Evolution of GC content in the histone gene repeating units from *Drosophila lutescens*, *D. takahashii* and *D. pseudoobscura*

Yuko Nakashima, Asako Higashiyama, Ayana Ushimaru,
Nozomi Nagoda and Yoshinori Matsuo*

Laboratory of Adaptive Evolution, Institute of Socio-Arts and Sciences, Tokushima University,
Tokushima 770-8502, Japan

(Received 18 February 2015, accepted 22 October 2015; J-STAGE Advance published date: 25 March 2016)

A subset of histone genes (H1, H2A, H2B and H4), which are encoded along with H3 within repeating units, were analyzed in *Drosophila lutescens*, *D. takahashii* and *D. pseudoobscura* to investigate the evolutionary mechanisms influencing this multigene family and its GC content. Nucleotide divergence among species was more marked in the less functional regions. A strong inverse relationship was observed between the extent of evolutionary divergence and GC content within the repeating units; this finding indicated that the functional constraint on a region must be associated with both divergence and GC content. The GC content at 3rd codon positions in the histone genes from *D. lutescens* and *D. takahashii* was higher than that from *D. melanogaster*, while that from *D. pseudoobscura* was similar. These evolutionary patterns were similar to those of H3 gene regions. Based on these findings, we propose that the evolutionary mechanisms governing nucleotide content at 3rd codon positions tend to eliminate A and T nucleotides more frequently than G and C nucleotides. These changes might be the consequence of negative selection and would result in GC-rich 3rd codon positions. In addition, interspecific differences in GC content, which exhibited the same pattern for all histone genes, could be explained by different selection efficiencies that result from changes in population size.

Key words: *Drosophila*, GC content, histone gene, population size effect

INTRODUCTION

The genome of an organism usually comprises various types of genes and elements; these include single-copy genes, duplicated genes, dispersed multigene families, clustered multigene families, repetitive elements and transposons, among others (Adams et al., 2000). The mechanisms underlying how such genes and elements are evolving must be studied to fully understand the evolution of genomes and populations. The histone genes represent a multigene family that forms a tandem repetitive cluster. Repeating units comprising the five replication-dependent histone genes (H1, H2A, H2B, H3 and H4) are tandemly repeated approximately 110 times in *Drosophila melanogaster* (Lifton et al., 1977). The members of this multigene histone family in *D. melanogaster* have evolved in a concerted fashion (Coen et al., 1982;

Matsuo and Yamazaki, 1989b; Kakita et al., 2003). Although the nucleotide sequences of the H3 genes have been analyzed in a wide range of *Drosophila* species (Matsuo, 2000a, 2000b, 2003, 2006; Tsunemoto and Matsuo, 2001), fewer studies have focused on the remainder of the repeated unit (Matsuo and Yamazaki, 1989a; Kremer and Hennig, 1990; Fitch and Strausbaugh, 1993; Schienman et al., 1998; Nagel and Grossbach, 2000; Tsunemoto and Matsuo, 2001; Kakita et al., 2003; Nagoda et al., 2005). DNA sequence data on the variability and divergence of multigene families have accumulated for the histone genes (Matsuo, 2006). Analyzing the histone genes from various *Drosophila* species could provide important information about mutation bias and population size, among other issues.

The evolutionary mechanisms that generate codon bias and GC content variation have been studied and discussed extensively (Ikemura, 1985; Bernardi and Bernardi, 1986; Li, 1987; Shields et al., 1988; Poh et al., 2012; Lawrie et al., 2013). Selective pressure and mutational bias are considered to be the main factors affecting GC content at

Edited by Kyoichi Sawamura

* Corresponding author. E-mail: matsuo.yoshinori@tokushima-u.ac.jp
DOI: <http://doi.org/10.1266/ggs.15-00018>

3rd codon positions (Moriyama and Hartl, 1993; Kliman and Hey, 1994; Akashi, 1995; Powell and Moriyama, 1997; Akashi et al., 2006; Heger and Ponting, 2007; Plotkin and Kudla, 2011). The factors affecting GC content at 3rd codon positions are relevant to concerted evolution (Matsuo, 2006). Studies of the GC content at 3rd codon positions within *Xdh*, *Adh* and several other genes in the *Drosophila saltans* group have shown that a low GC content in the *saltans* lineage can be explained by a shift in mutation pressure (Rodriguez-Trelles et al., 1999, 2000). However, analysis of the H3 gene region in *Drosophila* indicates that selection, not mutation bias, may be a major factor responsible for generating this codon bias (Matsuo, 2000b). Specifically, it was proposed that the effect of population size on selection efficiency is responsible for observed interspecific differences in GC content (Matsuo, 2000b, 2003, 2006). To identify and assess the contribution of population size to GC content, nucleotide sequence data for other histone genes and regions within the repeating unit should be analyzed. If the evolutionary pattern of the other histone genes is quite different from that of the H3 gene, the effect of population size would be small because all the genes in a genome are presumably affected simultaneously by any series of historical events experienced by a species. In this study, the histone genes linked within a repeating unit, H1, H2A, H2B and H4, are analyzed. Many genes should be studied to draw genome-wide conclusions.

To investigate the evolutionary mechanisms influencing the histone multigene family in *Drosophila*, DNA sequence data from as many *Drosophila* species as possible should be used. In this study, the H1, H2A, H2B and H4 histone genes within the repeating units were cloned from *D. lutescens*, *D. takahashii* and *D. pseudoobscura* and analyzed to investigate the evolutionary mechanisms that affect GC content of this gene family.

MATERIALS AND METHODS

***Drosophila* strains and PCR amplification** The *Drosophila* strains from *D. lutescens*, *D. takahashii* and *D. pseudoobscura* were kindly donated by Kyushu University, Japan. Genomic DNA was extracted from larvae with a DNA extraction kit (Sepa Gene Kit, Sanko Junyaku, Tokyo, Japan). Genomic DNA was amplified by polymerase chain reaction (PCR) with Takara Ex *Taq* (Takara Bio, Shiga, Japan) and the protocol of Saiki et al. (1985); briefly, each PCR cycle ($n = 40$) comprised denaturation at 94 °C for 1 min, annealing at 55 °C for 2 min, polymerization at 70 °C for 2 min, and an extension step of 5 sec. The nucleotide sequences of the primers used for PCR amplification were 5'-GGCGAGCGTGCTTAA-3' (H3F3), 5'-GGCGAGCGAGCTTAA-3' (H3F10) and 5'-TTGGTACGGGCCAT-3' (H3R20). The H3F3 and H3R20 primer pair was used for *D. pseudoobscura*, and the

H3F10 and H3R20 pair was used for *D. takahashii* and *D. lutescens*. Because the H3 genes of these species have already been investigated, primers were designed to amplify all of a typical single histone gene repeating unit except the H3 gene within the unit (Tsunemoto and Matsuo, 2001). Amplification of genomic DNA with these primers resulted in a single major band for each of the three *Drosophila* species. Arrangement, polarity and length of each histone gene within the clones derived from these bands were determined by sequencing, which confirmed that the clones were typical repeating units for the respective species. No protein-coding genes other than the five histones have been found in any of the cloned or analyzed repeating histone gene units (Matsuo and Yamazaki, 1989a).

Cloning and sequencing PCR products were cloned into the plasmid vector PCR2.1 (Invitrogen, Carlsbad, CA, USA) and sequenced with a Dye Terminator sequencing kit (Applied Biosystems, Foster City, CA, USA) and an ABI 310 sequencer (Sanger et al., 1977). The sequencing strategy for the histone gene repeating unit was similar to that described previously (Tsunemoto and Matsuo, 2001) with some modifications. As shown by Kakita et al. (2003), heterogeneity among histone gene clones from a single *Drosophila* strain, most of which reflects the genetic difference among family members, is 0.6–0.7%. This level of difference was much lower than the interspecific differences observed here, which were approximately 10%.

DNA analysis DNA sequences of the histone gene units from *D. lutescens*, *D. takahashii* and *D. pseudoobscura* were deposited in the DNA Data Bank of Japan (DDBJ). The accession numbers for the units are AB249649, AB249650 and AB249651 for *D. lutescens*, *D. takahashii* and *D. pseudoobscura*, respectively. The DNA sequences of other *Drosophila* histone gene units were obtained from DDBJ/GenBank: accession numbers X14215 for *D. melanogaster* (Matsuo and Yamazaki, 1989), X17072 and X52576 for *D. hydei* (Kremer and Hennig, 1990; Fitch and Strausbaugh, 1993), and AB192418 for *D. americana* (Nagoda et al., 2005). Analysis was done for *Drosophila* species whose histone genes have been experimentally studied, cloned and sequenced. In studying clusters of tandemly duplicated genes like the histone gene family, it is hard to know which genes (or repetitive units) are orthologous to each other. This problem arises because highly similar units are repeated many times, and variation in the number of repeats is found both within and between species. In most tandem clusters, as found here for the histone gene cluster, the family members show concerted evolution (Coen et al., 1982; Matsuo and Yamazaki, 1989b). Therefore, only one typical-length repeated unit was analyzed for each species. Multiple

sequence alignment and phylogenetic analysis were conducted using the Clustal W program (Thompson et al., 1994) from DDBJ with the NJ method (Saitou and Nei, 1987). Bootstrap values were calculated for 1,000 replications. Since the core histones are highly conserved, most nucleotide substitutions in the coding regions are synonymous changes. In such cases, percentage is a better measure of divergence.

RESULTS

Structure of the histone gene repeating units from *D. lutescens*, *D. takahashii* and *D. pseudoobscura*

To determine the structure and nucleotide sequence of the repeating units encoding the histone genes in *D. lutescens*, *D. takahashii* and *D. pseudoobscura*, PCR products amplified from genomic DNA were cloned into a plasmid vector. The H3 genes from these species have been analyzed previously; therefore, the primers used here for DNA amplification were designed to cover all of a single repeating unit except the H3 gene (Fig. 1). The nucleotide sequences of inserts of 4,785 bp, 5,071 bp and 5,042 bp (including the H3 gene being 5,196 bp, 5,482 bp and 5,453 bp) for *D. lutescens*, *D. takahashii* and *D. pseudoobscura*, respectively, showed that the H4, H2A, H2B and H1 genes were coded in this order within a single repeating unit (Fig. 1). For each species, the locations and transcriptional polarities of the four histone genes within the respective units were identical to those found in other *Drosophila* species (Lifton et al., 1977; Matsuo and Yamazaki, 1989a; Kremer and Hennig, 1990; Fitch and Strausbaugh, 1993; Nagel and Grossbach, 2000; Tsunemoto and Matsuo, 2001; Kakita et al., 2003; Nagoda et al., 2005).

Two spacer regions, one between the H4 and H2A genes in *D. pseudoobscura* and the other between the H1 and H3 genes in *D. takahashii*, were substantially larger than the corresponding regions in *D. melanogaster*; the difference in size between the *D. takahashii* spacer and the corresponding *D. lutescens* spacer was attributed to direct duplication of about 150 bp sequence (shown in light blue in Fig. 1). Although a number of small insertions and deletions were found in the spacer regions, the size differences between spacer regions between species were relatively small (Fig. 1). Similarly, the sizes of the coding regions were also highly conserved. The lengths of the H4, H2A and H2B genes were identical to those of *D. melanogaster*. However, compared to *D. melanogaster*, the H1 gene was 3 bp shorter in *D. pseudoobscura* and 18 bp shorter in *D. lutescens* and *D. takahashii*, but these differences in length did not result in any frameshift mutations. The amino acid sequences of H1 in these four species showed higher sequence divergence than that for the other histones.

Divergence, functional constraint and GC content of the regions in the histone gene repeating units

Comparisons of the nucleotide sequences of the repeating units from each of the three species to that of *D. melanogaster* revealed that divergence, represented as percent nucleotide difference, was larger in spacer regions than in coding regions (Fig. 2, Table 1). Moreover, the spacers in the 3' regions (downstream of the coding regions) between H4-H2A and H2B-H1 were more divergent than the spacers in the 5' regions (upstream of the coding region) between H3-H4 and H2A-H2B (Fig. 2, Table 1). The largest spacer, between H1-H3, showed the highest divergence. This observed pattern of divergence, in which functionally more important regions are more highly conserved than less important regions (Miyata and Yasunaga, 1981; Li, 1983), is consistent with the results obtained for the H3 gene region (Matsuo and Yamazaki, 1989b) and for the histone gene repeating units of the *D. melanogaster* species subgroup (Tsunemoto and Matsuo, 2001; Kakita et al., 2003). It is likely that negative selection was associated with the extent of nucleotide divergence between species (Kimura, 1983). A phylogenetic tree constructed with these repeating units that is consistent with trees constructed from the H3 gene (Matsuo, 2003) and other genes (Powell and DeSalle, 1995; Russo et al., 1995; Inomata et al., 1997) is shown in Fig. 3.

Separate comparisons for GC content in distinct regions of the repeating unit were conducted for each *Drosophila* species (Fig. 4). The results revealed that GC content was higher in the coding regions and lower in the spacer regions (Table 2). In the spacer regions, the GC content in the 5' regions of genes was higher than in the 3' regions (Table 2). While interspecific differences in GC content were apparent in the coding regions and to a lesser degree in the 5' regions, no such differences were observed in the 3' regions or in the longer H1-H3 spacer (Fig. 4, Table 2). A significant difference in the GC content between species was not found in the 3'-most region of the entire repeated unit, suggesting that GC content and nucleotide substitutions are in a steady state as was reported previously for the 3' region of the H3 gene (Matsuo, 2000b).

Furthermore, for longer evolutionary periods, the extent of divergence between species and the GC content of distinct regions within the repeating units were confirmed to be inversely related; specifically, large interspecies divergence was associated with lower GC content (Table 3). These findings corroborated those of a comparative study between two closely related species, *D. melanogaster* and *D. simulans* (Tsunemoto and Matsuo, 2001), and suggest that the GC content of distinct regions within the repeat unit is subject to functional constraint.

Interspecies differences in GC content at 3rd codon positions within histone genes

The GC content of

Fig. 1. Organization of the histone gene repeating unit in *D. melanogaster*, *D. pseudoobscura*, *D. lutescens* and *D. takahashii*. The numbers indicate the size (bp) of the regions or units. Arrows show the direction of transcription. The insertion found in the spacer between H1 and H3 in *D. takahashii* is a repetitive one (light blue color).

Fig. 2. Nucleotide divergence in the histone gene repeating unit of *D. lutescens*, *D. takahashii* and *D. pseudoobscura* compared to *D. melanogaster*. Divergences in the spacer H3-H4, H4 gene, spacer H4-H2A, H2A gene, spacer H2A-H2B, H2B gene, spacer H2B-H1, H1 gene and spacer H1-H3 are shown separately.

Table 1. χ^2 tests for heterogeneity of divergence with *D. melanogaster* in the coding vs. spacer regions, and in the 5' vs. 3' regions

		<i>D. melanogaster</i> – <i>D. lutescens</i>	<i>D. melanogaster</i> – <i>D. takahashii</i>	<i>D. melanogaster</i> – <i>D. pseudoobscura</i>
Coding region [#]	Different site (D)	177	203	278
	Identical site (I)	1635	1627	1534
	Total site (T)	1812	1830	1812
	Divergence (D/T)	0.098	0.111	0.153
Spacer region	Different site (D)	754	1346	1137
	Identical site (I)	1973	1964	1508
	Total site (T)	2727	3310	2645
	Divergence (D/T)	0.277	0.407	0.430
χ^2 test Coding - Spacer	χ^2 (d.f. = 1)	133***	326***	227***
5' region (H3-H4, H2A-H2B)	Different site (D)	78	116	158
	Identical site (I)	417	426	312
	Total site (T)	495	542	470
	Divergence (D/T)	0.158	0.214	0.336
3' region (H4-H2A, H2B-H1)	Different site (D)	235	399	332
	Identical site (I)	615	581	519
	Total site (T)	850	980	851
	Divergence (D/T)	0.277	0.407	0.390
χ^2 test 5' region – 3' region	χ^2 (d.f. = 1)	15.7***	37.4***	2.43

*** $P < 0.001$.[#] H3 is not included.

an entire coding region is affected by multiple factors such as coding proteins, constraint for amino acids and the presence of codon bias. However, the interspecies difference in GC content in the histone genes is mainly due to 3rd codon positions within the genes, because the amino acid sequences of histones are highly conserved between species. Therefore, the GC content at 3rd codon positions was subject to further analysis. The GC content at 3rd codon positions within histone genes in these *Drosophila* species was compared (Fig. 5). Marked differences in GC content at 3rd codon positions were evident between histone genes within a species; for example, the %GC values were higher for the H2B gene than the H1 gene, as found previously (Fitch and Strausbaugh, 1993; Tsunemoto and Matsuo, 2001). Nevertheless, interspecies differences were also apparent for histone genes (Table 4). For each histone gene studied (H4, H2A, H2B and H1), the GC content at 3rd codon positions in *D. lutescens* and *D. takahashii* was higher than that in *D. melanogaster*; although the interspecies differences for H2A were not significant, the differences for H4, H2B and H1 were significant at the 1% level (Table 4). For comparison, GC content data for the H3 genes in these species (Matsuo, 2003) are shown along with the corresponding data from *D. americana* (Nagoda et al., 2005) and *D. hydei* (Fitch and Strausbaugh, 1993) in Fig. 5. *D. americana*

Fig. 3. Phylogenetic relationships of the histone gene repeating units in *Drosophila*. Nucleotide sequences of the repeating unit except for the H3 gene were analyzed to generate a phylogenetic tree by the NJ method. Bootstrap values for 1000 replications are indicated.

Table 2. χ^2 tests for heterogeneity of GC content in the coding vs. spacer regions, in the 5' vs. 3' regions, and between species

	Species [#]	MEL	LUT	TAK	PSE	AME	HYD	χ^2 tests between species (d.f. = 5)
Coding region ^{##}	GC	917	999	992	958	884	861	35.4***
	AT	913	813	820	875	928	948	
	SUM	1830	1812	1812	1833	1812	1809	
	%GC	50.1	55.1	54.7	52.3	48.8	47.6	
Spacer region	GC	887	974	1051	1046	891	889	6.1
	AT	1915	1999	2208	2163	1911	2059	
	SUM	2802	2973	3259	3209	2802	2948	
	%GC	31.7	32.8	32.3	32.6	31.8	30.2	
χ^2 test Coding - Spacer	χ^2 (d.f. = 1)	159***	233***	245***	188***	134***	147***	
5' region (H3-H4, H2A-H2B)	GC	211	225	221	196	188	208	16.3*
	AT	311	281	284	288	347	369	
	SUM	522	506	505	484	535	577	
	%GC	40.4	44.5	43.8	40.5	35.1	36.0	
3' region (H4-H2A, H2B-H1)	GC	289	317	341	463	342	342	5.2
	AT	591	661	639	911	667	776	
	SUM	880	978	980	1374	1009	1118	
	%GC	32.8	32.4	34.8	33.7	33.9	30.6	
χ^2 test 5' region - 3' region	χ^2 (d.f. = 1)	8.21**	20.9***	11.4***	7.23**	0.24	5.17*	

*** $P < 0.001$; ** $P < 0.01$; * $P < 0.05$.[#] MEL: *D. melanogaster*; LUT: *D. lutescens*; TAK: *D. takahashii*; PSE: *D. pseudoobscura*; AME: *D. americana*; HYD: *D. hydei*.^{##} H3 is not included.

Table 3. Correlation coefficients between the divergence and GC content in the regions of the histone gene repeating unit

Interspecies divergence	GC mel ^a	GC lut	GC tak	GC pse
<i>D. melanogaster</i> - <i>D. lutescens</i>	-0.97* ^b	-0.97*		
<i>D. melanogaster</i> - <i>D. takahashii</i>	-0.97*		-0.97*	
<i>D. melanogaster</i> - <i>D. pseudoobscura</i>	-0.97*			-0.96*

^a GC mel, GC lut, GC tak and GC pse mean the GC content for *D. melanogaster*, *D. lutescens*, *D. takahashii* and *D. pseudoobscura*, respectively.^b Correlation coefficients between the interspecies divergence and GC content in the regions of the histone gene repeating unit. The regions studied are H3-H4 spacer, H4 gene, H4-H2A spacer, H2A gene, H2A-H2B spacer, H2B gene, H2B-H1 spacer, H1 gene and H1-H3 spacer.* Significant values with 7 d.f., $P < 0.001$.

and *D. hydei* were chosen for comparison because these species are distantly related to the *melanogaster* sub-group and DNA sequence data for their histone gene repeating units were available. For each histone gene, interspecific differences in GC content showed trends similar to those observed previously for the H3 gene (Fig. 5);

these findings indicated that the entire repeating unit is evolving in a similar manner.

DISCUSSION

Analysis of nucleotide divergence in the histone gene

Fig. 4. Comparison of the GC content of the histone gene repeating units in six *Drosophila* species: *D. melanogaster*, *D. lutescens*, *D. takahashii*, *D. pseudoobscura*, *D. americana* and *D. hydei*. The GC content of the spacer H3-H4, H4 gene, spacer H4-H2A, H2A gene, spacer H2A-H2B, H2B gene, spacer H2B-H1, H1 gene and spacer H1-H3 is compared separately for each of the six indicated species.

Fig. 5. Comparison of GC content at 3rd codon position of the histone genes in six *Drosophila* species. Data for the H3 genes are from Matsuo (2003).

repeating unit revealed that the coding regions of histone genes are more highly conserved than the sequences within the spacer regions. Furthermore, within the spacer regions, the 5' regions of the genes are more highly conserved than the 3' regions. It is thus apparent that the functional constraints associated with protein coding and with the regulation of gene expression are associated with the degree of divergence. The relationship between divergence and functional constraint suggested that negative selection is acting and that it has a marked influence on nucleotide divergence. The finding that the GC content is inversely related to the degree of divergence for each distinct region within the repeating unit was consistent with findings from comparisons of histone gene repeating units in two more closely related species, *D.*

melanogaster and *D. simulans* (Tsunemoto and Matsuo, 2001). These results suggested that negative selection has also affected GC content in these regions. The spacers have a low GC content, implying that, like other non-functional regions such as introns and transposons, they are AT-rich (Petrov and Hartl, 1999). Since functional constraints were expected to be weak in these regions, the effect of negative selection, if any, is likely to be small. Consequently, it is possible that the GC content in these regions is influenced by mutation bias, if such a bias exists. The AT richness in these regions suggests that the pattern of mutation is biased toward the A/T direction; this supposition is consistent with findings from the *Drosophila* genes (Petrov and Hartl, 1999). Furthermore, interspecies differences in the GC content of these less functionally constrained regions are relatively small and not significantly different, suggesting that the relationship between GC content and nucleotide substitution is nearly in a steady state. This conclusion is consistent with previous results on patterns of nucleotide substitution in the 3' region of the H3 gene (Matsuo, 2000; Tsunemoto and Matsuo, 2001). These results also indicated that the pattern of mutation is similar in the four *Drosophila* species studied (Matsuo, 2003). The coding and 5' regions of these histone genes are GC-rich; this finding indicated that the effect of negative selection results in the elimination of A or T more frequently than G or C, and that A or T in these regions must be more deleterious than G or C. It is not clear why GC content becomes high in the regions subject to strong functional constraint. One possible explanation for this finding is that when the base content is dissimilar to the mutation

Table 4. χ^2 tests for heterogeneity of GC content at 3rd codon position between histone genes, and between six species

Gene	Species [#]	MEL	LUT	TAK	PSE	AME	HYD	χ^2 test between species (d.f. = 5)
H4	GC ^{3rd}	58	66	64	55	42	44	19.5**
	AT ^{3rd}	45	37	39	48	61	59	
	SUM	103	103	103	103	103	103	
	%GC ^{3rd}	56.3	64.1	62.1	53.4	40.8	42.7	
H2A	GC ^{3rd}	67	78	82	70	62	63	10.8 (<i>P</i> = 0.06)
	AT ^{3rd}	57	46	42	54	62	61	
	SUM	124	124	124	124	124	124	
	%GC ^{3rd}	54.0	62.9	66.1	56.5	50.0	50.8	
H2B	GC ^{3rd}	78	94	93	75	77	70	17.9**
	AT ^{3rd}	45	29	30	48	46	53	
	SUM	123	123	123	123	123	123	
	%GC ^{3rd}	63.4	76.4	75.6	61.0	62.6	56.9	
H1	GC ^{3rd}	124	149	145	140	118	109	20.4***
	AT ^{3rd}	132	101	105	117	132	140	
	SUM	256	250	250	257	250	249	
	%GC ^{3rd}	48.4	59.6	58.0	54.5	47.2	43.8	
H3	GC ^{3rd}	79	89	85	75	61	66	17.4**
	AT ^{3rd}	57	47	51	61	75	70	
	SUM	136	136	136	136	136	136	
	%GC ^{3rd}	58.1	65.4	62.5	55.2	44.9	48.5	
Total	GC ^{3rd}	406	476	469	415	360	352	75.4***
	AT ^{3rd}	336	260	267	328	376	383	
	SUM	742	736	736	743	736	735	
	%GC ^{3rd}	54.7	64.7	63.7	55.9	48.9	47.9	
χ^2 test between histone genes	χ^2 (d.f. = 4)	8.6	10.5*	11.6*	1.8	13.2*	7.3	

*** *P* < 0.001; ** *P* < 0.01; * *P* < 0.05.[#] MEL: *D. melanogaster*; LUT: *D. lutescens*; TAK: *D. takahashii*; PSE: *D. pseudoobscura*; AME: *D. americana*; HYD: *D. hydei*.

bias, the probability of generating significant combinations of nucleotide sequence accidentally will be lower.

Selective pressure and mutational bias are considered as putative main factors affecting GC content at 3rd codon positions (Ikemura, 1985; Bernardi and Bernardi, 1986; Shields et al., 1988; Moriyama and Hartl, 1993; Kliman and Hey, 1994; Akashi, 1995; Powell and Moriyama, 1997; Heger and Ponting, 2007; Plotkin and Kudla, 2011). GC richness at 3rd codon positions is difficult to explain by mutation pattern alone because such patterns are likely to be biased toward A/T (Powell and Moriyama, 1997). Therefore, at 3rd codon positions, A or T must be eliminated at a higher frequency than are G or C, possibly

by negative selection, in the same way as has been observed in other regions, such as the 5' region. How, then, might interspecific differences in GC content at 3rd codon positions be explained? Studies of GC content at 3rd codon positions of *Xdh*, *Adh*, and several other genes in the *Drosophila saltans* group revealed that a low GC content in the *saltans* lineage could be explained by a shift in mutation pressure (Rodriguez-Trelles et al., 1999, 2000). However, as discussed above, the mutation pattern within the histone gene region may not differ among the species studied (Matsuo, 2003). Therefore, selection is the most likely explanation for the observed interspecific differences in GC content. Under the selection

hypothesis, there are two possible models for explaining the observed interspecific divergence in GC content at 3rd codon positions. One model is that the direction of codon bias differs between species, and that this difference results in differences in GC content at 3rd codon positions (Poh et al., 2012). The alternative model is that divergence in GC content at 3rd codon positions could be caused by differences in selection efficiency (Matsuo, 2003, 2006). Although the former model cannot be excluded completely, the second model is preferable for the following reason: it is reasonable to assume that the same selection pressure as in other genomic regions is also working on the synonymous sites in coding regions; therefore, less functional regions become more AT-rich than the more functional regions in most *Drosophila* species. The intensity of selection for GC content at 3rd codon position is essentially similar in meaning to codon choice and is likely to be small enough to be affected by the size of a population (Hartl et al., 1994). In a model depicting slightly deleterious mutations (Ohta, 1972), selection was observed to act effectively when populations were large. In such instances, A or T residues at 3rd codon positions would be eliminated, and these sites would effectively become GC-rich. Conversely, for smaller populations, selection does not act effectively, and GC content at 3rd codon positions has moved toward that which is more influenced by the mutation pattern, becoming slightly AT-rich. The former scenario is likely to have occurred in the ancestor of *D. takahashii* and *D. lutescens*, and the latter scenario in the ancestor of *D. hydei* and *D. americana*. Moreover, the effect of population size will not only be manifested in a single gene, but will have occurred simultaneously for all genes in a genome. This means that the aforementioned hypothesis can be tested by analyzing the GC content of other genomic genes in these species. Similar interspecific patterns in GC content at 3rd codon positions of all histone genes in a repeating unit as well as of other *Drosophila* genes (Matsuo, 2000a) would provide supporting evidence for the effect of population size. For determining the mechanism of concerted evolution of the histone multigene family, a negative selection and size effect on codon usage as well as gene conversion and unequal crossing-over will be important factors.

We would like to thank the anonymous referees for their important comments. This research was supported by a Grant-in-Aid for Scientific Research awarded to Y. M. from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

REFERENCES

- Adams, M. D., et al. (2000) The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195.
- Akashi, H. (1995) Inferring weak selection from patterns of polymorphism and divergence at “silent” sites in *Drosophila* DNA. *Genetics* **139**, 1067–1076.
- Akashi, H., Ko, W. Y., Piao, S., John, A., Goel, P., Lin, C. F., and Vitins, A. P. (2006) Molecular evolution in the *Drosophila melanogaster* species subgroup: frequent parameter fluctuations on the timescale of molecular divergence. *Genetics* **172**, 1711–1726.
- Bernardi, G., and Bernardi, G. (1986) Compositional constraints and genome evolution. *J. Mol. Evol.* **24**, 1–11.
- Coen, E., Strachan, T., and Dover, G. (1982) Dynamic of concerted evolution of ribosomal DNA and histone gene families in the *melanogaster* species subgroup of *Drosophila*. *J. Mol. Biol.* **158**, 17–35.
- Fitch, D. H. A., and Strausbaugh, L. D. (1993) Low codon bias and higher rates of synonymous substitution in *Drosophila hydei* and *D. melanogaster* histone genes. *Mol. Biol. Evol.* **10**, 397–413.
- Hartl, D. L., Moriyama, E. N., and Sawyer, S. A. (1994) Selection intensity for codon bias. *Genetics* **138**, 227–234.
- Heger, A., and Ponting, C. P. (2007) Variable strength of translational selection among 12 *Drosophila* species. *Genetics* **177**, 1337–1348.
- Ikemura, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2**, 13–34.
- Inomata, N., Tachida, H., and Yamazaki, T. (1997) Molecular evolution of the *Amy* multigenes in the subgenus *Sophophora* of *Drosophila*. *Mol. Biol. Evol.* **14**, 942–950.
- Kakita, M., Shimizu, T., Emoto, M., Nagai, M., Takeguchi, M., Hosono, Y., Kume, N., Ozawa, T., Ueda, M., Bhuiyan, Md. S. I., and Matsuo, Y. (2003) Divergence and heterogeneity of the histone gene repeating units in the *Drosophila melanogaster* species subgroup. *Genes Genet. Syst.* **78**, 383–389.
- Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*. Cambridge Univ. Press, Cambridge.
- Kliman, R. M., and Hey, J. (1994) The effects of mutation and natural selection on codon bias in the genes of *Drosophila*. *Genetics* **137**, 1049–1056.
- Kremer, H., and Hennig, W. (1990) Isolation and characterization of a *Drosophila hydei* histone DNA repeat unit. *Nucleic Acids Res.* **18**, 1573–1586.
- Lawrie, D. S., Messer, P. W., Hershberg, R., and Petrov, D. A. (2013) Strong purifying selection at synonymous sites in *D. melanogaster*. *PLoS Genetics* **9**, DOI: 10.1371/journal.pgen.1003527.
- Li, W. H. (1983) Evolution of duplicate genes and pseudogenes. In: *Evolution of Genes and Proteins*. (eds.: Nei, M., and Koehn, R. K.), pp. 14–37. Sinauer Associates, Inc., Sunderland, MA.
- Li, W. H. (1987) Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J. Mol. Evol.* **24**, 337–345.
- Lifton, R. P., Goldberg, M. L., Karp, R. W., and Hogness, D. S. (1977) The organization of the histone genes in *Drosophila melanogaster*: functional and evolutionary implications. *Cold Spring Harbor Symp. Quant. Biol.* **42**, 1047–1051.
- Matsuo, Y. (2000a) Evolutionary change of codon usage for the histone gene family in *Drosophila melanogaster* and *Drosophila hydei*. *Mol. Phylogenet. Evol.* **15**, 283–291.
- Matsuo, Y. (2000b) Molecular evolution of the histone 3 multigene family in the *Drosophila melanogaster* species subgroup. *Mol. Phylogenet. Evol.* **16**, 339–343.
- Matsuo, Y. (2003) Evolution of the GC content of the histone 3 gene in seven *Drosophila* species. *Genes Genet. Syst.* **78**, 309–318.
- Matsuo, Y. (2006) Genomic structure and evolution of the histone gene family in *Drosophila*. *Topics in Current Genetics*

- 2, 1–14.
- Matsuo, Y., and Yamazaki, T. (1989a) tRNA derived insertion element in histone gene repeating unit of *Drosophila melanogaster*. *Nucleic Acids Res.* **17**, 225–238.
- Matsuo, Y., and Yamazaki, T. (1989b) Nucleotide variation and divergence in the histone multigene family in *Drosophila melanogaster*. *Genetics* **122**, 87–97.
- Miyata, T., and Yasunaga, T. (1981) Rapidly evolving mouse α -globin related pseudo gene and its evolutionary history. *Proc. Natl. Acad. Sci. USA* **78**, 450–453.
- Moriyama, E. N., and Hartl, D. L. (1993) Codon usage bias and base composition of nuclear genes in *Drosophila*. *Genetics* **134**, 847–858.
- Nagel, S., and Grossbach, U. (2000) Histone H1 genes and histone gene clusters in the genus *Drosophila*. *J. Mol. Evol.* **51**, 286–298.
- Nagoda, N., Fukuda, A., Nakashima, Y., and Matsuo, Y. (2005) Molecular characterization and evolution of the repeating units of histone genes in *Drosophila americana*: coexistence of quartet and quintet units in a genome. *Insect Mol. Biol.* **14**, 713–717.
- Ohta, T. (1972) Population size and rate of evolution. *J. Mol. Evol.* **1**, 305–314.
- Petrov, D. A., and Hartl, D. L. (1999) Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. *Proc. Natl. Acad. Sci. USA* **96**, 1475–1479.
- Plotkin, J. B., and Kudla, G. (2011) Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* **12**, 32–42.
- Poh, Y. P., Ting, C. T., Fu, H. W., Langley, C. H., and Begun, D. (2012) Population genomic analysis of base composition evolution in *Drosophila melanogaster*. *Genome Biol. Evol.* **4**, 1245–1255.
- Powell, J. R., and DeSalle, R. (1995) *Drosophila* molecular phylogenies and their uses. *Evol. Biol.* **28**, 87–138.
- Powell, J. R., and Moriyama, E. N. (1997) Evolution of codon usage bias in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **94**, 7784–7790.
- Rodriguez-Trelles, F., Tarrio, R., and Ayala, F. J. (1999) Switch in codon bias and increased rates of amino acid substitution in the *Drosophila saltans* species group. *Genetics* **153**, 339–350.
- Rodriguez-Trelles, F., Tarrio, R., and Ayala, F. J. (2000) Fluctuating mutation bias and the evolution of base composition in *Drosophila*. *J. Mol. Evol.* **50**, 1–10.
- Russo, C. A., Takezaki, N., and Nei, M. (1995) Molecular phylogeny and divergence times of Drosophilid species. *Mol. Biol. Evol.* **12**, 391–404.
- Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., and Arnheim, N. (1985) Enzymatic amplification of β -globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**, 1350–1354.
- Saitou, N., and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.
- Sanger, F., Nicklen, S., and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
- Schienman, J. E., Lozovskaya, E. R., and Strausbaugh, L. D. (1998) *Drosophila virilis* has atypical kinds and arrangements of histone repeats. *Chromosoma* **107**, 529–539.
- Shields, D. C., Sharp, P. M., Higgins, D. G., and Wright, W. (1988) “Silent” sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Mol. Biol. Evol.* **5**, 704–716.
- Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.
- Tsunemoto, K., and Matsuo, Y. (2001) Molecular evolutionary analysis of a histone gene repeating unit from *Drosophila simulans*. *Genes Genet. Syst.* **76**, 355–361.