Comparative Genomic Study Reveals a Transition from TA Richness in Invertebrates to GC Richness in Vertebrates at CpG Flanking Sites: An Indication for Context-Dependent Mutagenicity of Methylated CpG Sites

Yong Wang and Frederick C.C. Leung*

School of Biological Sciences and Genome Research Centre, The University of Hong Kong, Pokfulam, Hong Kong, China.

Vertebrate genomes are characterized with CpG deficiency, particularly for GC-poor regions. The GC content-related CpG deficiency is probably caused by context-dependent deamination of methylated CpG sites. This hypothesis was examined in this study by comparing nucleotide frequencies at CpG flanking positions among invertebrate and vertebrate genomes. The finding is a transition of nucleotide preference of 5′ T to 5′ A at the invertebrate-vertebrate boundary, indicating that a large number of CpG sites with 5′ Ts were depleted because of global DNA methylation developed in vertebrates. At genome level, we investigated CpG observed/expected (obs/exp) values in 500 bp fragments, and found that higher CpG obs/exp value is shown in GC-poor regions of invertebrate genomes (except sea urchin) but in GC-rich sequences of vertebrate genomes. We next compared GC content at CpG flanking positions with genomic average, showing that the GC content is lower than the average in invertebrate genomes, but higher than that in vertebrate genomes. These results indicate that although 5′ T and 5′ A are different in inducing deamination of methylated CpG sites, GC content is even more important in affecting the deamination rate. In all the tests, the results of sea urchin are similar to vertebrates perhaps due to its fractional DNA methylation. CpG deficiency is therefore suggested to be mainly a result of high mutation rates of methylated CpG sites in GC-poor regions.

Key words: CpG deficiency, DNA methylation, invertebrate genomes, GC content

Introduction

CpG deficiency has been widely observed in vertebrate genomes (*1–3*) and is believed to be caused by DNA methylation (*4*, *5*). The degree of CpG deficiency is not uniform across a mammalian genome according to reports showing more significant CpG deficiency in GC-poor regions (*6–9*). Understanding this effect will help to reveal the fundamental reason for CpG mutation and DNA methylation. In a previous report, DNA methylation has been attempted to interpret CpG deficiency in TA-rich regions (*9*). However, the combined effect of DNA methylation and TA richness on deamination of methylated CpGs is still hypothetical up to now, due to lack of evidence.

***Corresponding author.**
E-mail: fcleung@hkucc.hku.hk

The connection between TA richness and methylated CpG mutations can be elucidated by comparative analyses between invertebrate and vertebrate genomes, in which DNA methylation level is clearly different. The pattern of DNA methylation in invertebrates is absent or fractional. In yeast and *Caenorhabditis elegans*, DNA methylation has not been detected. Both fruitfly and honeybee were found to have slight DNA methylation, and the difference between the two species is the major recognition site, CpT for fruitfly and CpG for honeybee (*10–12*). Fractional DNA methylation (20% at genome level) emerges in sea urchins and sea squirts (*13*). The DNA methylation in honeybees, sea urchins and sea squirts is similar to that in vertebrates in terms of recognition site and DNA methyltransferases involved, whereas it is restricted in transcribed regions

(*12*, *14*). Global methylation is observed only in vertebrates, and both genic and non-genic regions may be methylated. Among vertebrates, deamination rate of the methylated CpG sites is different because of positive correlation between the deamination rate and body temperature (*15*). Warm-blooded animals therefore show a higher CpG mutational rate than cold-blooded ones.

Because of the different methylation levels and corresponding CpG mutational rates between invertebrates and vertebrates, the hypothetical relationship between DNA methylation and the context dependence of CpG deficiency could be examined by comparing flanking nucleotide frequencies and local GC content of CpG sites among the organisms. If one observed a consistency for the nucleotide frequency and GC content between invertebrates and vertebrates, the context dependence might be a reflect of a compositional bias in all eukaryotic genomes; if an opposite tendency for the results was obviously exhibited, the CpG mutations caused by DNA methylation could be confirmed to be context-dependent. In this study, we selected four invertebrates and four vertebrates for the comparative genomics analyses. *Caenorhabditis elegans*, *Drosophila melanogaster* (fruitfly), *Apis mellifera* (bee), and *Strongylocentrotus purpuratus* (sea urchin) represent the invertebrates showing no DNA methylation, slight non-CpG methylation, slight CpG methylation, and fractional CpG methylation, respectively; Cold-blooded and warm-blooded vertebrates *Fugu rubripes* (pufferfish), *Danio rerio* (zebrafish), *Gallus gallus* (chicken), and *Homo sapiens* (human) were used because of different mutational rates at the methylated CpG sites.

Because DNA methylation pattern on repetitive elements is different between invertebrates and vertebrates, we used repeat-masked sequences for the analyses. In bee and sea urchin, reports demonstrate that repetitive elements are not the targets for methylation (*12*, *16*), whereas they are hypermethylated in vertebrates (*17*, *18*). This will create discrepancy in CpG frequency within the repetitive elements, and masking the elements may avoid making misleading results. Our results showed high 5′ T and 3′ A frequencies at CpG flanking positions in invertebrates (except sea urchin) and high 5′ A and 3′ T frequencies in vertebrates. We then performed genome-wide analyses to obtain observed/expected (obs/exp) values of CpG sites in regions specified with GC content. Our results indicate that the association between CpG deficiency and GC content is ascribed to context-dependent mutation of methylated CpG sites. Moreover, we reported CpG compositional biases in invertebrates.

Results

Nucleotide preference at CpG flanking positions

At CpG flanking positions, the separated four lines in Figure 1 show differences in frequency between A and T at identical positions, and also between G and C albeit not marginal in some species. The most marginal difference was generally observed at the positions ±1 and ±2. As the position is far from CpG, the difference becomes less remarkable. A consistent nucleotide preference for 5′ A and 3′ T was observed in at least ±1 and ±2 positions of all the vertebrates. The figures for the invertebrates show that T is more frequent than A at −1 and −2 positions, but less frequent at +1 and +2 positions. The figure for the sea urchin is an exception as it shows almost no difference in frequency between T and A, as well as between G and C. Therefore, the result of sea urchin is a transition of flanking nucleotide preference between invertebrates and vertebrates, and the transition correlates exactly with the emerging capacity of maintaining global CpG methylation in eukaryotes. All these findings suggest that methylated CpG sites with stretches of 5′ Ts and 3′ As mutate more easily in sea urchin and vertebrates.

Regarding G and C frequencies at the CpG flanking positions, we could not observe a consistent preference for G or C. For instance, G was favored at −1 and −2 positions over C in zebrafish, but a reverse preference was found at the same positions in human (Figure 1). The same is for the results of *C. elegans* and fruitfly.

Low GC content sequences show high CpG obs/exp ratios in invertebrates

The analyses were performed on sequences with bins of 500 bp, which had been classified into three groups according to GC content. High CpG obs/exp ratios were mostly shown in low GC group of the invertebrates, but in high GC group of the vertebrates (Figure 2). The high and low GC groups are referred to that containing sequences with GC content >50% and 20%–35% respectively. For fruitfly and bee, the peak

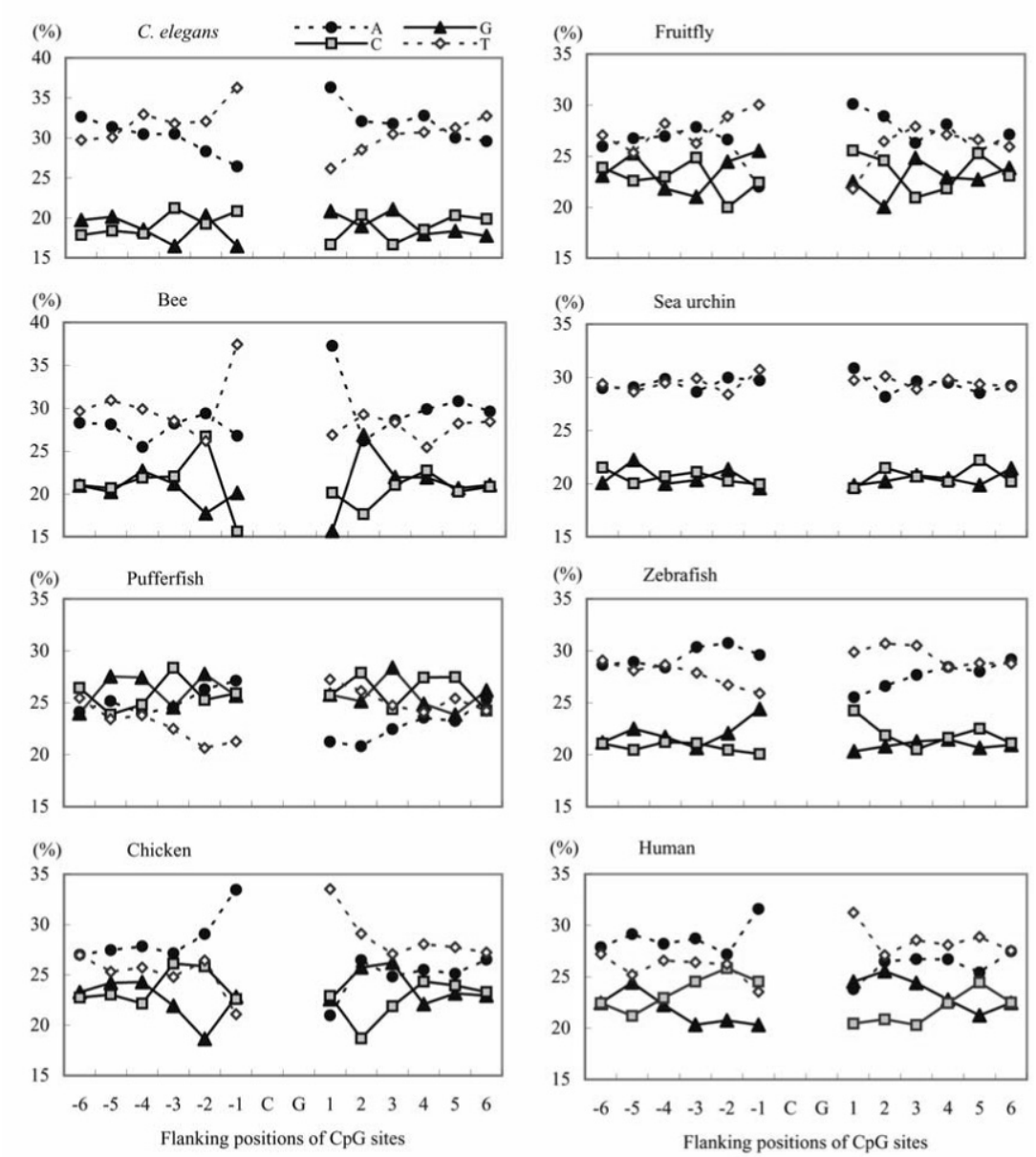

Fig. 1 Nucleotide frequency at CpG flanking positions. Nucleotide frequencies at CpG flanking positions were obtained from repeat-masked genomic sequences. Six 5′ flanking positions are labeled as −1 to −6; six 3′ flanking positions are labeled as 1 to 6.

of the percentage curve of high GC group is left-shifted compared with that of low GC group. Because of shortage of low GC content sequences, the curve of high GC group in *C. elegans* is not smooth, but the same conclusion as in fruitfly and bee can be drawn in the following figures for cumulative percentage (see below). The percentage curves in the four vertebrates are similar and characterized with left-shifted curves of low GC group compared with those in the invertebrates (except sea urchin). The difference is that the percentages of the low GC content sequences within low CpG obs/exp ratio ranges are relatively higher in chicken and human than in the fish species (Figure 2). More than 30% of low GC content sequences were found in extremely low CpG obs/exp ratios 0–0.1, indicating that GC-poor regions possess very few CpG sites in chicken and human genomes. Interestingly, the figure for sea urchin serves as a transitional pattern between the invertebrates and vertebrates, and is slightly similar to those for the vertebrates. Its three lines are nearly parallel and increase steadily, suggesting a high diversity of CpG obs/exp values in all the GC groups.

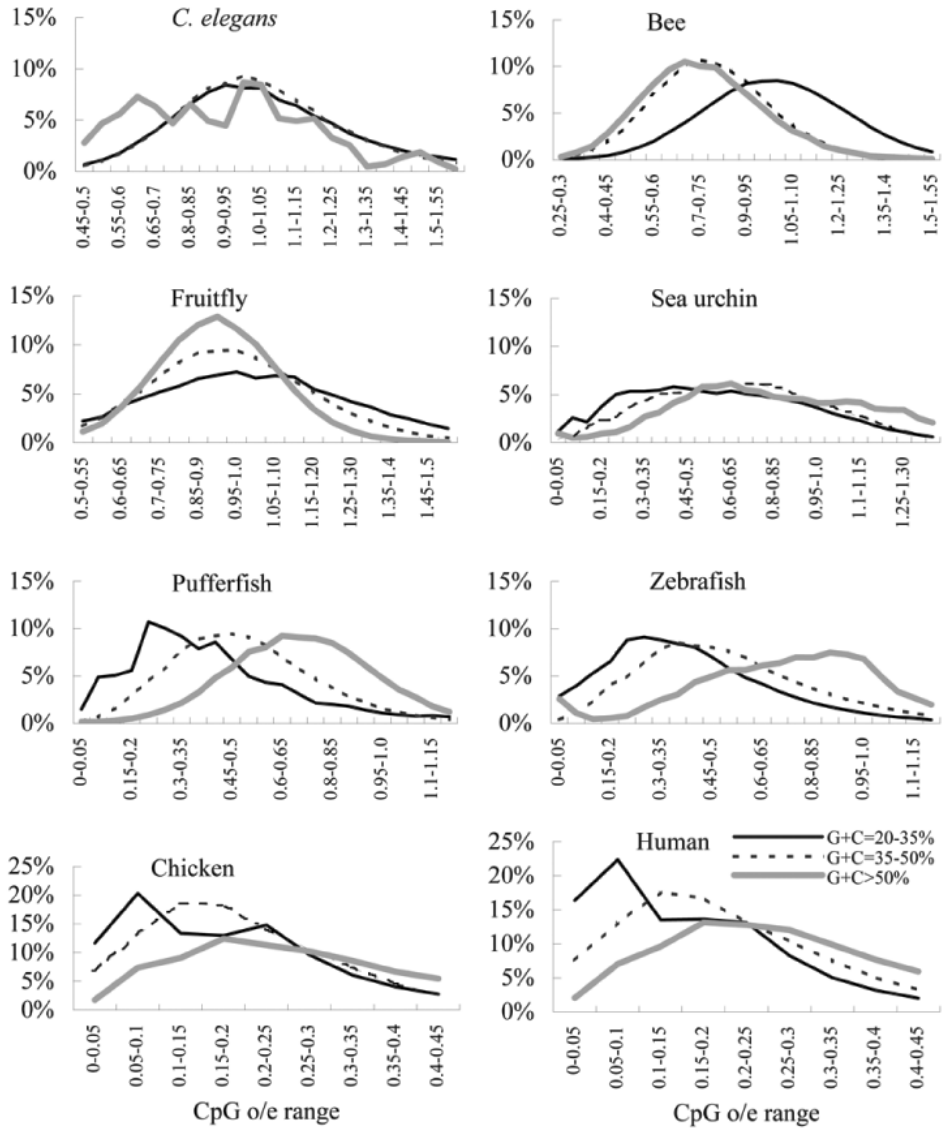

Fig. 2 Percentage of genomic fragments in different CpG obs/exp (o/e) ranges. CpG obs/exp values of all the genomic fragments in 500 bp unit were calculated. The fragments have been classified into three groups according to GC content. The axis x represents CpG obs/exp ranges in which the proportion of the fragments was shown. The three lines represent the proportions of the genomic segments for three GC groups: G+C=20%–35% (solid black line), 35%–50% (dashed line) and >50% (gray solid line).

We used Kruskal-Wallis test to evaluate the significance of the separation of the lines for different GC groups. Results show that the lines are significantly separated in the fish (P=0.0007 for pufferfish; P=0.015 for zebrafish) when the statistical analysis was performed using top 10 obs/exp values ranked on the basis of the proportion. We did not obtain the significant result for other species.

Cumulative percentage was obtained on the basis of the above mentioned percentages of CpG obs/exp ratios in individual GC content groups. The percentages for the ratios from range 0–0.05 were added up, until the cumulate of the percentages is approaching 100%. The fastest cumulative speed implicates in which range we could observe the highest percentage of CpG obs/exp ratio in a given GC group. We observed that the percentage of CpG obs/exp ratios was cumulated faster in low GC group than in high GC group for the invertebrates (except sea urchin), and the trend is converse in the vertebrates and sea urchin (see the steeping degree for the cumulative speed in Figure 3). For *C. elegans*, fruitfly and bee, the line for high GC content is left-shifted relative to the two for medium and low GC group, whereas it is right-shifted

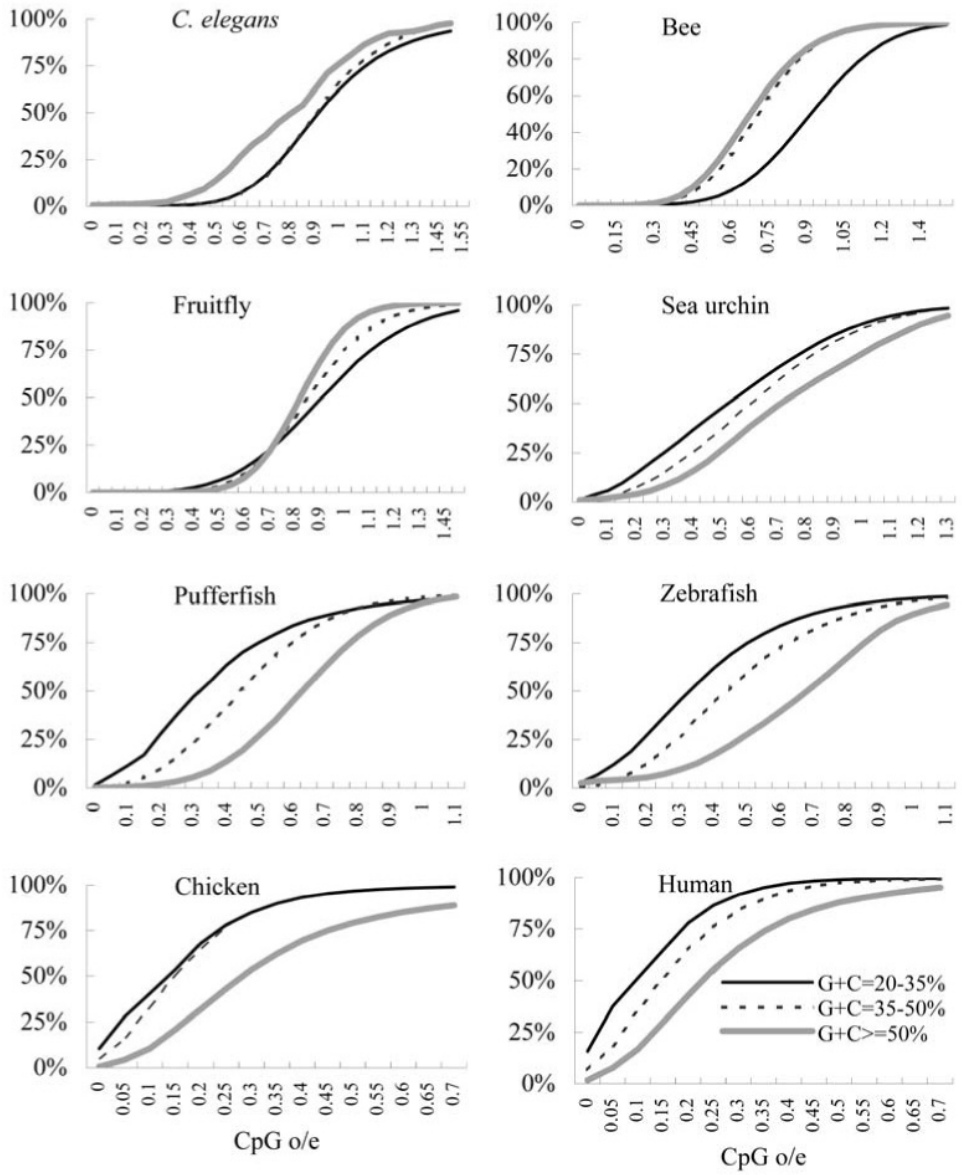

Fig. 3 Cumulative percent of genomic fragments in CpG obs/exp (o/e) ranges. All of the genomic fragments of 500 bp from 8 organisms were classified according to their GC level (G+C=20%–35%, 35%–50% and >50%) and CpG obs/exp values were then measured. In each GC level group, the percentage of the fragments belonging to individual CpG obs/exp ranges (interval is 0.05) was measured as shown in Figure 2, and then cumulative percentage was calculated for each of the CpG obs/exp ranges. The numbers on axis x represent the starting value of the CpG obs/exp ranges; the axis y denotes cumulative percentage of the genomic fragments.

in vertebrates as well as in sea urchin. From invertebrates to vertebrates, a conversion of CpG distribution in fragments with different GC contents is therefore exhibited.

The distance between the peaks of the three curves in Figure 2 suggests in what a degree that CpG obs/exp value differs in three GC groups. Because of the bell shape in most of the curves, we used the values at 50% of cumulative percentage to represent the peaks, and then the horizontal distance of the three curves at 50% allows us to estimate the difference of CpG frequency among the groups. The CpG obs/exp ratios for 50% cumulative percentage are shown in Table 1. The difference between >50% group and 20%–35% group is narrow in *C. elegans*, bee and fruitfly, and then turns out to be wide in sea urchin and the fish species. At last, it becomes narrow again in the warm-blooded species, accounting for the further left shifting of the curve for high GC group.

Table 1 CpG obs/exp ratios corresponding to 50% cumulative percentage of genomic fragments*

Species	CpG obs/exp ratios for 50% cumulative percentage		
	G+C=20%–35%	G+C=35%–50%	G+C>50%
C. elegans	0.92	0.92	0.81
Fruitfly	0.91	0.85	0.82
Bee	0.92	0.70	0.68
Sea urchin	0.52	0.60	0.70
Pufferfish	0.31	0.44	0.62
Zebrafish	0.33	0.43	0.67
Chicken	0.17	0.19	0.24
Human	0.09	0.18	0.22

*The G+C levels were used to classify genomic fragments. The obs/exp ratios denote CpG obs/exp values that correspond to the 50% cumulative percentage in Figure 2. More details are described in Figure 3.

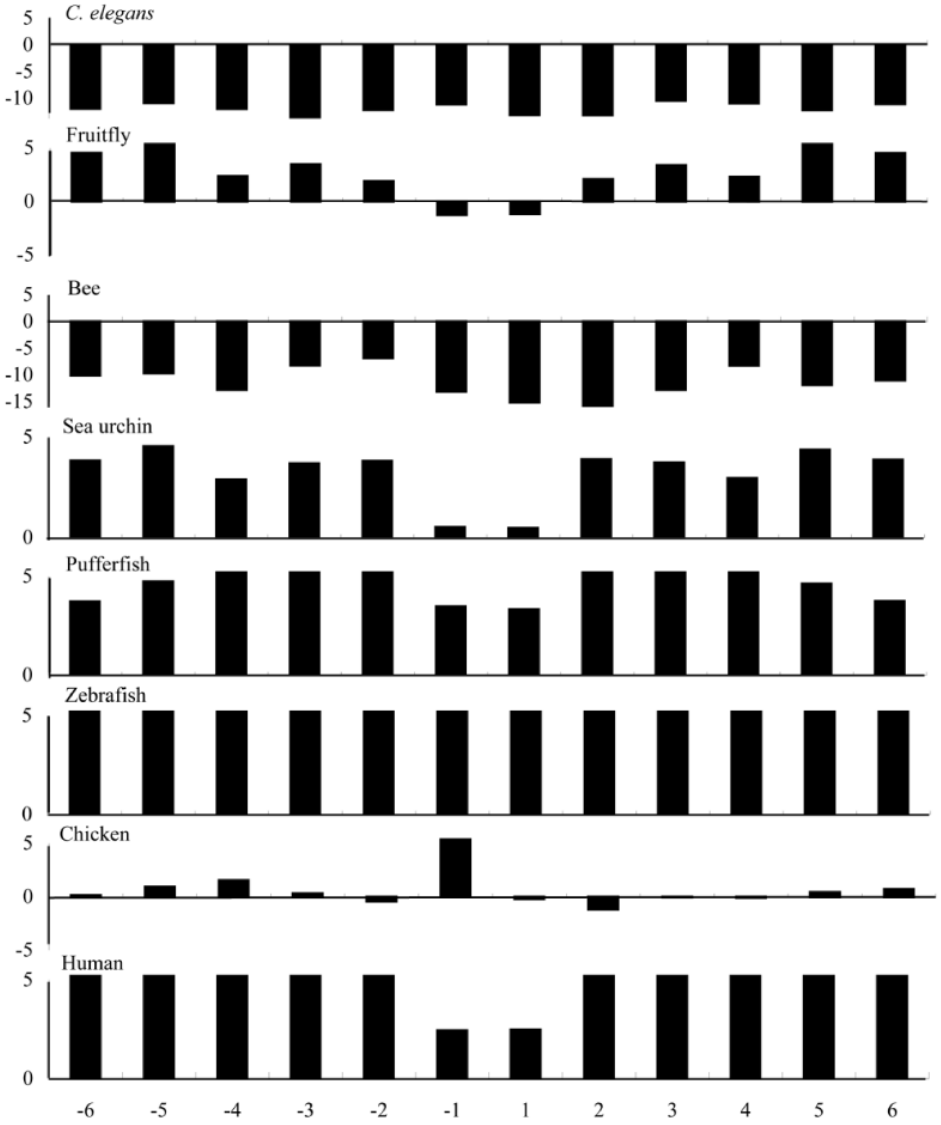

Fig. 4 G+C percent difference between real and Markov artificial sequences. The positions of −6 to +6 are the 12 flanking positions of CpG sites. The GC contents in percent were measured in real and Markov artificial sequences, and the GC content differences ($GC_{real}-GC_{artificial}$) were as shown of individual positions. The artificial sequences were Markov first-order sequences created using repeat-masked real genomic sequences.

GC content difference at CpG flanking positions

To confirm the conclusion from the above tests, we analyzed GC content at CpG flanking positions. The GC content was expected to be lower than genomic average in the invertebrates and higher than genomic average in the vertebrates and sea urchin. We used Markov sequences in the first order to generate artificial sequences and thus the original GC content at CpG flanking positions would be blurred to be genomic average. Twelve CpG flanking sites were investigated and the GC content difference between real and artificial sequences was measured to evaluate the GC enrichment at the local regions. The result is that the expectation for invertebrates was observed in *C. elegans* and bee sequences, and that for vertebrates was exhibited in all except chicken (Figure 4). Even though most of the fruitfly positions show GC enrichment, the two closing flanking positions of CpG sites display TA enrichment. For chicken, we found TA enrichment at ± 2 positions, and GC content difference between real and artificial sequences is not significant at half of the positions. It should be noted that the GC enrichment at -1 position is outstanding in chicken but is less significant in sea urchin. The inconsistency of the GC content differences among the species is understandable because our artificial sequences are the first-order Markov sequences and the GC contents at these positions more or less reflect the effect of neighboring nucleotide dependence at the CpG sites in real sequences.

Proportion of CpG islands in genomic fragments with a high GC content

CpG islands are CpG-rich regions retained after the impact of DNA methylation. The significance of CpG islands in genes and genomes has been extensively studied. Most of the studies focus on mouse and human genes and, in contrast, there are few reports on those in invertebrates and fish (*19*). We therefore addressed the frequency of CpG islands in these species.

The methods in this study allow for estimate of the proportion of CpG islands. The genomic fragments of 500 bp in >50% GC group with obs/exp values >0.6 may be regarded as CpG islands. Table 2 shows the percentage of the CpG islands in the fragments. Chicken has double the amount of CpG islands than human. However, because the association rate of these CpG islands on chicken promoters of genes (48%) is not higher than the rate in human (*20*, *21*), a large number of the chicken CpG islands under the criteria are not associated with promoters. Our results are therefore useful in adjusting the criteria to select the biologically functional CpG islands for a given species. The proportions in the other species are more than 50% (Table 2), implying that a majority of them are not functional as the CpG islands in mammals. A report has shown that the specific feature of CpG islands in fish and invertebrates is low GC content (*22*). This suggests a functional difference between the CpG islands in mammals, fish and invertebrates.

Table 2 Estimate of CpG islands in genomic fragments of high GC contents and high CpG obs/exp values*

Organism	CpG island
C. elegans	73%
Fruitfly	92%
Bee	83%
Sea urchin	62%
Pufferfish	56%
Zebrafish	61%
Chicken	15%
Human	7%

*The fragments in size of 500 bp and GC content >50% were selected by sliding bins across the genomes, and the percentage of CpG islands in criteria of length >500 bp, GC content >50% and CpG obs/exp values >0.6 could be estimated. The percentages may also be inferred from Figure 2.

Discussion

Context-dependent mutation of methylation CpG sites

In this study, we found that the regions with a high CpG obs/exp ratio were GC-poor in invertebrates and GC-rich in vertebrates. This supports our hypothesis that deamination process of methylated CpG sites in vertebrates is related to genomic TA richness. We also found that $5'$ T is not more robust than $5'$ A in affecting the CpG mutations. In previous studies, CpG mutation rates have been found to be positively related to TA richness, and variant hypotheses have been proposed to explain the correlation (*6*–*9*). This study provides solid evidence for the role of DNA

methylation in the relationship. DNA methylation seems to have a certain connection with genomic TA richness in the process of *de novo* methylation or subsequent CpG mutations. The specification of *de novo* methylation on GC-poor regions could be ruled out because GC-rich and GC-poor regions are uniformly methylated in human cells (*6*, *23*, *24*). Hence, the possibility is that methylated CpG sites in TA-rich regions mutate more easily.

The establishment of global DNA methylation in the fish species gives rise to rapid loss of CpG sites in GC-poor regions, explaining the left shifting of the lines for the low GC group and the significant separation of the three lines for different GC groups. In chicken and human, the left shifting is more obvious for all the lines and the separation of the lines becomes not significant again (Figure 3), as a result of higher mutational rate associated to higher body temperature (*15*, *25*). Intriguingly, the results for sea urchin, including the left-shift line of low-GC group and elevated GC content at CpG flanking positions, are similar to those for the vertebrates, confirming that the coincidence of CpG deficiency and low GC content is firstly observed in sea urchin, a species in which fractional methylation has been established.

To provide more evidence, we had measured nucleotide frequencies at TpG flanking sites. The same nucleotide preference as observed at CpG flanking positions was, however, not displayed. This is not surprising because CpG deficiency and TpG excess are not correlated at genome level (*26*). On the basis of these results, we propose that mutational rate of methylated CpG dinucleotide is accelerated by the flanking 5′ T and 3′ A nucleotides.

An open question is what makes the over-represented CpG sites in GC-poor regions. The possible interpretation is massive insertions of CpG sites with 5′ Ts and 3′ As in GC-poor regions. Genomic counting of TTCGAA and AACGTT patterns shows that the former is two-fold more than the latter in *C. elegans* and fruitfly (result not shown). Another example is the extreme CpG richness in GC-poor regions of the bee genome (*27*). We propose that CpGs were inserted into the genome in GC-poor regions through recombination or transposons. The evidence is that the CpG-containing codons in bee genes unmethylated are losing probably due to negative selection (result not shown).

In this study, we did not take the difference in methylation level between fish, chicken and human into account. Actually, fish species were found to show a higher methylation level than birds and mammals (*28*); even among fish species, DNA methylation level is variant as well (*29*). However, CpG is more deficient in chicken and human. Therefore, methylation level is not the essential factor for the degree of CpG deficiency, and body temperature seems to be more important in this aspect. A report has shown that deamination rate of methylation CpG sites is 20.6-fold lower in fish than in mammals (*15*). The high body temperature of chicken and human has greatly contributed to the more significant CpG deficiency, in spite of the low DNA methylation level compared to the fish.

CpG-related mutational bias resulting in GC-rich isochores

The inferred conclusion of our study acts as a complementary explanation for the origin of GC-rich isochores in mammals. Isochores are characterized by mosaic-like GC-rich and GC-poor regions (*30*), and defined to be a genomic fragment >100 kb in which sequences in overlapping 10 kb windows could not differ by >7% GC content (*31*). The isochores in current mammalian genomes are the result of extreme development of GC heterogeneity.

The driving force for isochore development in eukaryotes is under hot debate. Possible explanations can be natural selection, repair, mutational bias, changes in nucleotide pools, recombination, cytosine deamination, DNA flexibility, and biased gene conversion (*26*). Till date, no hypothesis can solely explain the phenomenon completely.

This study supports the mutational bias hypothesis. The context-dependent effect of mutagenicity on methylated CpG sites, as suggested in this study, leads to faster loss of CpG sites in GC-poor regions than in GC-rich regions, thus resulting in different mutational rates between GC-poor and GC-rich regions. Because CpG mutational rate is higher in GC-poor regions, new T and A nucleotides produced during deamination will make local GC content even lower. This is a positive feedback loop continuously putting mutational pressure on the remaining CpG sites in the regions and facilitating the formation and extension of GC-poor isochores. On the other hand, CpG mutations occur in GC-rich regions as well, even though a large number of them are supposed to be repaired by one or more mechanisms including biased gene conversion and recombination (*32*, *33*). Although the original GC content can hardly be main-

tained, decay of GC level in GC-rich regions will be unavoidable. The two processes enlarge the original difference in GC content between GC-poor and GC-rich regions, probably accounting for the isochore formation. This hypothesis is well supported by the left shifting of the lines especially for medium and high GC content groups in chicken and human, in comparison to the two fish species. All this suggests that we do not need to introduce a new mechanism enabling the original high GC contents to be even higher. The strong divergence of GC content in mammalian isochores is likely a result of heterogeneous mutational rates of CpG sites in variant GC contents.

Symmetrical distribution of pairing-nucleotides and Chargaff's second rule

An unexpected finding of the present study is a symmetrical distribution of pairing-nucleotides (A-T; G-C) at the CpG flanking positions (Figure 1). An example is that the frequency of T in any 5′ flanking positions co-varies with frequency of A in the corresponding 3′ flanking positions in the same distance to the CpG as the 5′ end T. Statistical tests showed that the covariance was highly significant in all of the species (Spearman test; $P<10^{-8}$). Actually, the figure illustrates how a eukaryotic genome abides to Chargaff's second rule. The rule states that A count and T count are almost equal on one strand of a long double-strand DNA, as well as G count and C count (*34*).

The covariance of the pairing nucleotides at the flanking positions of CpG and other self-complementary dinucleotides infers that the rule is ascribed to the strong potential of genomic fragments to form stem-loop or palindromic structures. Forsdyke (*35*) proposed a similar explanation to Chargaff's second rule on the basis of the finding by Qi and Cuticchia (*36*), in which the counts of oligonucleotides in a variety of genomes were close to those of their reversely complementary oligonucleotides. Forsdyke's point focused only on genome-wide stem-loop forming potential (*35*, *37*), whereas our results suggest that the potential of forming palindrome structures probably plays a more important role. The similar pattern as Figure 1 was also discovered for the other three self-complementary dinucleotides, TpA, ApT and GpC (results not shown). The only discrepancy was that nucleotide frequencies did not fluctuate as dramatically as at CpG flanking positions.

Such a correlation seems to be evidence of the origination of DNA from RNA world. In ancient ocean soup, the evolved RNA molecules folded into stable secondary structures that mostly comprised of stem-loop and palindromic structures. If DNA molecules were derived from double-stranded RNA molecules, the feature of inverse pairing could have been inherited by DNA molecules. These RNA-derived patterns were probably later blurred by fragment shufflings and development of functional elements. In this study, the patterns were re-constructed and visualized. Future work can hopefully answer the question that how the primary information from past era is preserved. Another probability is that our sequences still possess numerous unmasked, unidentified repetitive elements derived from RNA molecules. *Alu* repeat is one of such elements (*38*), and accounts for about 10% of the human genome (*21*).

Materials and Methods

Genomic sequences

A total of 330 Mb pufferfish scaffolds (assembly 5) were obtained from the IMCB website (http://www.fugu-sg.org); 720 Mb contigs (released in 2004-11-23) of sea urchin were from the website of Baylor College of Medicine (assembly 3); Bee 14.5 Mb shotgun sequences were downloaded from the NCBI (http://www.ncbi.nlm.nih.gov/genome/). The species-specific repeats of these DNA sequences have been masked. The whole genomes of *C. elegans*, fruitfly, zebrafish (Zv5), chicken (NCBI Build 2.1), and human (NCBI Build 35) were pooled from the NCBI. RepeatMasker (http://repeatmasker.org) was used to mask all species-specific repeats in the nematode, fruitfly, chicken, and human genomes. We located all the CpG sites and measured nucleotide frequencies at 12 flanking positions (6 at 5′ end and 6 at 3′ end). At last, nucleotide frequencies were calculated for each of the CpG flanking positions. The sequences contained many Ns because of repeat-masking and sequencing gap. If the flanking positions had Ns, the CpG site was neglected.

Calculating CpG obs/exp values in genomic regions of different GC contents

In this analysis, the genomic data of *C. elegans*, fruitfly, sea urchin, pufferfish, zebrafish, chicken, and human were used. A sliding window was used to

scan through the genomes and collect the information of sequences in bins of 500 bp, including GC content and CpG obs/exp ratio. CpG obs/exp ratio was calculated according to the formula: $F_{CpG}/(F_C{*}F_G)$, where F_{CpG} is frequency of CpG dinucleotide and F_C means frequency of C. The window containing more than 5 Ns (unknown nucleotide) was ignored. The information was first applied to classify the sequences in terms of their GC content. The three levels of GC content were: 20%–35%, 35%–50%, and >50%, representing low, medium, and high GC levels respectively. In each GC content group, the sequences were further grouped according to obs/exp values with an interval of 0.05. In a given obs/exp range, the proportion of the sequences to all was calculated. The obs/exp ranges of top 10 proportions were identified for each of the three GC content groups, and used for Kruskal-Wallis test. The test was applied to evaluate the significance of the separation of obs/exp values between the groups.

Markov first-order sequences

We used repeat-masked sequences to create Markov artificial sequences in the first order: *C. elegans* chromosome IV, fruitfly chromosome 2L, bee 14.5 Mb shotgun sequences, pufferfish 10 Mb scaffolds, zebrafish chromosome 22, chicken chromosome 3, and human chromosome 21. Due to masked repeats and sequencing gaps, our real sequences had interspersed stretches of Ns. To construct an artificial sequence in a proper length, the number of Ns was subtracted from full length of a given real sequence. The algorithm for creating the artificial sequence was that a nucleotide was added in position n+1 with the possibility determined by a nucleotide in position n in the real sequence. Therefore, a frequency matrix containing 16 possibilities was created for simulation. The possibility matrix could be calculated as follows. The amounts of A, G, C, and T in the closely successive positions of nucleotide X were calculated individually. The possibility of A following X in a real genomic sequence was calculated as N_{XpA}/N_X. The X points to A, G, C or T, and therefore the matrix hold 16 possibilities. In each extension procedure, a possibility was loaded depending on the nucleotide that showed in position X. The possibilities were determinant to the nucleotide frequencies in the following positions of nucleotide X. The first-order Markov sequence was extended until it was in the same size of the real sequence. We calculated GC content at 12 flanking positions of CpG sites (6 at 5′ end and 6 at 3′ end), and obtained GC content difference for each of the positions between the real and artificial sequences.

Authors' contributions

YW collected the datasets, conducted data analyses, and prepared the manuscript. FCCL supervised the project and co-wrote the manuscript. Both authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Karlin, S., *et al.* 1994. Heterogeneity of genomes: measures and values. *Proc. Natl. Acad. Sci. USA* 91: 12837-12841.
2. Karlin, S. and Burge, C. 1995. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* 11: 283-290.
3. Josse, J., *et al.* 1961. Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. *J. Biol. Chem.* 236: 864-875.
4. Bird, A.P. 1980. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* 8: 1499-1504.
5. Cooper, D.N. and Youssoufian, H. 1988. The CpG dinucleotide and human genetic disease. *Hum. Genet.* 78: 151-155.
6. Aissani, B. and Bernardi, G. 1991. CpG islands, genes and isochores in the genomes of vertebrates. *Gene* 106: 185-195.
7. Pesole, G., *et al.* 1997. Structural and compositional features of untranslated regions of eukaryotic mRNAs. *Gene* 205: 95-102.
8. Jabbari, K. and Bernardi, G. 1998. CpG doubles, CpG islands and Alu repeats in long human DNA sequences from different isochore families. *Gene* 224: 123-127.
9. Fryxell, K.J. and Moon, W.J. 2005. CpG mutation rates in the human genome are highly dependent on local GC content. *Mol. Biol. Evol.* 22: 650-658.
10. Lyko, F., *et al.* 2000. DNA methylation in *Drosophila melanogaster*. *Nature* 408: 538-540.
11. Gowher, H., *et al.* 2000. DNA of *Drosophila melanogaster* contains 5-methylcytosine. *EMBO J.* 19: 6918-6923.
12. Wang, Y., *et al.* 2006. Functional CpG methylation system in a social insect. *Science* 314: 645-647.

13. Tweedie, S., *et al.* 1997. Methylation of genomes and genes at the invertebrate-vertebrate boundary. *Mol. Cell. Biol.* 17: 1469-1475.
14. Simmen, M.W., *et al.* 1999. Nonmethylated transposable elements and methylated genes in a chordate genome. *Science* 283: 1164-1167.
15. Fryxell, K.J. and Zuckerkandl, E. 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol. Biol. Evol.* 17: 1371-1383.
16. Suzuki, M.M., *et al.* 2007. CpG methylation is targeted to transcription units in an invertebrate genome. *Genome Res.* 17: 625-631.
17. Rollins, R.A., *et al.* 2006. Large-scale structure of genomic methylation patterns. *Genome Res.* 16: 157-163.
18. Yoder, J.A., *et al.* 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* 13: 335-340.
19. Takai, D. and Jones, P.A. 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. USA* 99: 3740-3745.
20. International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432: 695-716.
21. International Human Genome Sequecing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.
22. Cross, S., *et al.* 1991. Non-methylated islands in fish genomes are GC-poor. *Nucleic Acids Res.* 19: 1469-1474.
23. You, Y.H., *et al.* 1998. Methylation of CpG dinucleotides in the lacI gene of the Big Blue transgenic mouse. *Mutat. Res.* 420: 55-65.
24. Caccio, S., *et al.* 1997. Methylation patterns in the isochores of vertebrate genomes. *Gene* 205: 119-124.
25. Shen, J.C., *et al.* 1994. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Res.* 22: 972-976.
26. Jabbari, K. and Bernardi, G. 2004. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene* 333: 143-149.
27. Honeybee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* 443: 931-949.
28. Jabbari, K., *et al.* 1997. Evolutionary changes in CpG and methylation levels in the genome of vertebrates. *Gene* 205: 109-118.
29. Varriale, A. and Bernardi, G. 2006. DNA methylation and body temperature in fishes. *Gene* 385: 111-121.
30. Bernardi, G., *et al.* 1985. The mosaic genome of warm-blooded vertebrates. *Science* 228: 953-958.
31. Nekrutenko, A. and Li, W.H. 2000. Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome Res.* 10: 1986-1995.
32. Eyre-Walker, A. and Hurst, L.D. 2001. The evolution of isochores. *Nat. Rev. Genet.* 2: 549-555.
33. Fullerton, S.M., *et al.* 2001. Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol.* 18: 1139-1142.
34. Chargaff, E. 1979. How genetics got a chemical education. *Ann. N Y Acad. Sci.* 325: 344-360.
35. Forsdyke, D.R. 2002. Symmetry observations in long nucleotide sequences: a commentary on the Discovery Note of Qi and Cuticchia. *Bioinformatics* 18: 215-217.
36. Qi, D. and Cuticchia, A.J. 2001. Compositional symmetries in complete genomes. *Bioinformatics* 17: 557-559.
37. Forsdyke, D.R. and Mortimer, J.R. 2000. Chargaff's legacy. *Gene* 261: 127-137.
38. Weiner, A.M., *et al.* 1986. Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annu. Rev. Biochem.* 55: 631-661.