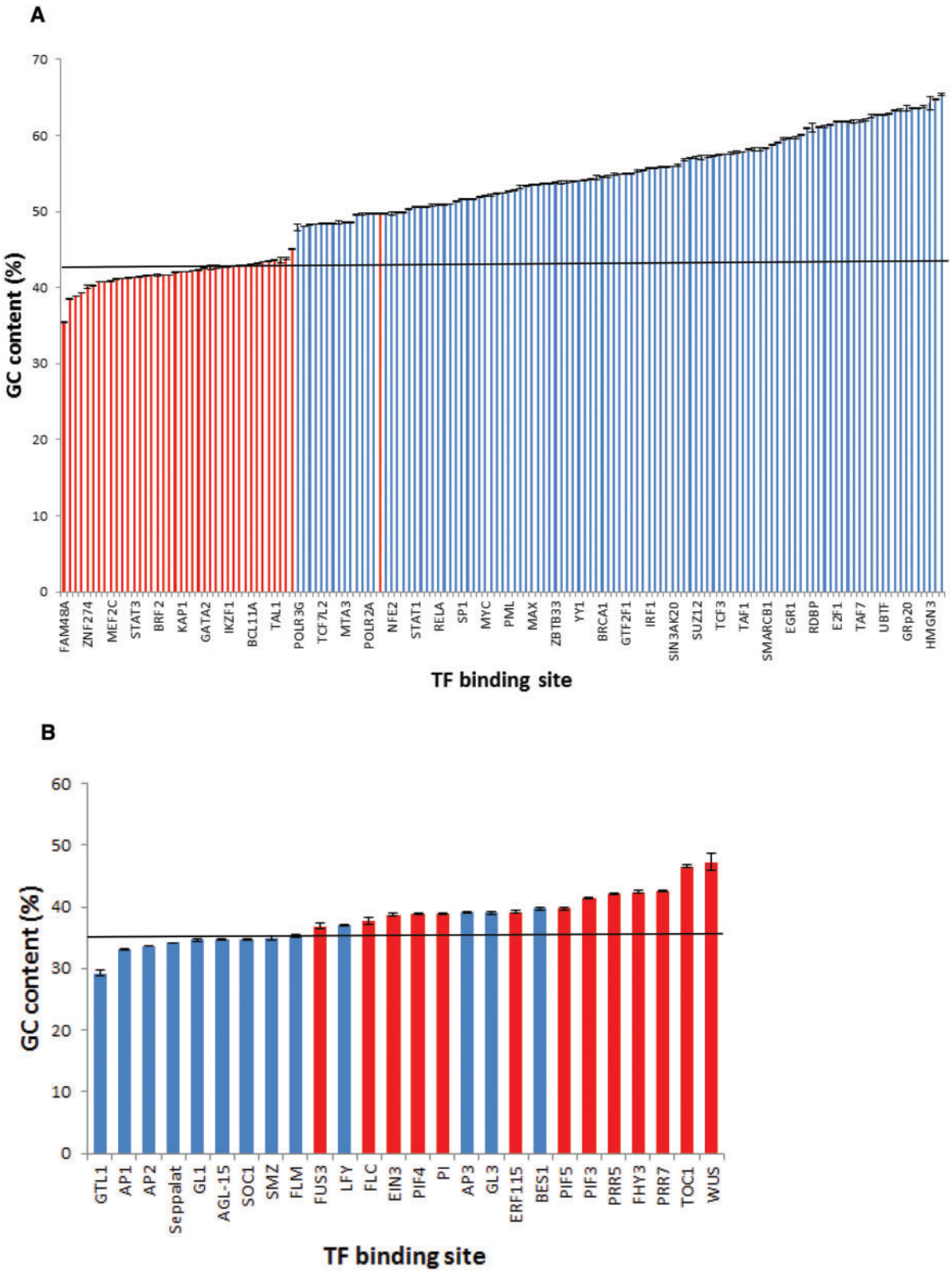

Fig. 7.—The GC content distribution for vertebrate and plant TF binding sites. (*A*) The GC content distribution for vertebrate (human) TF binding sites. The ubiquitous binding sites are shown in blue bars and tissue specific binding sites are represented in red color. The black horizontal line represents the non-coding genomic GC content for human genome (43.3%). (*B*) GC content distribution for plant (*A. thaliana*) TF binding sites. Ubiquitous binding sites are presented in blue color bars, whereas the tissue specific binding sites are provided in red color. The black horizontal line represents the non-coding genomic GC content for Arabidopsis genome (35.6%).

As for a high GC lineage, we expected that they should have evolved higher GC TFs with high GC binding sites. To this end, we tested the *A. thaliana* TF genes against all human, chicken, and fugu annotated genes. However, we found no significant difference in conservation of ubiquitous and tissue specific TFs with regards to plants. This implies that underrepresentation of tissue specific TFs among conserved TFs is a specific feature to the vertebrate lineage (supplementary fig. S8, Supplementary Material online).

Discussion

We identified lineage-common conserved non-coding regions for fungi, invertebrate, and non-mammalian vertebrate genomes that are shared among organisms in a particular lineage. The GC contents for the CNSs differed among lineages. The fungal and invertebrate CNSs were generally GC rich, whereas non-mammalian vertebrate CNSs were GC poor. In our previous study (Hettiarachchi et al. 2014), we found that plant CNSs are GC rich, showing similar characteristic as fungal and invertebrate CNSs. However, Babarinde and Saitou (2013) reported that mammalian CNSs were GC poor and their result shows similarity to our non-mammalian vertebrate CNSs. This low GC content in CNSs therefore appears to be a general feature that is shared by vertebrates. This result suggests that there seems to be a sudden transition of GC content preference in CNSs or in other words potential regulatory elements from plants, fungi, and invertebrates to vertebrates. Next question we addressed was what could be the plausible reason for this observed transition. To this end, we tried to determine the sequence properties of different transcription factors. We discovered in vertebrates the transcription factor binding sites related to tissue specific expression, transcription, and development are GC poor and binding sites for ubiquitous transcription factors such as SP1, NRF1, and E2F6 are GC rich.

The above-mentioned pattern we observed for GC content for transcription factor binding sites in vertebrates switched when we examined the transcription factors of plants. The ubiquitous transcription factors for plants seemed to be GC poor whereas the tissue specific and plant development and transcription regulation associated transcription factors are GC rich. This goes well in line with our observation for plant CNSs (Hettiarachchi et al. 2014) being GC rich and the identified CNSs in our previous analysis showed a highly enriched GO for transcription regulation and development.

In order to explain the origin of heterogeneity we decided to look into the evolutionary dynamics of the transcription factors. After examining the TF binding site data for vertebrates (human as the reference) we found that ubiquitous TF binding sites have higher GC compared with tissue specific binding sites with respect to the genomic GC content. In contrast, plant TF binding sites followed an opposite pattern where the tissue specific binding sites were GC rich and the ubiquitous binding sites were found to be GC poor. And, this heterogeneity in GC content with regards to the CNSs in lineages might be attributable to the tissue specific TFs, and in fact, we found that vertebrate tissue specific TFs are more lineage specific than the ubiquitous ones and that the tissue specific TFs were significantly less shared than the ubiquitous TFs. The lineage specific features should come from tissue specific TFs that are not shared across lineages. At this point, it becomes evident that because vertebrates evolved more tissue specific TFs that are lineage specific which are GC poor and more binding sites that are GC poor, the overall CNS GC content is also low.

Even though we expected that the high GC lineages to have evolved more high GC TFs with high GC binding sites, the scenario was different. After testing *A. thaliana* TF genes against human, chicken, and fugu for all annotated genes, we found no significant difference in conservation of ubiquitous and tissue specific TFs with regards to plants. This means that underrepresentation of tissue specific TFs among conserved TFs is a specific feature to the vertebrate lineage.

We also found that the GC content of the CNSs had a direct relation to the location of the CNSs in the genome. The low GC CNSs showed a higher probability to be located in open chromatin regions whereas high GC CNSs tend to be located in clearly positioned nucleosomes. The location of the CNSs is important because the CNSs located in open chromatin regions are easier to be accessed by the transcription factors, whereas the ones with high nucleosome occupancy are harder to be accessed due to the coiled nature of the region. This structural architecture of the CNSs should have a direct impact on the binding of the proteins to that region. Several studies have found that some binding sites actually require being located in coiled nucleosome regions for its proper regulation. Cirillo and Zaret (1999) reported that HNF3 liver enriched transcription factor binds to albumin gene enhancer region which is clearly located inside a nucleosome region. Binding of HNF3 stabilizes the nucleosome position which results in a very stable binding complex. Similarly TP53 transcription factor, which is known to be a roadblock against cancer, also binds to a high nucleosome occupancy region (Lidor Nili et al. 2010). Experimental evidence is needed to determine the function and the binding properties of the CNSs.

The GO analysis for nematode, Diptera, and bird CNSs target genes showed a high enrichment for regulation of transcription and development, transcription factor activity, and DNA binding, among others. This goes in line with numerous findings that are already documented and experimentally verified that CNSs are related to transcription regulation and development (Sandelin et al. 2004; Woolfe et al. 2005; Vavouri et al. 2007; Babarinde and Saitou 2013; Hettiarachchi et al. 2014).

In conclusion, we observed a GC content heterogeneity of the CNSs belonging to diverse lineages. Specifically we found that plants, invertebrates, and fungi were predominantly GC

rich and non-mammalian vertebrates were GC poor. The GC poor feature in the vertebrate lineage is specific due to its employment of many low GC binding sites that specifically occurred in this lineage. The Diptera CNSs behaved similar to vertebrate CNSs with respect to GC content. This is one enigmatic aspect in our findings which still needs further investigation.

Supplementary Material

Supplementary figures S1–S10 and tables S1–S4 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank Dr Kakutani Tetsuji, Dr Ikeo Kazuho, and Dr Timothy A. Jinam for their valuable suggestions and comments. We also wish to extend our sincere gratitude to Dr Isaac Adeyemi Babarinde for their constant support and comments. This study was partially supported by Grants-in-Aid for Scientific Research from Ministry of Education, Culture, Sports, Science and Technology, Japan 26251040 given to N.S.

Literature Cited

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Ashkenazy H, et al. 2012. FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic Acids Res. 40 (Web Server issue): W580–W584.

Babarinde IA, Saitou N. 2013. Heterogeneous tempo and mode of conserved noncoding sequence evolution among four mammalian orders. Genome Biol Evol. 5:2330–2343.

Babarinde IA, Saitou N. 2016. Genomic locations of conserved noncoding sequences and their proximal protein-coding genes in mammalian expression dynamics. Mol Biol Evol. 33:1807–1817.

Baxter L, et al. 2012. Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants. Plant Cell 24:3949–3965.

Bejerano G, et al. 2004. Ultraconserved elements in the human genome. Science 304:1321–1325.

Bilodeau S, Kagey MH, Frampton GM, Rahl PB, Young RA. 2009. SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. Genes Dev. 23:2484–2489.

Bogdanović O, et al. 2012. Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. Genome Res. 22:2043–2053.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10:1213–1218.

Casillas S, Barbadilla A, Bergman CM. 2007. Purifying selection maintains highly conserved noncoding sequences in Drosophila. Mol. Biol. Evol. 24:2222–2234.

Cirillo LA, Zaret KS. 1999. An early developmental transcription factor complex that is more stable on nucleosome core particles than on free DNA. Mol. Cell 4:961–969.

Clarke SL, et al. 2012. Human developmental enhancers conserved between deuterostomes and protostomes. PLoS Genet. 8:e1002852.

Costantini M, Clay O, Auletta F, Bernardi G. 2006. An isochore map of human chromosomes. Genome Res. 16:536–541.

Creyghton MP, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts development state. Proc. Natl. Acad. Sci. U S A. 107:21931–21936.

Drake JA, et al. 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat. Genet. 28:223–227.

Farrél D, Bellora N, Mularoni L, Messeguer X, Mar Albà M. 2007. Housekeeping genes tend to show reduced upstream sequence conservation. Genome Biol. 8:R140.

Gaffney DJ, et al. 2012. Controls of nucleosome positioning in the human genome. PLoS Genet. 8:e1003036.

Guo H, Moose SP. 2003. Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. Plant Cell 15:1143–1158.

Hebbes TR, Clayton AL, Thorne AW, Crane-Robinson C. 1994. Core histone hyper acetylation co-maps with generalized DNase I sensitivity in the chicken β-globin chromosomal domain. EMBO J. 13:1823–1830.

Heintzman ND, et al. 2009. Histone modifications at human enhancers reflect global cell- type-specific gene expression. Nature 459:108–112.

Hettiarachchi N, Kryukov K, Sumiyama K, Saitou N. 2014. Lineage-specific conserved noncoding sequences of plant genomes: their possible role in nucleosome positioning. Genome Biol. Evol. 6:2527–2542.

Heyndrickx KS, Van de Velde J, Wang C, Weigel D, Vandepoele K. 2014. A functional and evolutionary perspective on transcription factor binding in *Arabidopsis thaliana*. Plant Cell 26:3894–3910.

Inada DC, et al. 2003. Conserved noncoding sequences in the grasses. Genome Res. 13:2030–2041.

Jiang C, Pugh BF. 2009. Nucleosome positioning and gene regulation: advances through genomics. Nat. Rev. Genet. 10:161–172.

Kalmykova AI, Nurminsky DI, Ryzhov DV, Shevelyov YY. 2005. Regulated chromatin domain comprising cluster of co- expressed genes in *Drosophila melanogaster*. Nucleic Acids Res. 33:1435–1444.

Kannan MB, Solovieva V, Blank V. 2012. The small MAF transcription factors MAFF, MAFG and MAFK: Current knowledge and perspectives. Biochem Biophys Acta. 1823:1841–1846.

Kaplan N, et al. 2010. Nucleosome sequence preferences influence in vivo nucleosome organization. Nat. Struct. Mol. Biol. 17:918–920.

Kaplinsky NJ, Braun DM, Penterman J, Goff SA, Freeling M. 2002. Utility and distribution of conserved noncoding sequences in the grasses. Proc. Natl. Acad. Sci. U S A. 99:6147–6151.

Kataoka K. 2007. Multiple mechanisms and functions of maf transcription factors in the regulation of tissue-specific genes. J. Biochem. 141:775–781.

Kritsas K, et al. 2012. Computational analysis and characterization of UCE-like elements (ULEs) in plant genomes. Genome Res. 22:2455–2466.

Kundaje A, et al. 2012. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. Genome Res. 22:1735–1747.

Lee A, Kerk SY, Tan YY, Brenner S, Venkatesh B. 2010. Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. Mol. Biol. Evol. 28:1205–1215.

Lee J, Lee I. 2010. Regulation and function of SOC1, a flowering pathway integrator. J Exp Bot. 61:2247–2254.

Lidor Nili E, et al. 2010. p53 binds preferentially to genomic regions with high DNA-encoded nucleosome occupancy. Genome Res. 20:1361–1368.

Marinotti O, et al. 2013. The genome of Anopheles darlingi, the main neotropical malaria vector. Nucleic Acids Res. 41:7387–7400.

Matsunami M, Saitou N. 2013. Vertebrate paralogous conserved noncoding sequences may be related to gene expressions in brain. Genome Biol. Evol. 5:140–150.

Okubo K, Itoh K, Fukushima A, Yoshii J, Matsubara K. 1995. Monitoring cell physiology by expression profiles and discovering cell type-specific genes by compiled expression profiles. Genomics 30:178–186.
Polychronopoulos D, Sellis D, Almirantis Y. 2014. Conserved noncoding elements follow power-law-like distributions in several genomes as a result of genome dynamics. PLoS One 9:e95437.
Saber MM, Adeyemi BI, Hettiarachchi N, Saitou N. 2016. Emergence and evolution of Hominidae-specific coding and noncoding genomic sequences. Genome Biol. Evol. 8:2076–2092.
Saisawang C, Ketterman AJ. 2014. Micro-plasticity of genomes as illustrated by the evolution of glutathione transferases in 12 Drosophila species. PLoS One 9:e109518.
Sandelin A, et al. 2004. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics 5:99.
Seridi L, Ryu T, Ravasi T. 2014. Dynamic epigenetic control of highly conserved noncoding elements. PLoS One 9:e109326.
Takahashi M, Saitou N. 2012. Identification and characterization of lineage-specific highly conserved noncoding sequences in mammalian genomes. Genome Biol. Evol. 4:641–657.
Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30:2725–2729.
The UniProt Consortium. 2014. UniProt: a hub for protein information. Nucleic Acids Res. 43:D204–D212.
Tillo D, Hughes TR. 2009. G+C content dominates intrinsic nucleosome occupancy. BMC Bioinformat. 10:442.
Tserel L, et al. 2010. Genome-wide promoter analysis of histone modifications in human monocyte-derived antigen presenting cells. BMC Genomics 11:642.
Valouev A, et al. 2011. Determinants of nucleosome organization in primary human cells. Nature 474:516–520.
Vavouri T, Walter K, Gilks WR, Lehner B, Elgar G. 2007. Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. Genome Biol. 8:R15.
Wang A, et al. 2008. The transcription factor ATF3 acts as an oncogene in mouse mammary tumorigenesis. BMC Cancer 8:268.
Warnecke T, Batada NN, Hurst LD. 2008. The impact of the nucleosome code on protein-coding sequence evolution in yeast. PLoS Genet. 4:e1000250.
Washietl S, Machne R, Goldman N. 2008. Evolutionary footprints of nucleosome positions in yeast. Trends Genet. 24:583–587.
Woolfe A, et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. 3:e7.

Associate editor: Maria Costantini